MASTER'S DEGREE THESIS

Interuniversity Master in Statistics and Operations Research UPC-UB

Title: The association of Adherence to the Mediterranean Diet with Mortality in the EPIC-Spain cohort using Flexible Parametric Survival Models

Author: Yovaninna Alarcón Soto

Advisors: Catalina Bonet - Klaus Langohr

Department: Statistics and Operations Research

Academic year: 2015- 2016

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Interuniversity Master in Statistics and Operations Research UPC-UB

The association of Adherence to the Mediterranean Diet with Mortality in the EPIC-Spain cohort using Flexible Parametric Models

Yovaninna Alarcón Soto

Advisors:
Catalina Bonet
Klaus Langohr

Barcelona, June 6, 2016

Agradecimientos

Me gustaría agradecer a Catalina Bonet, por dirigirme a lo largo del proyecto, tener siempre nuevas ideas y puntos de vista que plantearme, y enseñarme mucho acerca de cómo hacer investigación. Siempre tuvo muy buena disposición a responder mis inquietudes, y puedo decir que aprendí mucho de ella.

A Klaus Langohr, por sus valiosos aportes y corrección, que contribuyeron a la mejora del proyecto. Por motivar además en mí las ganas de ser una mejor profesional.

Al Instituto Catalán de Oncología por facilitarme el dataset con el que trabajo, y darme la oportunidad de realizar el Trabajo Final de Máster en sus dependencias.

A Becas Chile por permitirme, a través de la beca de Magíster en el Extranjero cursar el Máster en Estadística e Investigación Operativa.

A mi familia: Sergio, Edith y Marioly, que aunque están en Chile, me han motivado desde lejos a ser siempre mejor persona y profesional. Por demostrarme amor infinito, no importando las distancias, y por ser mi motivación a seguir avanzando.

A mi familia postiza: Carmen, Jormary y Anil, que han sabido compartir conmigo sus vidas, y hacer de mi estadía por Barcelona una experiencia más enriquecedora.

A mis amigas Helen y Anil por ayudarme con la revisión de la gramática.

Contents

List of Figures

List of Tables

Abstract

Background: Modelling censored survival data is almost always done by Cox proportional-hazards regression. However, the use of Flexible Parametric Models for such data may has some advantages. For example, to introduce some flexibility in the shape of the survival curves.

Objective: To make a comparison between the Flexible Parametric Proportional-Hazards (PH) model and the traditional Cox model.

Design: This study included 41191 participants (62% female) aged 29-69 years recruited in 1992-96 from five Spanish region from the European Prospective Investigation into Cancer and Nutrition (EPIC-Spain cohort). The mean of follow-up was 18.5 years and 3646 deaths were identified. The Cox model and the Flexible Parametric PH model were used to determine the association between the Adherence to the Mediterranean Diet (mdscore) and mortality. In both models, it has been stratified by center, sex and age at recruitment and adjusted by smoke status, body mass index, physical activity, energy intake, waist circumference and educational level.

Results: Bigger adherence to the Mediterranean Diet implies less risk of mortality, according to the Flexible Parametric PH model (comparing subjects classified in the 3rd quartile of the mdscore with those in the first quartile HR=0.85, 95% CI [0.77,0.93]; 4th quartile HR=0.78, 95% CI [0.71,0.86]), similar estimates to those obtained by the stratified Cox Model (3rd quartile HR=0.84, 95% CI [0.76,0.92]; 4th quartile HR=0.76, 95% CI [0.69,0.84]).

Conclusions: The estimations obtained by the Flexible Parametric PH model are very similar to those obtained by the stratified Cox model. The Flexible Parametric model gives smooth survival curves and with more flexible shapes than those from Cox model. The implementation for delayed-entry models is correctly implemented only in `Stata`.

Keywords: Survival Analysis, proportional hazards, splines, flexible parametric model, adherence to the mediterranean diet score.

Resumen

Antecedentes: El modelamiento de datos de supervivencia censurados es casi siempre realizado mediante el modelo de regresión de Cox de riesgos proporcionales. Sin embargo, el uso de Modelos Flexibles Paramétricos tiene ciertas ventajas. Por ejemplo, para introducir cierta flexibilidad en las formas de las curvas de supervivencia.

Objetivo: Realizar una comparación entre el modelo Flexible Paramétrico de Riesgos Proporcionales y el tradicional modelo de Cox.

Diseño: Este estudio incluye 41191 participantes (62% mujeres) de entre 29-69 años reclutados entre 1992-96 pertenecientes a cinco regiones españolas de la cohorte European Prospective Investigation into Cancer and Nutrition (EPIC-España). La media de seguimiento es de 18.5 años y se identificaron 3646 muertes. El modelo de Cox y el modelo Flexible Paramétrico de Riesgos Proporcionales han sido utilizados para determinar la asociación entre la Adherencia a la Dieta Mediterránea (mdscore) y la mortalidad. En ambos modelos, se ha estratificado por centro, sexo y edad de reclutamiento y se ha ajustado además por estatus de fumador, índice de masa corporal, actividad física, ingesta energética, perímetro de cintura y nivel educacional.

Resultados: Una mayor adherencia a la Dieta Mediterránea deriva en menor riesgo de mortalidad, según el modelo Flexible Paramétrico de Riesgos Proporcionales (comparando individuos clasificados en el 3er cuartil de mdscore con aquellos en el 1er cuartil HR=0.85, 95% IC [0.77,0.93]; 4to cuartil HR=0.78, 95% IC [0.71,0.86]), estimaciones similares a las obtenidas por el modelo de Cox (3er cuartil HR=0.84, 95% IC [0.76,0.92]; 4to cuartil HR=0.76, 95% IC [0.69,0.84].

Conclusiones: Las estimaciones obtenidas para el modelo Flexible Paramétrico de riesgos proporcionales son muy similares a aquellas obtenidas a través del ajuste del modelo de Cox. El modelo Flexible Paramétrico provee curvas de supervivencia suavizadas y con formas más flexibles que las obtenidas luego de ajustar el modelo de Cox. La implementación para modelos de entrada retardada está correctamente implementada solo en `Stata`.

Palabras clave: Análisis de supervivencia, riesgos proporcionales, splines, modelo flexible paramétrico, índice de adherencia a la dieta mediterránea.

Resum

Antecedents: La modelització de les dades de supervivència censurades, gairebé sempre es realitza mitjançant el model de regressió de Cox de riscos proporcionals. No obstant això, l'ús de models flexibles paramètrics d'aquestes dades podria tenir alguns avantatges. Per exemple, per introduir certa flexibilitat en la forma de les corbes de supervivència.

Objectiu: Realitzar una comparació entre el model Flexible Paramètric de riscos proporcionals i el tradicional modelo de Cox.

Disenny: Aquest estudi inclou 41191 participants (62% dones) d'entre 29-69 anys d'edat reclutats entre 1992-1996 pertanyents a cinc regions de la cohort European Prospective Investigation into Cancer and Nutrition (EPIC-Espanya). La mitjana de seguiment és 18.5 anys i es van identificar 3646 morts. El model de Cox i el model Flexible Paramètric de Riscos Proporcionals han estat utilitzats per determinar l'associació entre l'Adherència a la Dieta Mediterrània (mdscore) i la mortalitat. En ambdós models, s'ha estratificat per centre, sexe i edat de reclutament i s'ha ajustat per estatus de fumador, índex de massa corporal, activitat física, ingesta d'energia, perímetre de cintura i nivell educacional.

Resultats: Una major adherència a la Dieta Mediterrània deriva en menor risc de mortalitat, segons el model Flexible Paramètric de Riscos Proporcionals (comparant els subjectes classificats en el 3er quartil de l'mdscore amb aquells en el primer quartil HR=0.85, 95% IC [0.77,0.93]; 4t quartil HR=0.78, 95% IC [0.71,0.86]), estimacions similars a les obtingudes pel model de Cox (3er quartil HR=0.84, 95% IC [0.76,0.92]; 4t quartil HR=0.76, 95% IC [0.69,0.84]).

Conclusions: Les estimacions per al model Flexible Paramètric de riscos proporcionals són molt similars a aquelles obtingudes a través de l'ajust del model de Cox. El model Flexible Paramètric proveeix de corbes de supervivència suavitzades i amb formes més flexibles que les obtingudes després d'ajustar el model de Cox. La implementació per a models d'entrada retardada està correctament implementada només en `Stata`.

Paraules clau: Anàlisi de supervivència, riscos proporcionals, splines, model

Notation

AIC	Akaike Information Criterion
BIC	Bayes Information Criterion
BMI	Body Mass Index
cm	Centimeter
CI	Confidence Interval
$H(t)$	Cumulative Hazard Function
df	Degrees of Freedom
$F(t)$	Distribution Function
EPIC	European Prospective Investigation into Cancer and Nutrition
$h(t)$	Hazard Function
HR	Hazard Ratio
K-M	Kaplan-Meier
kcal	Kilocalories
kg	Kilograms
PLRT	Partial Likelihood Ratio Test
MD	Mediterranean Diet
$f(t)$	Probability Distribution Function
PH	Proportional Hazards
PHA	Proportional Hazards Assumption
RP	Royston-Parmar
S.E.	Standard Error
$S(t)$	Survival Function
WHO	World Health Organization

Chapter 1

Introduction

The Cox model (Cox, 1972) has played a vital role in applied survival analysis during the last three decades. The model and its software implementations have popularized survival analysis and made it accessible to researchers in varied disciplines, who are not necessarily statisticians. It has been so successful that it is probably used in most practical analyses of the effects of covariates on survival.

Royston-Parmar models (Royston and Parmar, 2002) also known as Flexible Parametric Models, in some important aspects go beyond the Cox model and beyond the standard parametric survival models. Weibull, loglogistic, and lognormal models are generalized to proportional hazards (PH), proportional odds (PO), and probit-scaled Royston-Parmar (RP) models, respectively. In this study, the Flexible Parametric PH model is used, which overcomes the problems of potentially poor fit of standard parametric models and of the "noisy" estimates of the hazard and survival functions associated with the Cox model and with non-parametric estimators such as the Kaplan-Meier estimator (Royston and Lambert, 2011).

According to what has been said previously, the main objective of this study is to make a comparison between the Flexible Parametric Proportional-Hazards model and the traditional Cox Model.

In order to do this, data from the EPIC-Spain cohort was analysed. European Prospective into Cancer and Nutrition (EPIC) is a big prospective study about the relationships between diet, anthropometric measures, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases.

According to the diet per each individual, the density of intake of vegetables, fruits, legumes, fish, cereal, olive oil, wine, meat and dairy products was used to obtain four different versions of the Adherence of Mediterranean Diet index. These four versions were compared, and the most appropriate of them

was chosen as the main variable of interest (mdscore) to explain the mortality. Using this mdscore, both, Flexible Parametric PH and Cox models were fitted, stratifying by sex, geographical center an age at recruitment. The models also included other covariates such as smoke status, body mass index, physical activity, energy intake, waist circumference and educational level.

In Chapter 2, the database with which this study is developed, corresponding to the EPIC-Spain cohort, its recruitment, the dietary and lifestyle questionnaires, its follow-up and assessment of vital status of each individual is presented.

In Chapter 3, the Mediterranean Diet, its definition and some previous results about its importance to increase longevity is described. Also, this chapter presents different definitions of the score for Adherence to the Mediterranean Diet (mdscore), which is the main variable for this study.

In Chapter 4, the statistical methods which are used are presented, starting with some basic concepts, Cox Model, stratified Cox Model, Royston-Parmar Models and their implementations, mainly in Stata.

In Chapter 5, the variables of the study, its categories, and the specification of the reference categories are explained. Also the population of the study is described, as the total of exclusions, number of events and follow-up time.

In Chapter 6 the main results of the study are given. The results are divided in sections such as choosing the best version of index of Adherence to the Mediterranean Diet, models fitted (Cox and Flexible Parametric PH model) and comparison between models.

In Chapter 7 the main conclusions and discussion are presented, also the further research according to this study.

Chapter 2

EPIC-Spain

The European Prospective Investigation into Cancer and Nutrition (EPIC) is one of the largest studies in the world, with more than half a million (521000) participants recruited across 10 European countries: Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and the United Kingdom, and followed for almost 15 years, whose methodological details have been published previously (Riboli et al., 2002; Riboli and Kaaks, 1997; Bingham and Riboli, 2004). The present study uses the data from the Spanish cohort.

EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases. EPIC researchers are active in all fields of epidemiology, moreover important contributions have been made in nutritional epidemiology using biomarker analysis and questionnaire information, as well as genetic and lifestyle investigations (EPIC study, 2016).

2.1 Recruitment

Concerning the Spanish cohort, the EPIC study is being conducted in five regions: Asturias, Gipuzkoa, Navarra, Murcia, and Granada (for number of participants by Spanish center and its percentage distribution to the global study see Table 2.1). Recruitment began in 1992-1993, and it was finished in 1996. The cohort in Spain consists of 41438 participants with interviews on diet, and 39880 participants with blood samples available (EPIC Spain, 2016).

Information on diet, lifestyle factors, anthropometric measurements and a blood sample were obtained at baseline. At recruitment, all participants gave their informed consent, and the study was approved by the Medical Ethical Committee of the Bellvitge Hospital (Barcelona) (Buckland et al., 2009).

Table 2.1: Number of participants by Spanish center and its percentage distribution to the global study.

	Males	Females	Number of participants	Percentage of EPIC cohort
Asturias	3083	5459	8542	1.64%
Granada	1796	6083	7879	1.51%
Murcia	2685	5831	8516	1.63%
Navarra	3908	4176	8084	1.55%
Gipuzkoa	4158	4259	8417	1.61%
Spain	15630	25808	41438	7.95%

Source: http://epic.iarc.fr/centers/spain.php

2.2 Dietary questionnaire

Information on the usual diet over the last 12 months was collected at recruitment by means of an interview-administered computerised version of a dietary history questionnaire that had been previously validated in Spain and used at all centers. The questionnaire was open, but was structured by meals and included a list of 662 common foods and recipes from each region. The recipes were broken down into simple foods, and the frequency of consumption of them at least twice a month was recorded, taking seasonal variability into account. To determine the quantity of each food actually consumed, the portion size of each food item was determined by several methods: a set of 35 photographs, natural units, household measurement or geometric figures (EPIC Group of Spain, 1997a). A specific food composition table was used to calculate the daily energy and nutrient intake (Slimani et al., 2007; EPIC Group of Spain, 1997b).

A dietary calibration study was conducted over a sub sample who completed a standardized 24-hour diet recall, so dietary data were scaled by using an additive calibration. For further information, see Kaaks and Riboli (1997).

2.3 Assessment of lifestyle variables

Information on lifestyle and other health-related factors were obtained by an interviewer-administered questionnaire at recruitment. It included questions on education and socio-economic status, history of previous illnesses, history of tobacco use, physical activity, medical and reproductive history. Measurements of height, weight, and hip and waist circumferences were taken using standardized procedures.

2.4 Follow-up and assessment of the vital status

Follow-up consists of a computerized version of a follow-up questionnaire. In 1996-1999, 40755 participants were interviewed by phone, using this computerized method. Follow-up for the identification of cancer cases is done every 4 years. It is based on a computerized record linkage programme that links EPIC files with the population cancer registries of Asturias, Gipuzkoa, Granada, Murcia, and Navarra. Other sources of information (hospital discharge databases, pathology reports) are also being used. The main source of mortality data is the National Mortality Registry of the National Institute of Statistics (Instituto Nacional de Estadística, INE). Other regional sources are used, and in some centers letters to the members of the Spanish cohort are being sent each year to update the vital status. By December 2007 (record linkage done in 2010), 3646 new cases of cancer had been diagnosed and 1972 participants had died.

Chapter 3

Adherence to the Mediterranean Diet

3.1 Mediterranean Diet

Although different regions in the Mediterranean basin have their own diets, it is appropriate to consider them as variants of a single entity, the Mediterranean Diet (MD). Indeed, the dietary patterns that prevail in the Mediterranean have many common characteristics, most of which stem from the fact that olive oil occupies a central position in all of them. Olive oil is important not solely for its own health benefit; it is also associated with the consumption of large quantities of vegetables in the form of salads and equally large quantities of legumes in the form of cooked food. Thus, it might be convenient, if not wholly accurate, to define the MD as the dietary pattern around in the olive-growing areas of the Mediterranean region in the late 1950s and early 1960s, when the consequences of World War II were overcome, but the fast-food culture had not yet invaded the area (Trichopoulou and Lagiou, 1997).

The MD is globally recognised as a healthy dietary model and also the United Nations Educational, Scientific and Cultural Organization (UNESCO) declares the MD as intangible cultural heritage of humanity (UNESCO, 2016).

The traditional MD pattern is characterized by the daily use of olive oil, an abundance of plant foods such as fruit and vegetables, nuts and seeds, legumes and cereals (that in the past were largely unrefined), the consumption of fish and seafood (depending on the proximity of the sea), moderate-to-low intake of dairy products mostly from fresh cheese and yoghurt, moderate alcohol mostly in the form of wine, and a less frequent consumption of meat and meat products.

Numerous epidemiological studies have explored the health benefits of the MD and evidence consistently shows that individuals who adhere to the MD have healthier ageing and a longer life expectancy (Pérez-López et al., 2009; Roman et al., 2008). This is related, in part, to its role in preventing major chronic diseases such as cardiovascular disease, certain cancers, type 2 diabetes and also some neurodegenerative diseases, as supported by findings from observational studies (Roman et al., 2008).

The MD's favourable fatty acid profile, high fibre content, wide variety of antioxidants and phytochemicals, other still unidentified biologically active compounds and their synergistic interactions can explain some of its beneficial effects on health (Pérez-López et al., 2009).

3.2 Measurement of Adherence to the Mediterranean Diet

Over two decades ago the key elements of the MD were grouped into an a-priori MD score (Trichopoulou et al., 1995), to reflect the level of adherence to this dietary pattern. Various versions of this and other MD scores are now widely used to study the relationship between the MD pattern and different health parameters. Prospective studies have shown that following the MD is associated with a decrease in overall mortality (Mitrou et al., 2007; Sofi et al., 2008; Trichopoulou et al., 2003, 2005). A Mediterranean-like diet has also been reported to have a beneficial effect on mortality in countries outside the Mediterranean basin (Mitrou et al., 2007).

In this study, each participant's degree of adherence to a MD was evaluated by using four different versions of Mediterranean Diet score (mdscore), which are variation of the original Mediterranean Diet score (for further information, see Trichopoulou et al. (1995, 2003)). These versions are based on the intake of 9 key components of this diet; seven of these components presumed to fit the Mediterranean Diet: fruit (including nuts and seeds but excluding fruit juices), vegetables (excluding potatoes), legumes, cereals (including whole-grain and refined flour, pasta, rice, other grains, and bread (69.5% of total cereals)), fresh fish (including seafood), olive oil and wine (in the original version total alcohol consumption was used, in this case is only wine, because is the most consumed in Spain). The last 2 components presumed not to fit the Mediterranean Diet are total meat (including processed meat) and dairy products (including low-fat and high-fat milk, yoghurt, cheese, cream desserts, and dairy and non-dairy creams). Each mdscore component apart from wine was divided for daily energy intake (kcal/day) without taking into account alcohol. In the case of wine consumption was divided for total daily alcohol intake (kcal/day).

Nutrient intake divided by total energy intake is called "nutrition density approach" and is the traditional method for accounting for total energy intake in nutritional studies and epidemiologic analyses. The nutrition density method has several advantages: it can be calculated directly for an individual without the use of any statistical models, it is familiar to nutritionist as a measure of dietary composition, and it has been used in national dietary guidelines (Willet et al., 1997).

Taking this into account, these are the four different versions of the index. All of these four versions for mdscore were calculated with `R` project (For `R` code see A) . The definitions are given below.

Let X_i, $i = 1, \ldots, 8$ be the nutritional density of each component, that is

$$X_i = \frac{\text{consumption of } i\text{th component (g/day)}}{\text{energy intake (kcal/day)}}$$

for each individual.

In the case of wine, its nutritional density is given by

$$X_9 = \frac{\text{consumption of wine (g/day)}}{\text{total alcohol intake (kcal/day)}}$$

Then, let X_1 be the nutritional density of vegetables, X_2 fruits, X_3 legumes, X_4 fish, X_5 cereals, X_6 olive oil, X_7 meat, X_8 dairy products and X_9 wine. Also considering α_i as

$$\alpha_i = \begin{cases} 1 & \text{if component fits to the MD} \\ -1 & \text{if component does not fit} \end{cases}$$

$\mathbf{mdscore}_{sum}$

This index consists only of the sum of the 9 nutritional densities, the seven components presumed to fit a Mediterranean Diet are added, and the 2 presumed not to fit are subtracted:

$$\begin{aligned} mdscore_{sum} &= \sum_{i=1}^{9} \alpha_i X_i \\ &= X_1 + X_2 + X_3 + X_4 + X_5 + X_6 - X_7 - X_8 + X_9 \end{aligned}$$

mdscore$_{cdf}$

By participant, the cumulative distribution function (cdf) was considered for each nutritional density of the score ($X_{cdf_1}, \ldots, X_{cdf_9}$), it enables the range of each component varying between 0 and 1. Then, the seven components which presumed to fit a Mediterranean Diet are added, and the 2 presumed not to fit are subtracted.

$$\begin{aligned} mdscore_{cdf} &= \sum_{i=1}^{9} \alpha_i X_{cdf_i} \\ &= X_{cdf_1} + X_{cdf_2} + X_{cdf_3} + X_{cdf_4} + X_{cdf_5} + X_{cdf_6} - X_{cdf_7} \\ &\quad -X_{cdf_8} + X_{cdf_9} \end{aligned}$$

Also this score was divided into quartiles, as a second version of itself. A value of 0, 1, 2 and 3 was assigned to the first, second, third and fourth quartiles of intake.

mdscore$_{sd}$

For this index, the mean and the standard deviation of the centers of Granada and Murcia are obtained, due to a more Mediterranean condition. Then, for each participant, all the 9 nutritional densities were standardized according to the previous mean and standard deviation. After this, each value, let say X_{sd}, was categorized as follows:

$$X_{sd} = \begin{cases} -2 & \text{if } x < -2 \\ -1 & \text{if } x \in [-2, -1) \\ 0 & \text{if } x \in [-1, 1] \\ 1 & \text{if } x \in (1, 2] \\ 2 & \text{if } x > 2 \end{cases}$$

The mdscore$_{sd}$, is defined as follows:

$$\begin{aligned} mdscore_{sd} &= \sum_{i=1}^{9} \alpha_i X_{sd_i} \\ &= X_{sd_1} + X_{sd_2} + X_{sd_3} + X_{sd_4} + X_{sd_5} + X_{sd_6} - X_{sd_7} \\ &\quad -X_{sd_8} + X_{sd_9} \end{aligned}$$

$\mathbf{mdscore}_{ter}$

This index was based on the article of Buckland et al. (2009). Each of the 9 nutritional densities were divided into tertiles. Values of 0, 1 and 2 were assigned to the first, second and third tertiles of intake if the component fit to the MD. Values of 2, 1 and 0 were assigned to the first, second and third tertiles of intake if the component does not fit to the MD. For each participant, the points received from each of the 9 components ($X_{ter_1}, \ldots, X_{ter_9}$) were summed to give an individual mdscore, as in the previous versions:

$$\begin{aligned} mdscore_{ter} &= \sum_{i=1}^{9} X_{ter_i} \\ &= X_{ter_1} + X_{ter_2} + X_{ter_3} + X_{ter_4} + X_{ter_5} + X_{ter_6} + X_{ter_7} \\ &\quad + X_{ter_8} + X_{ter_9} \end{aligned}$$

Chapter 4

Methods

In this chapter, basic concepts such as the definitions of the survival function, hazard function and cumulative hazard function are presented. Cox model and stratified Cox model are also explained, emphasizing the weaknesses of these models, leading to the definition of Flexible Parametric models. Since the Flexible Parametric model is the central thread of this work, its theoretical basis and its implementation in the Stata software are explained with greater detail.

4.1 Basic Concepts

Let be T the time until the event of interest, ε. In the present work the event ε is defined as death. Formally, T is a non-negative random variable. All the concepts described below can be found in Gómez et al. (2015).

The distribution of the model for T can be characterized by the survival function, $S(t)$, the hazard function, $h(t)$, or the cumulative hazard function, $H(t)$. Each of them serves to illustrate different aspects of the distribution of T.

4.1.1 Survival Function

The survival function is denoted by S and corresponds to the probability of an individual surviving beyond time t (experiencing the event after time t). It is defined as

$$S(t) = Pr(T > t)$$

for $t \geq 0$.

Basic properties

- $S(0) = 1$ and $S(\infty) = 0$,
- $S(t)$ is a monotonically decreasing function,
- if T is continuous, $S(t)$ is continuous and strictly decreasing.

The survival function can take different forms, but basically all start from 1, decrease monotonically and converge to zero when t tends to infinity.

4.1.2 Hazard Function

The hazard function when T is a random absolutely continuous variable, is defined by:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr[t \leq T < t + \Delta t | T \geq t]$$

Intuitively $h(t)\Delta t$ can be seen if we know any individual has survival time t, as the probability that ε occurs in $(t, t + \Delta t]$.

The hazard function describes any aspect of a probability distribution, and would be estimated by the proportion of people who fail at time t among those who had not previously failed.

The risk function expressed as the risk changes over time containing the same information as the survival but in terms of its speed (or rate) of change. When the risk is high, survival declines quickly, whereas if the risk is zero the survival curve is flat.

Basic properties, for T absolutely continuous

- $h(t)$ is a non-negative function,
- $\int_0^s h(u)du < \infty$ for any $s > 0$ and $\int_0^\infty h(u)du = \infty$,
- $h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t))$.

Due to the dynamic nature of survival data, a characterization of the distribution by the hazard function is generally very convenient since this function does not change with condition because it is already conditioned on survival time.

4.1.3 Cumulative Hazard Function

The cumulative hazard function, $H(t)$, when T is absolutely continuous, is defined by

$$H(t) = \int_0^t h(u)du$$

It is very useful graphically, also technically but does not have an intuitive interpretation.

Basic Properties

- $H(t)$ is a non-negative function,
- $H(t)$ is a monotonic increasing function,
- $H(t) = -\ln S(t)$, or equivalently, $S(t) = \exp\{-H(t)\}$.

4.2 Censoring and Truncation

One difficulty of survival analysis is the incomplete information on the survival of some individuals, that is to say, the exact time until the event occurs is not observed, either because the event of interest ε occurs before the person enters the study, or because when the study ends ε has not happened yet, and in general because all the knowledge is that ε has occurred within a certain time interval. These peculiar characteristics of survival studies are known under the name of censoring.

There are various categories of censoring, such as right-censoring, left-censoring, and interval-censoring. In this study the interest is on right-censoring and left-truncation, due to the nature of the survival database.

Right censoring

Individuals are followed until the specific event ε occurs or until the end of the study. If ε is observed during the study, the time until ε is known. If for an individual event ε has not occurred during the course of the study, the observation is said to be right-censored. The right-censoring can occur due to any of the following circumstances:

- Finishing the study at a predetermined time. The aim is to extract conclusions on the time to failure (death) from the data collected. This situation occurs in clinical trials because, for example, some individuals survived after the end of it.

- The monitoring of some individuals has been lost (lost to follow-up). These individuals are observed only during part of the period of observation. This can be due to many reasons, for example, a change of residence or a change of hospital. In such situations, it has to follow the individual to ensure that the loss is not related with the disease.
- The event is produced by other than the cause of interest.

Then, let be C_R a pre-specified time, for an arbitrary individual the random variables (Y, δ) are defined as

$$\begin{aligned} Y &= \min\{T, C_R\} \\ \delta &= \begin{cases} 1 & \text{if } T \leq C_R: \text{ not-censored data} \\ 0 & \text{if } T > C_R: \text{ censored data} \end{cases} \end{aligned}$$

The random variable δ is the indicator of not-censoring, although is usually known as the indicator of censoring.

Left-truncation

Left truncation occurs in a cohort when subjects at risk prior to baseline do not remain observable until the start of follow-up, it is said L.

Left-truncated data are also known as delayed data entries. If, for example, L is the time of truncation, only individuals with $T \geq L$ are observed.

4.3 The Kaplan-Meier estimator of S(t)

There are different ways to estimate the survival function $S(t)$, one of these is the Kaplan-Meier (K-M) estimation (Kaplan and Meier, 1958). This is a non-parametric estimator that takes into account the right-censoring.

Let $T_1, T_2, \ldots, T_n$ be a sample of the population of interest and (Y_i, δ_i) defined as before.

The notation $t_{\max}$ is the maximum of the observations, ie: $t_{\max} = Y_{(n)}$ and t_{last} for maximum of the failure times, that is to say: $t_{last} = max\{Y_i; \delta_i = 1\}$.

Let be $Y_{(1)} < Y_{(2)} < \ldots Y_{(n)}$ the order statistic and δ_i the value of δ for $Y_{(i)}$. The K-M estimator supports the general form for all $t \leq t_{\max}$, that is, for all t in the range of data existing:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < Y_{(1)} \\ \prod_{i:Y_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) & \text{if } t \geq Y_{(1)} \end{cases}$$

where d_i is the number of deaths at time $Y_{(i)}$, and n_i is the number of subjects at risk just before $Y_{(i)}$.

If the last ordered observation is censored, then the $\lim_{t\to\infty} \hat{S}(t) > 0$ and the estimator is not well defined. One suggestion is given by the redefinition of $\hat{S}(t) = 0$ for $t \geq Y_{(n)}$.

4.4 The Cox Proportional Hazards Model

In this section, the widely used multiplicative hazards model (Cox, 1972), often called the proportional hazards model is presented.

As before, let T denote the time to some event. The data, based on a sample of size n, consists of the triple $(T_i, \delta_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ where T_i is the time on study for the ith patient, δ_i is the event indicator for the ith patient ($\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if the lifetime is right-censored) and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{is})^t$ is the vector of covariates or risk factors for the ith individual at time t which may affect the survival distribution of T. Here the vectors $\boldsymbol{x}_{ik}$, $k = 1, \ldots, s$, are covariates whose fixed values are known at time 0, such as sex, center, educational level, etc.

Let $h(t|\boldsymbol{x})$ be the hazard rate at time t for an individual with risk vector $\boldsymbol{x}$. The basic model (Cox, 1972) is as follows:

$$h(t|Z) = h_0(t)c(\beta^t \boldsymbol{x})$$

where $h_0(t)$ is an arbitrary baseline hazard rate, $\beta = (\beta_1, \ldots, \beta_s)^t$ is a parameter vector, and $c(\beta^t \boldsymbol{x})$ is a known function. This is called a semi-parametric model because a parametric form is assumed only for the covariate effect. The baseline hazard rate is treated non-parametrically. Because $h(t|\boldsymbol{x})$ must be positive, a common model for $c(\beta^t \boldsymbol{x})$ is

$$c(\beta^t \boldsymbol{x}) = \exp(\beta^t \boldsymbol{x}) = \exp\left(\sum_{k=1}^{s} \beta_k \boldsymbol{x}_k\right)$$

yielding

$$h(t|X) = h_0(t)\exp(\beta^t \boldsymbol{x}) = h_0(t)\exp\left(\sum_{k=1}^{s} \beta_k \boldsymbol{x}_k\right) \tag{4.1}$$

The Cox model is often called a proportional hazards model because, if we look at two individuals with covariate values $\boldsymbol{x}$ and $\boldsymbol{x}^*$, the ratio of their hazard rates is:

$$\frac{h(t|\boldsymbol{x})}{h(t|\boldsymbol{x}^*)} = \frac{h_0(t)\exp[\sum_{k=1}^{s} \beta_k \boldsymbol{x}_k]}{h_0(t)\exp[\sum_{k=1}^{s} \beta_k \boldsymbol{x}_k^*]} = \exp\left[\sum_{k=1}^{s} \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_k^*)\right] \tag{4.2}$$

which is a constant. So, the hazard rates are proportional. The quantity (4.2) is called the relative hazard (hazard ratio) of an individual with risk factor $\boldsymbol{x}$ having the event as compared to an individual with risk factor $\boldsymbol{x}^*$. In particular, if $\boldsymbol{x}_1$ indicates the body mass index effect ($\boldsymbol{x}_1 = 1$ if the person has obesity and $\boldsymbol{x}_1 = 0$ if the body mass index is normal) and all other covariates have the same value, then, $h(t|\boldsymbol{x})/h(t|\boldsymbol{x}^*) = \exp(\beta_1)$, is the instantaneous risk of having the event in t if the individual have obesity relative to the risk of having the event if the individual have normal weight given that $T \geq t$.

4.4.1 Estimation of the model parameters

The standard way to estimate the β_j coefficients is by maximizing the partial likelihood function called $L(\beta)$. The estimators obtained comply with generally good properties of the maximum likelihood method.

Suppose there are r times to the event ε (death), $n-r$ times of censoring and no ties. Denote by $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ the r ordered death times, by $R_j = R(t_{(j)}) = \{i : Y_i \geq t_{(j)}\}$ the set of all individuals at risk of dying at time $t_{(j)}$, that is to say, the set of all those individuals who are alive and not censored at time $t_{(j)}$- and by n_j =card(R_j) the number of individuals at risk in $t_{(j)}$. Denote also the set containing all the information in the sample $\Gamma = \{(Y_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ as before.

The basic principle of the deduction of the partial likelihood function resides in the fact that knowledge of r death times $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ with the labels $e_1, e_2, \ldots, e_r$ indicating which individual corresponds the death, is equivalent to the original data (this is certainly true if there is no censoring). For more details, see (Gómez et al., 2015).

The partial likelihood function is defined as

$$L(\beta_1, \ldots, \beta_s) = \prod_{j=1}^{r} P\{e_j = i|\Gamma_j\} = \prod_{j=1}^{r} P\{\boldsymbol{x}_{(j)} = x_{(j)}|\Gamma_j\}$$

and is interpreted as the product, for each time of death, of the conditional probabilities that the individual whose vector of covariates is $x_{(j)}$ dies at time $t_{(j)}$ knowing death has occurred among n_j individuals at risk at time $t_{(j)}$. Therefore, the partial likelihood is equal to

$$L(\beta_1, \ldots, \beta_s) = \prod_{j=1}^{r} \frac{\exp\{\beta' x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\beta' x_{lj}\}},$$

or equivalently

$$L(\beta_1, \ldots, \beta_s) = \prod_{i=1}^{n} \left(\frac{\exp\{\beta' x_i\}}{\sum_{l \in R(Y_i)} \exp\{\beta' x_l\}} \right)^{\delta_i}, \tag{4.3}$$

and its logarithm can be expressed as

$$\ln L(\beta_1, \ldots, \beta_s) = \sum_{i=1}^{n} \delta_i \left(\beta' x_i - \ln \sum_{l \in R(Y_i)} \exp\{\beta' x_l\} \right) \tag{4.4}$$

Using the partial likelihood, the dependence of the underlying risk function $h_0(t)$ is removed. Note the partial likelihood function can be rewritten as

$$L(\beta_1, \ldots, \beta_s) = \prod_{j=1}^{r} \frac{\exp\{\sum_{k=1}^{s} \beta_k x_{(j)k}\}}{\sum_{l \in R(t_{(j)})} \exp\{\sum_{k=1}^{s} \beta_k x_{jk}\}},$$

where the numerator only depends on the information for individual experiencing the event, while the denominator depends on all those who have it not experienced yet. The estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_s)$ is obtained maximizing the partial likelihood (4.3), or equivalently maximizing the logarithm of the partial likelihood (4.4). The maximization of this function using numerical methods provides the corresponding estimators. It can be shown that the maximum likelihood $\hat{\beta}$ obtained from maximizing the partial likelihood is asymptotically unbiased, efficient and normal. For more details, see Gómez et al. (2015).

4.4.2 Residual Analysis in the Cox Model

Diagnostic procedures for model checking are known as essential parts of a modelling process. Many of these procedures are based on residuals. In survival analysis, when a Cox model is established, different types of residuals can be considered for different purposes: Schoenfeld residuals for checking the proportional hazards assumption for a covariate, Score residual for the determination of influential observations and Deviance residuals used to examine overall test of the goodness-of-fit of a Cox model. For more detail see Gómez et al. (2015).

4.5 Stratified Cox Model

An alternative to extending the Cox model to deal with non-proportional hazards is to stratify over the covariates that do not satisfy the proportional hazards assumption (PHA). In essence, stratification involves fitting a model that has a different baseline hazard in each stratum, although the effect of other covariants on survival is the same.

The set of observed data is given by

$$D = \{Y_i, \delta_i, Z_i, [\boldsymbol{x}_i, 0 \leq t \leq Y_i], i = 1, 2, \ldots, n\}$$

Where, as before, Y_i is the time on study for patient ith, δ_i is the censorship indicator, $\boldsymbol{x}_i$ is the vector of covariates for the same individual and Z_i is the

stratum to which the individual belongs. Lets assume Z_i is a categorical variable with d values.

The hazard function for individuals within stratum j ($j = 1, \ldots, d$) is given by

$$\begin{aligned} h_j(t|\boldsymbol{x}) &= \exp\{\beta'\boldsymbol{x}\}h_{0j}(t) \\ &= \exp\{\beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \ldots + \beta_s\boldsymbol{x}_s\}h_{0j}(t). \end{aligned}$$

By this model it is assumed that the regression coefficients are the same for each stratum but the underlying hazard functions may differ among themselves and not need to be related.

The partial likelihood for stratum j is denoted by $L_j(\beta)$, as in (4.3), using only individuals for stratum j. The joint partial likelihood function is built as the product of the partial likelihood $L_j(\beta)$, that is to say,

$$L(\beta) = L_1(\beta) \cdot L_2(\beta) \cdot \ldots \cdot L_s(\beta)$$

The methods for estimation and hypothesis test are deduced in a simple way from the logarithm of the joint partial likelihood function, i.e., from

$$\ln L(\beta) = \ln L_1(\beta) + \ln L_2(\beta) + \ldots + \ln L_s(\beta).$$

4.6 Royston-Parmar Models

In this section, a parametric approach to survival analysis is described. This approach introduces flexibility in the shapes of survival functions that can be modelled. Royston-Parmar (RP) models are also known as flexible parametric models. These models are a generalization of the Weibull, loglogistic and lognormal models. The Weibull generalization gives the proportional hazards (PH) RP model, the loglogistic generalization gives the proportional odds RP model, and the lognormal generalization gives the probit-scaled RP model. In this study, and in order to make the comparison with the Cox Model, it is more appropriated work with the Flexible Parametric PH model.

4.6.1 Flexible Parametric Proportional Hazards model

A common parametric model for survival data is the Weibull model. The Weibull model is a proportional hazards model, but is often criticized for its lack of flexibility in the shape of the baseline hazard function, which is either monotonically increasing or decreasing (Lambert and Royston, 2009).

The cumulative hazard function, $H(t)$, for a Weibull distribution is

$$H(t) = \lambda t^{\gamma_1} \tag{4.5}$$

for some $\gamma_1 > 0$, where γ_1 is a shape parameter. The Weibull hazard function

$$h(t) = \frac{dH(t)}{dt} = \lambda\gamma_1 t^{\gamma_1 - 1}$$

is constant when $\gamma_1 = 1$ (that is, the exponential sub-case), monotonic increasing when $\gamma_1 > 1$, and monotonic decreasing when $\gamma_1 < 1$.

The expression (4.5) can be written in logarithmic form as:

$$\ln[H(t)] = \ln(\lambda) + \gamma_1 \ln(t) = \gamma_0 + \gamma_1 \ln(t) \tag{4.6}$$

Where $\gamma_0 = \ln(\lambda)$. As you see $\ln[H(t)]$ is a sum of two components: a constant (γ_0) and a linear function of log time ($\gamma_1 \ln(t)$). If a covariate vector, $\boldsymbol{x}$, is included in the survival model, the basic Weibull model (4.6) can be written with covariates $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s)^t$ and parameter vector $\beta = (\beta_1, \ldots, \beta_s)^t$

$$\ln[H(t|\boldsymbol{x}_i)] = \gamma_0 + \gamma_1 \ln(t) + \boldsymbol{x}_i\beta \tag{4.7}$$

Thus the log baseline cumulative hazard function $\gamma_0 + \gamma_1 \ln(t)$ with covariates additive on this scale. The basic idea of the Flexible Parametric approach is to relax the assumption of linearity of log time by using restricted cubic splines.

From the model (4.7) is necessary to highlight:

- Under the proportional hazard assumption the covariates can still be interpreted as (log) hazard ratios since proportional hazards also imply proportional cumulative hazards.
- The cumulative hazard as a function of log time is generally a stable function, in fact, in all Weibull models it is a straight line. It is easier to accurately capture the shape of more stable functions.
- It is easy to transform to the survival and hazard functions

$$S(t) = \exp[-H(t)] \qquad h(t) = \frac{d}{dt}H(t)$$

4.6.2 Restricted Cubic Splines

Splines are flexible mathematical functions defined by piecewise polynomials, with some constraints that ensure the overall curve is smooth. The point at which the polynomials join are called knots. Regression splines are useful as they can be incorporated into any regression model with a linear predictor.

Let $s(x)$ denote a non-linear spline function of order P for covariate x, with K knots at $k_1 < \ldots < k_K$. The spline function, $s(x)$, can be written with no continuity restrictions as follows

$$s(x) = \sum_{j=0}^{P} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{iP}(x - k_i)_+^P$$

Note the use of the "+" notation, where

$$u_+ = \begin{cases} u & \text{if } u > 0 \\ 0 & \text{if } u \leq 0 \end{cases}$$

Notice that the presence of a $\beta_{iP}(x - k_i)_+^P$ term allows a discontinuity at knot k_i for $s_{(j)}$, and its absence forces the continuity of $s_{(P)}$ at k_i.

Some type of splines are piecewise, linear, cuadratic, cubic, etc. Cubic splines ($P = 3$) are the most common type of spline used in practice. Higher degree polynomials are generally not needed, because if there were a complicated shape between knots (for example, with more than turning points), then further knots could be added rather than fitting a higher degree polynomial.

A cubic splines can be written as

$$s(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{i3}(x - k_i)_+^3$$

Thus the number of parameters in a standard regression spline model is $K + 4$. The fitted function is forced to be continuous and to have continuous 1^{st} and

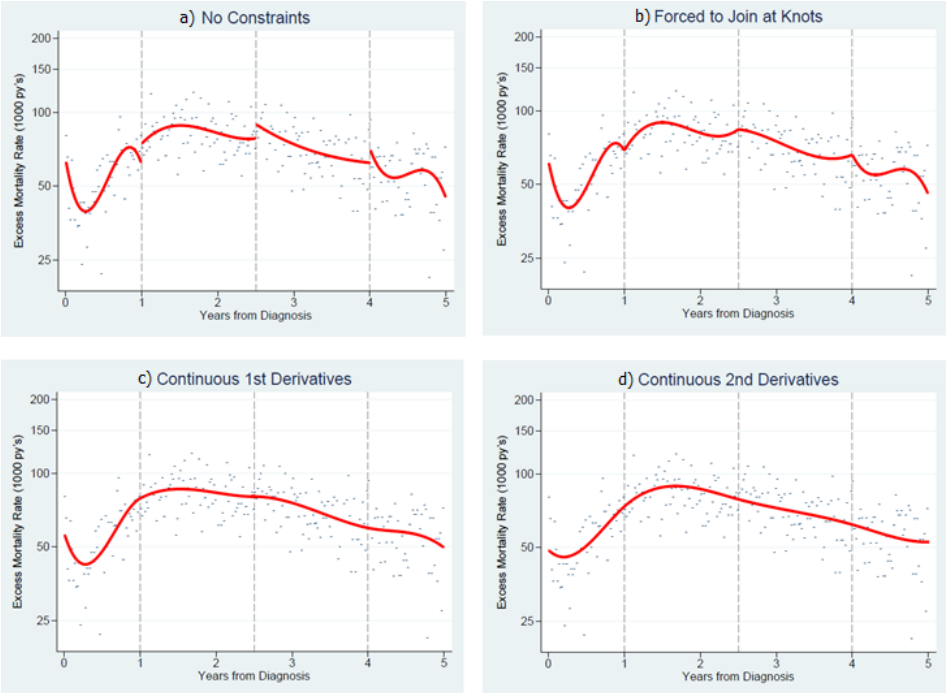

Figure 4.1: Example of using cubic spline functions with increasingly stringent continuity restrictions. Source: Royston and Lambert (2011).

2^{nd} derivatives. To understand this better, an illustrative example can be seen.

Figure 4.1(a) shows the fit of a cubic spline function with no constraints- that is, no continuity restrictions. The vertical lines show the locations of three knots (at 1, 2.5 and 4 years) that divide the time scale into four intervals. Within each interval, a cubic polynomial function has been fit. This is clearly not a sensible approach because in some intervals there is evidence of overfitting. If the function changes smoothly over time, it is needed that the estimated functions join at the knots.

This reason introduces the first continuity restriction, that can be seen in Figure 4.1(b). It is still cubic polynomials fitting within each interval, but the functions are forced to join at the knot locations. As you can see, there is a slight improvement: the function is now continuous. However, the function has a number of kinks and sharp turning points. By using a different number or different locations of the knots, the kinks and turning points are likely to occur in different places, and for that reason a second continuity restriction is introduced.

Now the first derivative of the spline functions is also force to agree at the knots.

Remember the first derivative is the gradient of the function, and thus one can hope to remove the kinks and bumps. The fitted function with this continuity restriction can be seen in Figure 4.1(c). This function is much better and is clearly smoother. However, suppose that the vertical lines showing the knot positions were removed: you could probably guess where at least two of the knots were located. This is not desirable, and for this reason a final continuity restriction is introduced.

As well as forcing the function and the first derivatives to agree at the knots, the final continuity restriction is to force the second derivative to agree at the knots. The second derivative is the rate of change in the gradient, and thus the function should become even smoother. You can see this in Figure 4.1(d), where the fitted function is much smoother and truly does look continuous.

The **restricted cubic splines** (Durrleman and Simon, 1989) are a simple extension of the cubic splines defined above. The difference is that the function is forced (restricted) to be linear before the first knot and after the last knot. The first and last knots are known as the boundary knots. For this study the boundary knots are the minimum and maximum of the uncensored survival times. However, knot location is usually not particularly crucial. All you need to fit a restricted cubic spline function for a covariate x is to include new variables in the linear predictor. The new variables are transformations of x. Let $s(x)$ be the restricted cubic spline function. Let's consider $K-2$ interior knots $k_1, \ldots k_{K-2}$ and also two boundary knots, $k_{\min}$ and $k_{\max}$, then $s(x)$ can be written as a function of parameters γ and some newly created variables $r_1, \ldots, r_{K-1}$, giving

$$s(x) = \gamma_0 + \gamma_1 r_1 + \gamma_2 r_2 + \ldots + \gamma_{K-1} r_{K-1}$$

The derived variables r_j are calculated as follows

$$\begin{aligned} r_1 = &\quad x \\ r_j = &\quad (x-k_j)_+^3 - h_j(x-k_{\min})_+^3 - (1-h_j)(x-k_{\max})_+^3 \end{aligned}$$

for $x = \ln(t)$ and $j = 2, \ldots, K-1$,

$$h_j = \frac{k_{\max} - k_j}{k_{\max} - k_{\min}} \tag{4.8}$$

For algebraic details see Appendix B of Royston and Parmar (2002).

4.6.3 Flexible Parametric Models: Incorporating Splines

As the model is on the log cumulative hazard scale, model (4.7) can be written as

$$\ln[H(t|\boldsymbol{x}_i)] = \ln[H_0(t)] + \boldsymbol{x}_i\beta$$

with $\ln[H_0(t)] = \gamma_0 + \gamma_1 \ln(t)$.

A restricted cubic spline (rcs) function of $\ln(t)$, with knots, k_0, can be written, $s(\ln(t)|\gamma, k_0)$.

$$\ln[H(t|\boldsymbol{x}_i)] = \eta_i = s(\ln(t)|\gamma, k_0) + \boldsymbol{x}_i\beta \tag{4.9}$$

For example, with 3 knots:

$$\ln[H(t|\boldsymbol{x}_i)] = \eta_i = \gamma_0 + \gamma_1 r_{1i} + \gamma_2 r_{2i} + \boldsymbol{x}_i\beta$$

Expression (4.9) can be transform to the survival and hazard scales:

$$S(t|\boldsymbol{x}_i) = \exp(-\exp(\eta_i)) \qquad h(t|\boldsymbol{x}_i) = \frac{ds(\ln(t)|\gamma, k_0)}{dt}\exp(\eta_i)$$

The hazard function involves the derivatives of the restricted cubic splines functions. However, these are easy to calculate,

$$s'(x) = \gamma_1 r'_1 + \gamma_2 r'_2 + \ldots + \gamma_{K-1} r'_{K-1}$$

Where

$$\begin{aligned} z'_1 &= 1 \\ z'_j &= 3(x-k_j)^2_+ - 3\lambda_j(x-k_{\min})^2_+ - 3(1-\lambda_j)(x-k_{\max})^2_+ \end{aligned}$$

When choosing the location of the knots for the restricted cubic splines, it is useful to have some default locations. As it has been mentioned, a sensible choice for the boundary knots $k_{\min}$, $k_{\max}$ is the smallest and largest uncensored log survival-times. Fortunately, optimal knot positioning does not appear to be critical for a good fit.

Royston and Parmar (2002) suggested knot positions based on empirical centiles of the distribution of log time, as given in the following table:

Table 4.1: Position of internal knots for modelling the baseline distribution function in RP models. Knots are positions on the distribution of uncensored log event-times.

Knots	df	Centiles
1	2	50
2	3	33, 67
3	4	25, 50, 75
4	5	20, 40, 60, 80
5	6	17, 33, 50, 67, 83
6	7	14, 29, 43, 57, 71, 86
7	8	12.5, 25, 37.5, 50, 62.5, 75, 87.5
8	9	11.1, 22.2, 33.3, 44.4, 55.6, 66.7, 77.8, 88.9
9	10	10, 20, 30, 40, 50, 60, 70, 80, 90

The positions are recommended by Durrleman and Simon (1989) for rcs.

In many applications, models with more than 3 knots (that is, with more than 4 degrees of freedom) are not recommended, because the resulting curves are potentially unstable.

4.6.4 Likelihood function and parameter estimation

Suppose that the sample comprises n independent observations $\{t_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, \ldots, n$, where δ_i is 0 for right-censored observation and 1 for an observed event. Let the likelihood for the ith observation be L_i, so that the likelihood for the whole sample is $\prod_{i=1}^{n} L_i$.

The contribution for the ith individual to the log likelihood function for a parametric survival model is given by

$$\ln L_i = \delta_i \ln h(t_i) + \ln S(t_i); \tag{4.10}$$

with late entry at t_{0i}, it becomes $\delta_i \ln h(t_i) + \ln S(t_i) - \ln S(t_{0i})$. Let $\eta_i = s(\ln t_i; \gamma) + \boldsymbol{x}_i\beta$ and its first derivative be $\eta' = d\eta_i/dt_i = ds(\ln t_i; \gamma)/dt_i = t_i^{-1} ds(\ln t_i; \gamma)/d(\ln t_i)$. In detail, (4.10) becomes the following for PH models,

$$Li = \begin{cases} \eta_i' \exp(\eta_i - \exp \eta_i) & \text{for an observed event,} \\ \exp(-\exp \eta_i) & \text{for a censored observation} \end{cases}$$

The expression for L_i is the density function at t_i, for observed events and the estimated survival probability at t_i for censored observations. Also

$$\begin{aligned}\frac{ds(\ln t_i;\gamma)}{d(\ln t_i)} &= \gamma_1 + \sum_{j=2}^{K-2} \frac{dr_j(\ln t_i)}{d(\ln t_i)} \\ &= \gamma_1 + \sum_{j=2}^{K-2} \gamma_j \{3(\ln t_i - k_j)_+^2 - 3\lambda_j(\ln t_i - k_{\min})^2 \\ &\quad - 3(1-\lambda_j)(\ln t_i - k_{\max})_+^2\}\end{aligned}$$

4.6.5 Comparing models

A useful criterion of model fit is the Akaike information criterion (AIC), defined as the deviance (minus twice the maximized log likelihood) plus $2m$, where m is the dimension of the model (that is, the number of fitted parameters). It can be used the AIC for comparing models with a different number of knots when using splines. The candidate model with the lowest AIC may be preferred. These models do not have to be nested.

An alternative to the AIC is the Bayes information criterion (BIC), which is the deviance penalized by adding $m \log n$, where m is the model dimension and n is the sample size. In survival analysis, n is interpreted as the number of events rather than the number of individuals. The model that minimizes the BIC among a set of candidates is said to be "best" in the sense that the BIC asymptotically selects the true model, provided that model is one of the candidate models.

Because parametric models are fit by maximum likelihood, AIC and BIC values can be used. Therefore, the comparison between parametric models is quite easy. The parameters of a Cox model, however, are estimated by maximum partial likelihood and the AIC or BIC values for this models are not comparable with those from parametric models, i.e., you can compare different Cox models on the same dataset using partial-likelihood versions of AIC or BIC, but you cannot compare a Cox model with a parametric model.

However, in the interest of parsimony and of reducing over-fitting, this criterion should not be applied mechanically.

4.7 Stratification in Flexible Parametric PH models

In this study a stratified Cox Model was fitted, and for that reason, it is necessary also to stratify in the Flexible Parametric model to enable a proper comparison. You also need to know that this section within the study is absolutely new, because it is not developed before, or at least is not documented yet.

Let Z_j, $j = 1, \ldots, D$ be the variables that do not satisfy the proportional hazards assumption. In this study, each of these variables are categorical, with d_j categories each, then these covariates can be defined as

$$\begin{aligned} Z_{1l_1}, &\quad l_1 = 1, \ldots, d_1 \\ Z_{2l_2}, &\quad l_2 = 1, \ldots, d_2 \\ \vdots &\quad \vdots \\ Z_{Dl_D}, &\quad l_D = 1, \ldots, d_D \end{aligned}$$

As in the Cox Model, in the Flexible Parametric PH model, if one or more variables do not satisfy the proportional hazard assumption, they can be considered as stratum variables.

Note that the terms *stratum variables* and *stratum* could be confused. To be clear, stratum variables can be for example Sex (Male, Female) and Center (Center$_1$, Center$_2$) while different combinations of these variables generate each stratum: "Male-Center$_1$", "Male-Center$_2$", "Female-Center$_1$", "Female-Center$_2$". In this example, there are 2 stratum variables (2 categories each) and for that reason there are 4 strata.

Due to stratification appears then a new set of splines for each stratum variable Z_j, which allows to have different baseline curves for every stratum (combinations of these variables). This new set of splines is defined by

$$\sum_{j=1}^{D} s(\ln(t)|\delta_{jl_j}, k_j) \tag{4.11}$$

Where k_j is the set of knots corresponding to every stratum variable. Incorporating the expression (4.11) in the model (4.9):

$$\ln H_{(Z_{1l_1},\ldots,Z_{Dl_D})}(t|\boldsymbol{x}_i) = s(\ln(t)|\gamma, k_0) + \sum_{j=1}^{D} s(\ln(t)|\delta_{jl_j}, k_j) + \boldsymbol{x}_i\beta$$

We can define a "reference stratum", which is the one that considers the category of each stratum variable with which the common spline is made. Moreover, if the reference stratum is considered, the model is reduced to

$$\ln H_{(Z_{1l_1},\ldots,Z_{Dl_D})}(t|\boldsymbol{x}_i) = s(\ln(t)|\gamma, k_0) + \boldsymbol{x}_i\beta$$

Where $s(\ln(t)|\gamma, k_0)$ is a common spline for all the possible strata.

It is important here to note that since this methodology is new, there is no goodness of fit test developed or at least documented.

4.8 Software Implementation

All analyzes were performed using two softwares: `R` and mainly `Stata`.

For the definitions of the different versions of index, their summary and the fit of a Cox model with each of them, the statistical software `R` Project was used. The semi-parametric estimation and their residual analysis were performed through the package `survival` (Jackson, 2016).

To get the baseline estimation of the Cox Model, command `stcox` from `Stata` was used (Corporation, 2003). With option `basesurv(newvar)` in the model, estimates of the baseline survival function can be obtained. When the model is fitted with the `strata()` option, estimates of the baseline functions for each stratum are obtained.

Mathematically, the baseline hazard contribution $h_i = (1-\alpha_i)$ (Kalbfleisch and Prentice, 2002) is defined at every analytic time t_i, at which a failure occurs and is undefined (or, if you prefer, 0) at other times. Stata stores h_i in observations where a failure occurred and missing values in the other observations.

The baseline survivor function $S_0(t)$ is defined at all values of t: its estimate changes its value when failures occur and, at times when no failures occur, the estimated $S_0(t)$ is equal to its value at time time of the last failure.

For the Flexible Parametric models two packages are implemented in `R`: `flexsurv` (Jackson, 2016) maintained by Christopher Jackson, and `rstpm2` (Clements and Liu, 2016), maintained by Mark Clements. The latter tries to emulate the command `stpm2` implemented in `Stata`, but has some problems. For example, it does not work with left-truncated data. Concerning the first, `flexsurv` presents some convergence problems, due to initial values definition, and actually, the investigation group working with the maintainer are trying to solve this issue.

According to this, Flexible Parametric models were fitted with `Stata`, with command `stpm2`. You can see the principal explanation for this command in Appendix B. Syntax used here was:

`stpm` *covariates*, `tvc`(*stratum variables*) `df() dftvc() scale(hazards)`

In option `tvc` you can specify all the stratum variables, in option `df()` you must specify the degrees of freedom for the main basal spline (corresponding to the knots + 1, for example for 3 knots you need to put `df(4)`), in option `dftvc()` you must specify the knots (+1) for every stratum variables (the number can be different (`dftvc(2 3 4)` for example) or the same per each of them (`dftvc(2)`); and for this study, the scale used was hazards.

As a brief history, Patrick Royston wrote `stpm` in 2001 (Royston, 2001). Chris Nelson extended the methodology in `stpm` to relative survival (Nelson et al., 2007) in `strsrcs`. In 2009, Paul Lambert and Patrick Royston wrote `stpm2` to improve the modelling of time-dependent effects (before of this, they tended to be over parametrised) (Lambert and Royston, 2009). The combination of these methods for standard and relative survival make it easier to obtain useful predictions. Also `stpm2` is much faster than `stpm`, especially with large datasets.

An illustrative example is presented for educational purposes only, using a database other than the interest for this work data. This example appears in Royston and Lambert (2011). Data available in `http://www.pauldickman.com/survival/`.

Example: Rotterdam breast cancer data

Sauerbrei et al. (2007) analyzed data from 2982 patients with primary breast cancer whose records were included in the tumour bank at Rotterdam, The Netherlands. Follow-up time ranged from 1 to 231 months (median: 107 months). They analysed relapse-free survival, defined as the time from primary surgery to disease recurrence or death from breast cancer. Times to death from other causes were treated as censored. For this analysis, they censored the event and censoring times at 120 months (10 years). With the relapse-free survival outcome, 1477 events were observed in the interval up to 120 months.

Figure 4.2 shows the estimated survival for the data.

The survival curves indicate a median time to event of about eight years. The Kaplan-Meier curve shows a slight downturn after about nine years, which is not reflected in the survival curve from `stpm2`.

To show how spline functions are computed, the example illustrate in Figure 4.2 will be deconstructed.

The number of interior knots is m=2. The smallest and largest uncensored events time in the Rotterdam breast cancer data are 0.104 and 9.99 years, so the

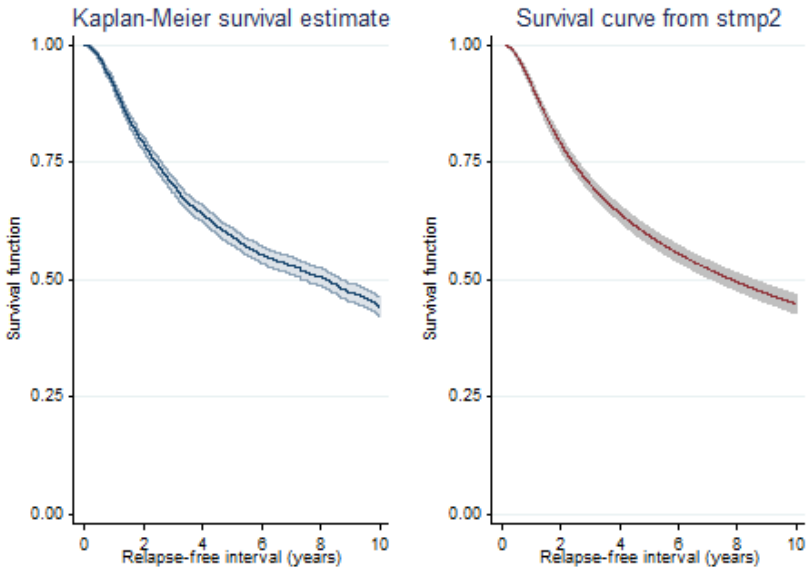

Figure 4.2: Rotterdam breast cancer data. Survival curves are estimated by the Kaplan-Meier method (sts graph) and stpm2. Source: Royston and Lambert (2011)

boundary knots are $k_{\min} = \ln(0.104) = -2.26$ and $k_{\max} = \ln(9.99) = 2.30$. The 33rd and 67th centiles of the uncensored event times are 1.577 and 3.571 years, so $k_1 = 0.46$, $k_2 = 1.27$. From this, and replacing in the formula (4.8) are obtained $\gamma_1 = (k_{\max} - k_1)/(k_{\max} - k_{\min}) = 0.404$ and $\gamma_2 = (k_{\max} - k_2)/(k_{\max} - k_{\min}) = 0.226$. The crude Stata code to calculate the spline basis functions $r_1(\ln t)$ and $r_2(\ln t)$ is prsented below:

```
scalar kmin = -2.26
scalar kmax = 2.30
scalar k1 = 0.46
scalar k2 = 1.27
scalar lambda1 = 0.404
scalar lambda2 = 0.226
generate lnt = ln(_t)
generate r0 = cond(lnt > kmin, (lnt-kmin)^3, 0)
generate r3 = cond(lnt > kmax, (lnt-kmin)^3, 0)
generate r1 = cond(lnt > k1, (lnt-k1)^3, 0) - lambda1*r0 - (1-lambda1)*r3
generate r2 = cond(lnt > k2, (lnt-k2)^3, 0) - lambda2*r0 - (1-lambda2)*r3
```

The maximum likelihood estimates of the parameters γ_0, γ_1, γ_2 and γ_3 are -1.7864, 2.5114, 0.1558 and -0.0392 respectively. If the command `stpm2` (The

RP function in `Stata`) is applied with the following syntax

```
stpm2, df(3) scale(hazard) nolog noorthog
```

the same estimations are obtained.

Table 4.2: Example: Flexible Parametric model fitted with Rotterdam breast cancer data.

	coef	**S.E.**	**Z**	**p-value**	**95%CI**
rcs_1	2.5114	0.17	14.92	<0.001	2.18 ; 2.84
rcs_2	0.1558	0.04	3.60	<0.001	0.07 ; 0.24
rcs_3	-0.0392	0.05	-0.80	0.426	-0.14 ; 0.06
cons	-1.7864	0.10	18.66	0.001	-1.97 ; -1.60

The estimated log cumulative-hazard function can be calculated by

```
generate lnH = -1.7864 + 2.5114 * lnt + 0.1558 * r1 - 0.0392 * r2
```

But also, the syntax `predict lnH, xb` gives the log cumulative hazard function directly.

Use of the `noorthog` option prevents orthogonalization of the spline functions to ensure that the estimates of γ_0, γ_1, γ_2 and γ_3 agree with those calculated from first principles. In practice, it is preferable not to apply the `noorthog` option, because it tends to make the estimates of the spline model parameters slower to compute and numerically less stable.

Spline basis functions: problem and solution

The basis functions are highly correlated, which can sometimes cause `stmp2` difficulties in estimating the parameters of the model quickly and reliably. For example, suppose that we have 100 observations of $\ln t$ equally spaced on the interval [0.01,1]. We place $m = 3$ knots at 0.25, 0.5 and 0.75 and use the auxiliary command `frac_spl` to compute the untransformed spline basis functions. These turn out to have the following correlation matrix:

	$\ln t$			
$\ln t$	1.000			
$v_1(\ln t)$	-0.922	1.000		
$v_2(\ln t)$	-0.942	0.997	1.000	
$v_3(\ln t)$	-0.966	0.987	0.996	1.000

A simple solution to this problem is to transform the basis functions linearly so that, after transformation, the correlations are zero. Gram-Schmidt orthogonalization is one way to do this, and it is available through the Stata command

`orthog`. Orthogonalization is available in `stpm2`, although the option `noorthog` is provided for compatibility with earlier versions and for pedagogic reasons (Royston, 2004).

Chapter 5

Variables and Population under Study

5.1 Variables of study

In Chapter 6 the results of analyses are presented. For this study, initially a stratified Cox model was fitted (stratified by sex, center and age of recruitment) considering as the main variable the Adherence to the Mediterranean Diet (mdscore). Other variables included in the models were: smoke status, body mass index, physical activity, energy intake, waist circumference and educacional level. Then, a Flexible Parametric model was used, considering the same strata and the same variables of interest. Variables are explained below.

As was mentioned above, the main variable of interest is the mdscore, for more information read Section 3.2. The stratum variables are:

- **Sex** classified according to male or female.
- **Center** according to the 5 centers mentioned before: Asturias, Gipuzkoa, Navarra, Murcia and Granada.
- **Age at recruitment** classified in four categories: < 40, $[40, 50)$, $[50, 60)$, ≥ 60 years.

Other variables, potential confounders, considered in the Cox and Flexible Parametric models are (in blue color the reference category for each of the categorical variables):

- **Smoke Status** summarized in ten categories: never smoker, current smoker of 1-15 cigarettes/day, current smoker of 16-25 cigarettes/day, current smoker of $\geq$26 cigarettes/day, former smoker ($<$10 years since cessation), former smoker (10-20 years since cessation), former smoker ($>$20

years since cessation), smoker of cigar/pipe ocassional, current or former smoker (information missing), unknown status of smoke intensity.

- **Body Mass Index (BMI)** calculated as kg/m^2, was used to classify the subjects into four categories: <18.5 (underweight) , $[18.5, 25)$ (normal weight), $[25, 30)$ (overweight), ≥ 30 (obesity) (Consultation, 2011). Reference category is the union of the two first categories, that is, < 25 kg/m^2 (under and normal weight) due to only few individuals in the first (only 40 individuals, of which 38 have mild thinness $[17, 18.5)$ kg/m^2, 2 moderate thinness $[16, 17)$ kg/m^2 and nobody have severe thinness < 16 kg/m^2).

- **Physical Activity**. Different domains of physical activity gathered: occupational physical activity, housekeeping activities, and leisure-time physical activity taking into account seasonal variation (Haftenberger et al., 2002). Using this information, subjects were classified according to a synthetic index into four categories: inactive, moderately inactive, moderately active and active.

- **Energy Intake**, corresponding to energy intake (kilocalories/day). This variable was centered to the mean (this process is automatic for the Cox Model in `Stata`, but not for Flexible Parametric Model). So, the reference here is the mean. Adjustment for total energy is usually appropriate in epidemiologic studies to control for confounding, reduce extraneous variation, and estimate the effect of dietary interventions (Willet et al., 1997).

- **Waist Circumference**, to assess abdominal obesity, waist circumference was used as a dichotomous variable, according to Consultation (2011): men < 102 cm and women < 88 cm were considered as normal, whereas men ≥ 102 cm and women ≥ 88 cm were considered as having abdominal obesity.

- **Educational Level** classified according to six categories: none, primary school, technical or professional training, secondary school, university degree and unknown.

5.2 Study Population

The data used in this study correspond to the EPIC-Spain cohort and was recruited between 1992 and 1996 from 3 regions in the north (Asturias, Gipuzkoa and Navarra) and 2 regions in the south (Granada and Murcia) of Spain.

To assess the quality of the reported dietary data, the ratio between the energy intake and the estimated energy requirement was calculated for each participant. Individuals with values higher or lower than the mean plus or minus

three standard deviations of the log-transformed ratio were considered as having an implausible diet.

From the initial sample of 41437 individuals with 3676 deaths- are excluded: 8 individuals for exclusion in the study (people who have decided to leave the study) and 238 for having implausible values in regards to diet. Therefore, in total have been excluded from the original sample 246 individuals among whom there were 30 deaths. Thus the evaluable population is 41191 individuals.

These 41191 individuals are healthy volunteers (25612 females), aged 29-69 years, of different social and educational levels, and were recruited mostly from among blood donors (about 60%). The study population covered a diverse range of socioeconomic levels and different geographic areas.

The mean of follow-up was 18.5 years, with 3646 deaths registered.

A description of the events (deaths) and the total of individuals are presented in the table below.

Table 5.1: Baseline characteristics and the number of events of the study population

Cohort Characteristics	**n (%)**	**Total Events (%)**
Total	41191	3646
Sex		
Men	15579 (37.82)	2139 (58.67)
Women	25612 (62.18)	1507 (41.33)
Center		
Asturias	8514 (20.67)	717 (19.67)
Granada	7763 (18.85)	638 (17.50)
Murcia	8458 (20.53)	756 (20.74)
Navarra	8065 (19.58)	742 (20.35)
Gipuzkoa	8391 (20.37)	793 (21.75)
Age at Recruitment (years)		
<40	5423 (13.17)	120 (3.29)
[40,50)	17538 (42.58)	837 (22.96)
[50,60)	13082 (31.76)	1425 (39.08)
≥60	5148 (12.50)	1264 (34.67)
BMI (kg/m^2)		
< 25	9106 (22.11)	565 (15.50)
[25,30)	19716 (47.86)	1640 (44.98)

Table 5.1: (continued)

Cohort Characteristics	**n (%)**	**Total Events (%)**
≥30	12369 (30.03)	1441 (39.52)
Waist circumference		
Normal	23724 (57.60)	1703 (46.71)
Abdominal Obesity	17467 (42.40)	1943 (53.29)
Physical Activity		
Inactive	5437 (13.20)	642 (17.61)
Moderately inactive	8614 (20.91)	981 (26.91)
Moderately active	23550 (57.17)	1748 (47.94)
Active	3590 (8.72)	275 (7.54)
Educational Level		
None	14176 (34.42)	1564 (42.90)
Primary	15938 (38.69)	1261 (34.59)
Technical/professional	3391 (8.23)	243 (6.66)
Secondary	2663 (6.47)	224 (6.14)
University	4744 (11.52)	319 (8.75)
Unknown	279 (0.68)	35 (0.96)
Smoke Status		
Never	22827 (55.42)	1623 (44.51)
Current 1-15 c/d	4671 (11.34)	343 (9.41)
Current 16-25 c/d	2657 (6.45)	316 (8.67)
Current ≥ 26 c/d	1049 (2.55)	217 (5.95)
Former <10 years	4199 (10.19)	427 (11.71)
Former 11-20 years	2116 (5.14)	195 (5.35)
Former >20 years	882 (2.14)	100 (2.74)
Cigar/Pipe occasional	2649 (6.43)	410 (11.25)
Current/Former, missing	119 (0.29)	14 (0.38)
Unknown	22 (0.05)	1 (0.03)
Energy Intake (kcal/day)	2137.95 (688.95)*	

c/d: cigarettes per day
* values are mean and standard deviation in this case.

It is also important to see the distribution of the individuals per every stratum, you can see this in the table below

Table 5.2: Individuals and events per stratum.

	Sex			
	Men		**Women**	
	Events	**Total**	**Events**	**Total**
Asturias				
< 40 years	1	70	28	1180
$[40, 50)$ years	124	1613	75	2267
$[50, 60)$ years	149	949	118	1460
≥ 60 years	143	443	79	532
Granada				
< 40 years	5	94	16	1080
$[40, 50)$ years	41	723	75	2186
$[50, 60)$ years	84	607	125	1831
$[50, 60)$ years	117	353	175	889
Murcia				
< 40 years	10	122	22	1240
$[40, 50)$ years	77	1238	76	2192
$[50, 60)$ years	168	936	128	1745
$[50, 60)$ years	136	377	139	608
Navarra				
< 40 years	0	15	19	731
$[40, 50)$ years	118	1970	58	1688
$[50, 60)$ years	249	1431	85	1299
$[50, 60)$ years	154	487	59	444
Gipuzkoa				
< 40 years	3	76	16	815
$[40, 50)$ years	135	1894	58	1767
$[50, 60)$ years	230	1613	89	1211
$[50, 60)$ years	195	568	67	447

In the next Chapter (Chapter 6) all graphics are presented for the center Granada, because it is the center that has events in all the sub-categories, and also is more "Mediterranean".

Chapter 6

Results

The results of the present study have been divided in three sections. The first section is about choosing the best version of the Adherence to the Mediterranean Diet Score. The second section is the description of the stratified Cox Model and the Flexible Parametric PH Model fitted. The third section is about the comparison between the Flexible Parametric PH Model and the stratified Cox Model.

6.1 Adherence to the Mediterranean Diet score (mdscore)

A description of the different versions of the Mediterranean Diet Adherence score is given in Section 3.2 and the `R` code is presented in Appendix A. In Table 6.1 the range for every score can be seen, the minimum value represents the lowest Adherence to the MD and the maximum the highest Adherence. Also, the mean, standard deviation and their quartiles are presented.

Table 6.1: Summary for different versions of Mediterranean Diet Adherence score.

Score	**min; max**	**mean (sd)**	$\mathbf{Q_1}$	$\mathbf{Q_2}$	$\mathbf{Q_3}$
mdscore_{sum}	-0.79 ; 2.78	0.91 (0.73)	0.24	0.67	1.65
mdscore_{cdf}	-0.91 ; 6.06	2.63 (0.92)	2.00	2.63	3.26
mdscore_{sd}	-9 ; 12	0.25 (2.08)	-1	0	2
mdscore_{ter}	0 ; 17	8.67 (2.58)	7	9	10

Choosing the best version of the Adherence to the Mediterranean Diet Score

Once created and described the different indices, different Cox models have been fitted. The model 0 (m_0) corresponds to a Cox model stratified by center, sex and age at recruitment, considers all the covariates of the study except mdscore (that is to say, considers smoke status, BMI, physical activity, energy intake, waist circumference and educational level) for every individual in the study. The following models add to these covariates a different version of the Adherence to the Mediterranean Diet score in each case.

In Table 6.2 the Akaike Criterion (AIC) and Bayesian Criterion (BIC) are presented in order to select the model, and the p-values associated with the Partial Likelihood Ratio Test, upon comparing each model to the initial model m_0.

Table 6.2: Comparison of Cox models with different versions of mdscore.

Model	**df**	**AIC**	**BIC**	**PLRT**
m_0	21	46781.35	46962.50	
m_1 (with mdscore$_{sum}$)	22	46771.13	46960.90	< 0.001
m_{2a} (with mdscore$_{cdf}$)	22	46749.96	46939.73	< 0.001
m_{2b} (with mdscore$_{cdf}$*)	24	46752.77	46959.80	< 0.001
m_3 (with mdscore$_{sd}$)	22	46747.05	46936.82	< 0.001
m_4 (mdscore$_{ter}$)	22	46760.93	46950.70	< 0.001

* considering mdscore$_{cdf}$ divided into quartiles.

It can be seen that the smallest values of AIC and BIC presented are for models m_{2a} and m_3. It can also be seen in Table 6.3 that the hazard ratios associated with this variable of interest are the same (for complete output for these models see Appendix C).

Table 6.3: Hazard ratio and confidence interval for different versions of mdscore from Cox Model.

Score	**HR (95% IC)**	**p-value**
mdscore$_{sum}$	0.92 (0.87 − 0.96)	< 0.001
mdscore$_{cdf}$ (ref: 1st quartile)		
2nd quartile	0.93 (0.85 − 1.02)	0.129
3rd quartile	0.84 (0.76 − 0.92)	< 0.001
4th quartile	0.76 (0.69 − 0.84)	< 0.001
mdscore$_{cdf}$	0.90 (0.86 − 0.93)	< 0.001
mdscore$_{sd}$	0.90 (0.86 − 0.93)	< 0.001
mdscore$_{ter}$	0.97 (0.96 − 0.98)	< 0.001

It has finally been decided to consider for both models (stratified Cox model and Flexible Parametric PH model) the version of index called mdscore_{cdf}, which henceforth will be called only mdscore to facilitate notation. The decision has been taken, seeing that this variable proposes greater ease of use, to be treated both continuously and categorically. Compared to mdscore_{sd} it also gives greater ease of extrapolation, in case you want to work with data from any center in Spain or abroad.

6.2 Models fitted

In this section the Cox model fitted and its interpretation is given. Same work is done with the Flexible Parametric PH Model approach, in order to have a first impression to make some comparison.

6.2.1 Semi-Parametric Estimation of the Survival Function

A Cox model stratified by center, sex and age of recruitment, considering the main variable of interest, mdscore, as a categorical variable (divided into quartiles) has been fitted. Also all other covariates presented in Chapter 5 has been included in this model. The consideration of mdscore as a categorical variable is due to easy interpretation (Model fitted with mdscore as a continuous variable can be seen in Appendix C).

The following table shows the output of this model:

Table 6.4: Cox Model fitted with mdscore as a categorical variable.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore (ref: 1st quartile)						
2nd quartile	-0.07	0.93	0.04	-1.52	0.129	0.85 − 1.02
3rd quartile	-0.18	0.84	0.04	-3.63	<0.001	0.76 − 0.92
4th quartile	-0.27	0.76	0.04	-5.47	<0.001	0.69 − 0.84
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.95	<0.001	1.37 − 1.76
Current 16-25 c/d	0.84	2.31	0.16	12.46	<0.001	2.03 − 2.64
Current ≥ 26 c/d	1.42	4.14	0.33	17.95	<0.001	3.55 − 4.84
Former < 10 years	0.42	1.52	0.09	6.86	<0.001	1.35 − 1.71
Former 11-20 years	0.23	1.26	0.10	2.84	0.005	1.07 − 1.47
Former > 20 years	0.06	1.07	0.11	0.59	0.554	0.86 − 1.31
Cigar/Pipe, occasional	0.43	1.54	0.10	6.62	<0.001	1.35 − 1.74
Current/Former, missing	0.38	1.46	0.39	1.40	0.163	0.86 − 2.47
Unknown	-0.44	0.65	0.65	-0.44	0.662	0.09 − 4.60

Table 6.4: (continued)

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.43	0.001	0.75 − 0.93
≥ 30	0.03	1.03	0.07	0.51	0.610	0.91 − 1.17
Physical Activity						
Inactive	0.05	1.05	0.06	0.98	0.328	0.95 − 1.17
Moderately inactive	0.07	1.07	0.05	1.59	0.112	0.98 − 1.17
Active	-0.03	0.97	0.06	-0.48	0.634	0.85 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-2.76	0.006	1.00 − 1.00
Waist Circumference						
Abdominal Obesity	0.11	1.11	0.05	2.38	0.018	1.02 − 1.22
Educational Level						
None	0.02	1.02	0.04	0.44	0.662	0.94 − 1.10
Technical/Professional	-0.13	0.88	0.06	-1.77	0.077	0.77 − 1.01
Secondary	0.01	1.01	0.08	0.20	0.845	0.88 − 1.17
University	-0.17	0.84	0.06	-2.63	0.009	0.74 − 0.96
Unknown	0.19	1.21	0.21	1.12	0.262	0.87 − 1.70

c/d: cigarettes per day

In the above table, considering a level of significance of 0.05, you can see that the HR of the second quartile of the main variable (mdscore) is less than 1, but the difference is not that big to speak about significant difference from the reference category (which in this case is the first quartile). However, when talking about the third or fourth quartile there is significant difference from the baseline, that is to say there is a significant difference in the survival of individuals who best adhere to the Mediterranean Diet compared with the least adherence.

We note further that:

- For the third quartile, the sign of the coefficient is negative (-0.18), which means that individuals who are in the third quartile of adhesion (high adherence) to the Mediterranean Diet have a lower instantaneous risk of mortality compared with those of least adherence. The hazard ratio associated is equal to 0.84, this means that there is a 19% increased instantaneous risk of mortality in individuals with less adherence to the MD (1st quartile) compared with those who belong to the 3rd quartile of adhesion.
- For the fourth quartile, the sign of the coefficient is negative (-0.27), which means that individuals who are in the fourth quartile of adhesion (those most adhere) to the Mediterranean Diet have less instantaneous risk of mortality for which less adhere. The hazard ratio associated is equal to

0.76, this means that there is a 32% increased instantaneous risk of mortality in individuals with less adherence to the MD (1st quartile) compared with those with greater adherence (4th quartile).

It is also necessary to point out that the goodness of fit of this model has been verified. The assumption of proportional hazards was performed by the test proposed by Grambsch and Therneau (1994) and also through the graphical analysis of the Schoenfeld residuals, furthermore the influence of each individual (dfbeta residuals), and the overall fit (with deviance residuals) (for further information, see Appendix C) have been verified.

6.2.2 Flexible Parametric PH Model

The description of the Flexible Parametric PH model is given in Section 4.6. From the description, it is necessary to remember that there are many possible combinations of number of knots, for both the splines for the baseline and the splines associated with each stratum variable. The Table 6.5 presents the AIC and BIC criteria, for each combination performed. In each case it has been stratified by center, sex and age of recruitment. It has been fitted considering mdscore as the main variable of interest and considering all the other covariates used in the previous Cox model (smoke intensity, BMI, physical activity, energy, waist circumference and educational level).

Table 6.5: Knots combinations for Flexible Parametric PH Models.

Model	df	dftvc	AIC	BIC	df	Observation
1	2	2	11123.51	11503.06	44*	convergence achieved
2	3	2	11222.20	11567.24	40*	100 iterations
3	**4**	**2**	**11114.50**	**11502.66**	**45**	**convergence achieved**
4	5	2	11110.34	11507.14	47*	convergence achieved
5	2	3	11282.56	11705.23	49*	100 iterations
6	3	3	11283.78	11715.08	50*	100 iterations
7	4	3	11263.37	11720.55	53	100 iterations
8	5	3	11271.49	11728.66	53*	100 iterations
9	2	4	11291.69	11774.75	56*	100 iterations
10	3	4	11268.61	11760.29	57*	100 iterations
11	4	4	11279.01	11762.06	56*(1)	100 iterations
12	5	4	11276.21	11767.89	57*	100 iterations
13	2	5	11298.51	11824.69	61*	100 iterations
14	3	5	11294.84	11846.90	64*	100 iterations
15	4	5	11301.19	11861.88	65*	100 iterations
16	5	5	11285.49	11854.81	66*(2)	100 iterations

(1) rcs Age $\geq 60_4$ omitted because of collinearity.
(2) rcs Age $\geq 60_5$ omitted because of collinearity.
* degrees of freedom are incorrect in the output of Stata.

In black font you can see the selected model (model 3). Of the sixteen fitted models, only three of them achieved the convergence, the others had problems

to achieve the convergence due to the identification of initial values. When the autor of this study contacted Paul Lambert (one of the creators of this model), he recommended using the option `initstrata`*(varname)* available in `stpm2`, but still has not been documented. Not knowing how this works, it is left as a further analysis.

In addition, you can see the degrees of freedom that are obtained by running the command `estat ic` in `Stata` (this command also delivers AIC and BIC) and in 14 of the 16 models, they are incorrect. This is understandable when the model does not reach convergence, because the delivered results are approximations of real estimates. But this situation also occurs in models which reach the convergence (models 1 and 4, for example), for that reason there are some bugs in this command.

The selected model has the smallest BIC, achieved the convergence and has no problem in its output. The fitted model selected is presented in the table below

Table 6.6: Flexible Parametric PH Model fitted with mdscore as a categorical variable

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore (ref: 1st quartile)						
2nd quartile	-0.07	0.93	0.04	-1.45	0.148	0.85 − 1.02
3rd quartile	-0.17	0.85	0.04	-3.45	0.001	0.77 − 0.93
4th quartile	-0.25	0.78	0.04	-5.03	<0.001	0.71 − 0.86
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.95	<0.001	1.37 − 1.76
Current 16-25 c/d	0.83	2.30	0.15	12.47	<0.001	2.01 − 2.62
Current $\geq$ 26 c/d	1.43	4.17	0.32	18.23	<0.001	3.57 − 4.86
Former < 10 years	0.41	1.51	0.09	6.77	<0.001	1.34 − 1.70
Former 11-20 years	0.23	1.25	0.10	2.80	0.005	1.07 − 1.47
Former > 20 years	0.04	1.04	0.11	0.41	0.682	0.85 − 1.29
Cigar/Pipe, occasional	0.48	1.62	0.10	7.55	<0.001	1.43 − 1.83
Current/Former, missing	0.33	1.39	0.37	1.21	0.226	0.82 − 2.35
Unknown	-0.49	0.61	0.61	-0.49	0.625	0.09 − 4.36
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.46	0.001	0.75 − 0.92
$\geq$ 30	0.03	1.03	0.07	0.51	0.609	0.91 − 1.17
Physical Activity						
Inactive	0.07	1.07	0.06	1.30	0.195	0.97 − 1.19
Moderately inactive	0.07	1.07	0.05	1.53	0.125	0.98 − 1.17
Active	-0.03	0.97	0.06	-0.47	0.638	0.85 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-2.52	0.012	1.00 − 1.00

Table 6.6: (continued)

	coef	HR	S.E.	Z	p-value	95%CI
Waist Circumference						
Abdominal Obesity	0.10	1.10	0.05	2.16	0.031	1.01 − 1.20
Educational Level						
None	0.01	1.00	0.04	0.20	0.845	0.93 − 1.09
Technical/Professional	-0.11	0.90	0.06	-1.52	0.128	0.78 − 1.03
Secondary	0.01	1.01	0.08	0.20	0.841	0.88 − 1.17
University	0.19	0.83	0.05	-2.90	0.004	0.73 − 0.94
Unknown	0.14	1.15	0.20	0.81	0.419	0.82 − 1.61
rcs						
rcs_1	0.68	1.98	0.21	6.34	<0.001	1.60 − 2.45
rcs_2	0.00	1.00	0.03	-0.07	0.946	0.93 − 1.07
rcs_3	-0.02	0.98	0.01	-1.82	0.069	0.96 − 1.00
rcs_4	-0.02	0.98	0.01	-3.00	0.003	0.97 − 0.99
rcs Center						
rcs Granada_1	0.06	1.06	0.03	2.25	0.025	1.01 − 1.12
rcs Granada_2	-0.02	0.98	0.01	-1.92	0.054	0.97 − 1.00
rcs Murcia_1	0.09	1.09	0.03	3.14	0.002	1.03 − 1.15
rcs Murcia_2	-0.02	0.98	0.01	-3.16	0.002	0.96 − 0.99
rcs Navarra_1	0.13	1.14	0.03	4.32	<0.001	1.07 − 1.21
rcs Navarra_2	-0.03	0.97	0.01	-3.12	0.002	0.96 − 0.99
rcs Gipuzkoa_1	-0.01	0.99	0.03	-0.29	0.771	0.94 − 1.05
rcs Gipuzkoa_2	0.00	1.00	0.01	0.09	0.927	0.98 − 1.02
rcs Sex						
rcs Female_1	-0.34	0.71	0.02	-11.51	<0.001	0.67 − 0.76
rcs Female_2	-0.07	0.93	0.02	-3.88	<0.001	0.89 − 0.96
rcs Age						
rcs $[40, 50)_1$	0.08	1.08	0.11	0.75	0.456	0.88 − 1.32
rcs $[40, 50)_2$	-0.02	0.98	0.02	-0.82	0.410	0.95 − 1.02
rcs $[50, 60)_1$	0.12	1.13	0.13	1.09	0.275	0.91 − 1.41
rcs $[50, 60)_2$	-0.05	0.95	0.03	-1.79	0.074	0.90 − 1.01
rcs $\geq 60_1$	0.08	1.09	0.12	0.73	0.466	0.87 − 1.36
rcs $\geq 60_2$	-0.10	0.90	0.04	-2.26	0.024	0.82 − 0.99
constant	-2.48	0.08	0.01	-28.30	<0.001	0.07 − 0.10

c/d: cigarettes per day
rcs: restricted cubic spline

It is evident that the output of the Parametric Flexible PH Model presents new estimates (which did not have the Cox model) for restricted cubic splines. These estimates are important for the development of the model.

Like when fitting the Cox model above, it can be seen that the second quartile has an HR< 1 but is not significant when compared with the first quartile (p-value $= 0.148$) with respect to the hazards, using a level of significance of 0.05. However, the relationship is significant with the top two quartiles of adhesion (3rd quartile p-value $= 0.001$ and 4th quartile p-value < 0.001), that is to say greater Adherence to the Mediterranean Diet reduced risk of mortality, in fact:

- The sign of the coefficient is negative (-0.17) for the third quartile, this means that individuals with high adherence (in the 3rd quartile) have a lower instantaneous risk of mortality compared with those with lower Adherence (1st quartile). The hazard ratio associated is 0.85, this means that there is a 18% increased instantaneous risk of mortality in individuals with less Adherence to the MD (1st quartile) compared with those belong to the 3rd quartile of Adherence.
- The sign of the coefficient is negative also (-0.25) for the 4th quartile, which means that individuals who are in the fourth quartile of Adherence (those most adhere) to the Mediterranean Diet have less instantaneous risk of mortality for which less adhere. The hazard ratio associated is equal to 0.78, this means that there is a 28% increased instantaneous risk of mortality in individuals with less Adherence to the MD (1st quartile) compared with those with greater adherence (4th quartile).

6.3 Comparison between the Cox and the Flexible Parametric PH Model

6.3.1 Global comparison

As we have seen, both models agree fairly in terms of the hazard ratio and their respective confidence intervals of 95%. To make this even more clearly, a comparative table for both models is presented below:

Table 6.7: Comparison of hazard ratios between Cox and Flexible Parametric PH Model.

	Cox Model HR (95%CI)	**Flexible Parametric*** **HR (95%CI)**
mdscore (ref: 1st quartile)		
2nd quartile	0.93 (0.85 − 1.02)	0.93 (0.85 − 1.02)
3rd quartile	0.84 (0.76 − 0.92)	0.85 (0.77 − 0.93)
4th quartile	0.76 (0.69 − 0.84)	0.78 (0.71 − 0.86)
Smoke Status		
Current 1-15 c/d	1.55 (1.37 − 1.76)	1.55 (1.37 − 1.76)
Current 16-25 c/d	2.31 (2.03 − 2.64)	2.30 (2.01 − 2.62)

Table 6.7: (continued)

	Cox Model HR (95%CI)	**Flexible Parametric*** **HR (95%CI)**
Current $\geq$ 26 c/d	4.14 (3.55 − 4.84)	4.17 (3.57 − 4.86)
Former $<$ 10 years	1.52 (1.35 − 1.71)	1.51 (1.34 − 1.70)
Former 11-20 years	1.26 (1.07 − 1.47)	1.25 (1.07 − 1.47)
Former $>$ 20 years	1.07 (0.86 − 1.31)	1.04 (0.85 − 1.29)
Cigar/Pipe, ocassional	1.54 (1.35 − 1.74)	1.62 (1.43 − 1.83)
Current/Former, missing	1.46 (0.86 − 2.47)	1.39 (0.82 − 2.35)
Unknown	0.65 (0.09 − 4.60)	0.61 (0.09 − 4.36)
BMI (kg/m^2)		
$[25, 30)$	0.83 (0.75 − 0.93)	0.83 (0.75 − 0.92)
$\geq$ 30	1.03 (0.91 − 1.17)	1.03 (0.91 − 1.17)
Physical Activity		
Inactive	1.05 (0.95 − 1.17)	1.07 (0.97 − 1.19)
Moderately inactive	1.07 (0.98 − 1.17)	1.07 (0.98 − 1.17)
Active	0.97 (0.85 − 1.10)	0.97 (0.85 − 1.10)
Energy Intake (kcal/day)	1.00 (1.00 − 1.00)	1.00 (1.00 − 1.00)
Waist Circumference		
Abdominal Obesity	1.11 (1.02 − 1.22)	1.10 (1.01 − 1.20)
Educational Level		
None	1.02 (0.94 − 1.10)	1.00 (0.93 − 1-09)
Technical/Professional	0.88 (0.77 − 1.01)	0.90 (0.78 − 1.03)
Secondary	1.01 (0.88 − 1.17)	1.01 (0.88 − 1.17)
University	0.84 (0.74 − 0.96)	0.83 (0.73 − 0.94)
Unknown	1.21 (0.87 − 1.70)	1.15 (0.82 − 1.61)

c/d: cigarettes per day.
* restricted cubic splines omitted.

From the values presented, the estimates for energy intake in both models can attract the attention. Remember that this variable must be included in the model as a possible confounder (see definition in Chapter 5).

6.3.2 Comparison with two specific cases

Two cases relating to specific individuals profiles are presented in order to compare survival through the Cox model and the Flexible Parametric PH model. In order to see how the Flexible Parametric PH model fitted works in detail, the calculations of the survival will be performed manually.

Example 1

Let's consider a subject having the following characteristics:

- **mdscore (categorical)**: 3rd quartile
- **Center:** Asturias
- **Sex:** man
- **Age at recruitment:** < 40 years
- **Smoke Status:** former 11-20 cigarettes per day
- **BMI:** $[25, 30)$ kg/m^2
- **Physical Activity:** moderately active (reference category)
- **Energy Intake (centered):** -694.7063
- **Waist circumference:** normal (reference category)
- **Educational level:** technical or professional

With these features, the estimated survival probability by Cox model is 0.99 at $t = 58$ years.

With the Flexible Parametric Model, since the subject belongs to the "reference stratum", the fitted model reduces to:

$$\ln[H(t|\boldsymbol{x}_i)] = \underbrace{s(\ln(t)\gamma, k_0)}_{(2)} + \underbrace{\boldsymbol{x}_i\beta}_{(1)}$$

Part (1) of the formula is estimated with the information of the model directly, which can be summarized by the following table:

	$\hat{\beta}$	$\boldsymbol{x}_i$
mdscore: 3rd quartile	-0.17	1
Smoke Status: Former 11-20 c/d	0.23	1
BMI: $[25, 30)$ kg/m^2	-0.18	1
Energy Intake (centered)	0.00	-694.71
Educational Level: Tech/Prof	-0.11	1

regarding Part (2), we must remember that a model has 4 degrees of freedom (this also mean 5 knots) for the baseline function.

Then, considering the quartiles, according to the table below, the percentiles 0, 25, 50, 75 and 100 of the not-censored times were used

Knots	**Position**	**Value**
k_1	0 (min)	ln(36.2)
k_2	25	ln(61.2)
k_3	50	ln(68.6)
k_4	75	ln(75.2)
k_5	100 (max)	ln(87.2)

Remember that

$$s(\ln(t)|\gamma, k_0) = \gamma_0 + \gamma_1 r_{1i} + \gamma_2 r_{2i} + \gamma_3 r_{3i} + \gamma_4 r_{4i}$$

(see further details in Section 4.6)

where

$$\begin{aligned} r_1 &= \ln(t) \\ r_j &= (\ln t - k_j)_+^3 - h_j(\ln t - k_{\min})_+^3 - (1-h_j)(\ln t - k_{\max})_+^3 \end{aligned}$$

with $h_j = \dfrac{k_{\max} - k_j}{k_{\max} - k_{\min}}$.

replacing, the values of z can be obtained, however remember that internally `Stata` takes an additional step using the Gram-Schmidt orthogonalization:

	$\hat{\gamma}$	r_i
rcs_1	0.68	-1.23
rcs_2	-0.002	-0.26
rcs_3	-0.02	-0.96
rcs_4	-0.02	0.65
cons	-2.48	

Thus,

$$\ln[H(t|\boldsymbol{x}_i)] = -3.50$$

Corresponding to the predicted survival: $\hat{S}(t) = 0.97$. As you can see this value is very close to that obtained by Cox model (0.99).

In the second example, difficulty is added to the model: the stratum considered is different from the "reference stratum".

Example 2

Let's consider an individual with:

- **mdscore (categorical)**: 1st quartile
- **Center:** Granada
- **Sex:** Woman
- **Age at recruitment:** ≥ 60
- **Smoke Status:** no smoker (reference category)
- **BMI:** ≥ 30 kg/m^2
- **Physical Activity:** moderately active (reference category)
- **Energy Intake (centered):** -512.6188
- **Waist circumference:** abdominal obesity (reference category)
- **Educational level:** primary

With these characteristics, the estimated probability survival is 0.78 after Cox model fitted at $t = 80.3$ years.

For this individual profile, the Flexible Parametric Model fitted corresponds to

$$\ln[H(t|\boldsymbol{x}_i)] = \underbrace{s(\ln(t)\gamma, k_0)}_{(2)} + \underbrace{\sum_{j=1}^{D} s(\ln(t)|\delta_{jl_j}, k_j)}_{(3)} + \underbrace{\boldsymbol{x}_i\beta}_{(1)}$$

Part (1) of the formula is obtained by

	$\hat{\beta}$	$\boldsymbol{x}_i$
BMI: ≥ 30 kg/m^2	0.03	1
Energy Intake (centered)	0.00	-662.59
Waist Circumference: abdominal obesity	0.10	1

For part (2) similar to the previous example, in this case the values for r_i are

	$\hat{\gamma}$	r_i
rcs_1	0.68	1.48
rcs_2	-0.002	-1.12
rcs_3	-0.02	-0.84
rcs_4	-0.02	-0.10
cons	-2.48	

For part (3), it can be extended for this individual:

$$\begin{aligned}\sum_{j=1}^{D} s(\ln(t)|\delta_{jl_j}, kj) &= \delta_{121}r'_{121} + \delta_{221}r'_{221} + \delta_{341}r'_{341} \\ &\quad +\delta_{122}r'_{122} + \delta_{222}r'_{222} + \delta_{342}r'_{342}\end{aligned}$$

and can be summarized in the following table:

	$\hat{\delta}$	r'
rcs Granada_1	0.06	1.48
rcs Granada_2	-0.02	-1.21
rcs Female_1	-0.34	1.48
rcs Female_1	-0.07	-1.21
rcs $\geq$ 60 years_1	0.08	1.48
rcs $\geq$ 60 years_2	-0.10	-1.21

In this way

$$\ln[H(t|\boldsymbol{x}_i)] = -1.32$$

Corresponding to the predicted survival: $\hat{S}(t) = \exp(-\exp(-1.32)) = 0.77$, very similar to the value for estimated survival with Cox Model (0.78).

6.3.3 Baseline curves comparison

Figure 6.1 (on page 58) shows the graphs of different baseline curves for each stratum (in this case, and considering that there are 40 stratum, it has decided to set the center of Granada arbitrarily), both through nonparametric estimation, by Kaplan-Meier (blue), semi-parametric after adjusting the Cox model (in red), and parametric, through the application of Flexible Parametric PH model (green). The `Stata` code used to obtain these graphics can be seen in Appendix E.3.1.

You can see that all curves are quite similar, although, the curve estimated by the Flexible Parametric PH model is more smooth.

6.3.4 Comparison of survival curves

Commonly, the prognostic index, $x\hat{\beta}$, of a model is used as a summary for each individual of the information from the covariates. Often this index is categorized into groups and Kaplan-Meier survival curves are calculated to display the group-specific estimation. With the Flexible Parametric PH model the model-based mean survival curve for each group can also be computed, which is a smooth function of t (as opposed to the K-M curves, which are more or less jagged step functions as we saw before).

To get the mean curve, the survival curve is evaluated for each individual at a fixed set of time points and average these values at each time point.

An example is shown for Granada center in Figure 6.2 (on page 59). Here, four groups were created by categorizing the prognostic index at its 25th, 50th and 75th centiles. `Stata` code can be seen in Appendix E.3.2.
Figure 6.2 (on page 59) shows that the survival curves estimated by the Flexible Parametric PH model are smoother than the Kaplan-Meier survival curves predicted from a Cox model.

6.4 Other applications of the Flexible Parametric PH model

One of the great advantages of this model is its capacity of prediction. Its implementation in `Stata` also has a lot of power, making predictions very quickly. This special feature is presented below, however, it is necessary to clarify that is not the initial objective of this study, and that the type of comments that emerge from the analysis of these graphs can not be viewed lightly, because before making prediction is necessary to validate the predictive model.

Survival probabilities for individuals

A quantity of interest is conditional survival, which answers, for example, questions like "I have survived seventy years; what are my chances of surviving five years more?". Mathematically this corresponds to

$$\begin{aligned} Pr(T > 75|T > 70; x) &= Pr(T > 75 \text{ and } T > 70; x)/Pr(70; x) \\ &= Pr(T > 75; x)/Pr(T > 70; x) \\ &= S(75; x)/S(70; x) \end{aligned}$$

that is, the survival probability at seventy five years divided by that at seventy years. The result for the Granada center is shown in Figure 6.3 (on page 60). `Stata` code is available in Appendix E.3.3.

In this case only 4 graphics are shown, because the population was restricted only to who reached the 70 years. The median conditional seventy five-year probability is 0.92 for men (in both categories of age at recruitment) and 0.96 for women (also in both categories of age).

It is also necessary to insist that this is one of the advantages that this model presents, but you can not take lightly to predict, especially when the observed time is exceeded.

Survival probabilities across the risk spectrum

Rather than fixing times of interest and examining the distribution of survival probabilities, an alternative is to plot estimated probabilities against t at specified centiles of the distribution of the prognostic index (mdscore as continuous variable), for example, at the 10th, 20th, ... , 90th centiles. The plots give an impression of the available range of discrimination, showing what may happen to individuals at the extremes of the risk spectrum and in the middle. An example for Granada center is shown in Figure 6.4 (`Stata` code is in Appendix E.3.4).

Other applications of this model, but that can not be exemplified with the dataset from this study, are for example:

- Ease of working with time-dependent variables, in fact `stpm2` was created to work efficiently with this type of variables.
- Obtaining relative survival estimates. Relative survival is used extensively in population-based cancer studies to measure patient survival correcting for causes of death not related to the disease of interest. Relative survival provides a measure of net mortality, i.e. the probability of death due to cancer in the absence of other causes. (Lambert et al., 2010)
- Quantifying differences between two populations (for further information, see, for example Lambert et al. (2011)).

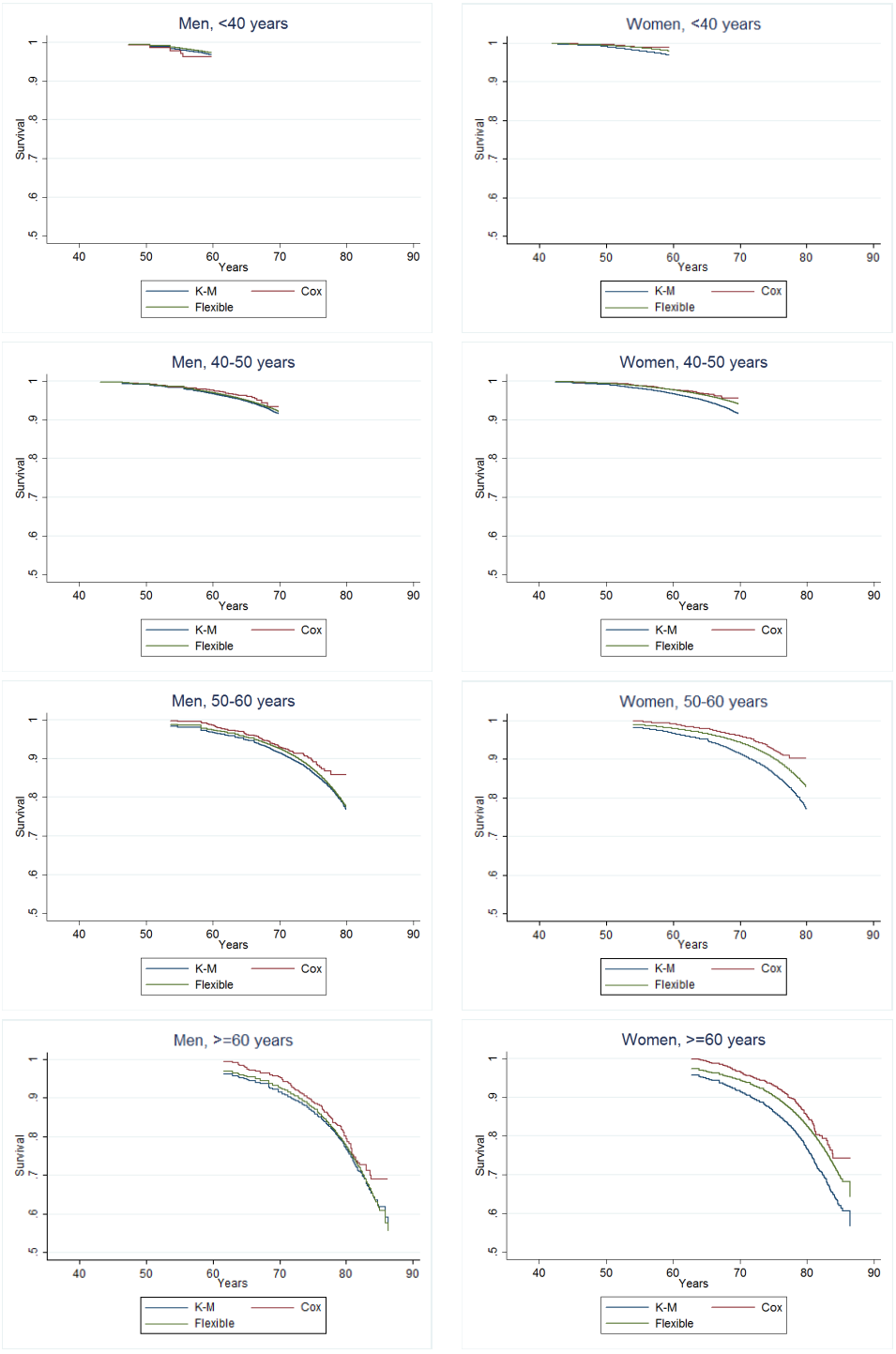

Figure 6.1: Baseline curves for center of Granada, by Sex and Age at Recruitment.

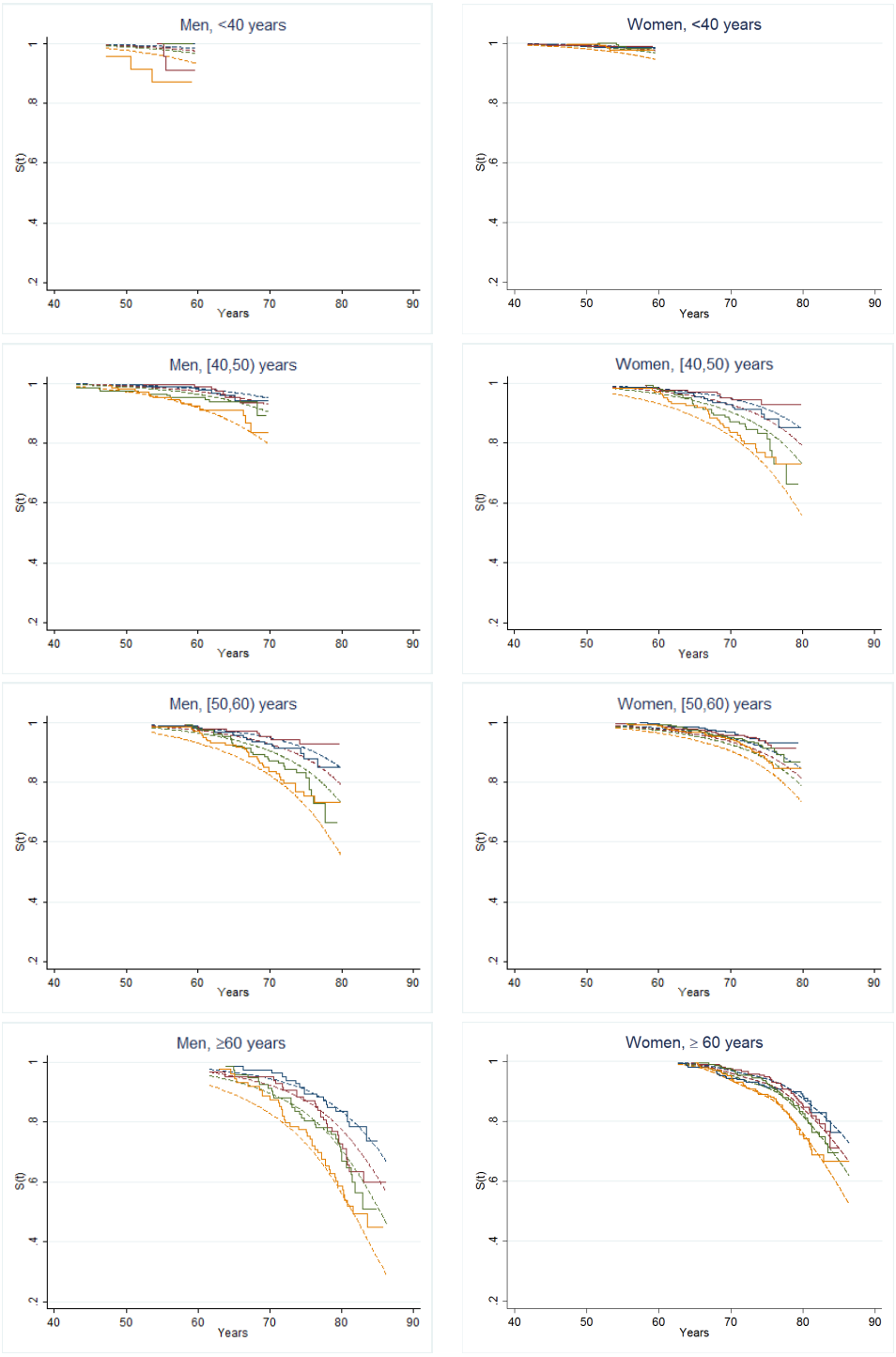

Figure 6.2: Kaplan-Meier curves (jagged lines) and mean survival curves (dashed lines) in 4 prognostic groups.

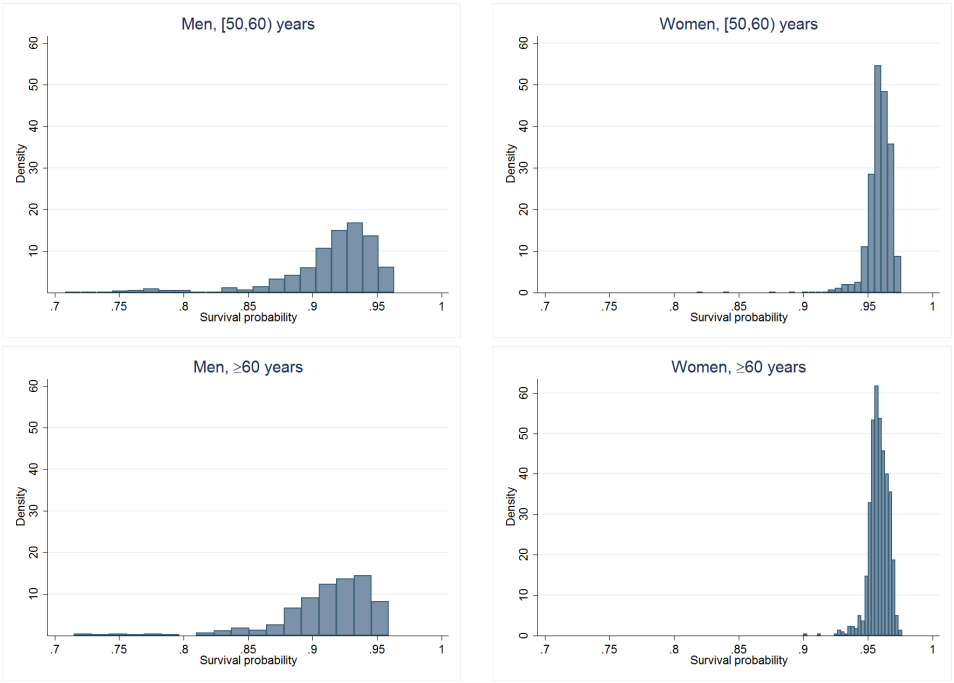

Figure 6.3: Conditional survival probabilities at 75 years, given no mortality at seventy years, for Granada center, by Sex and Age at recruitment.

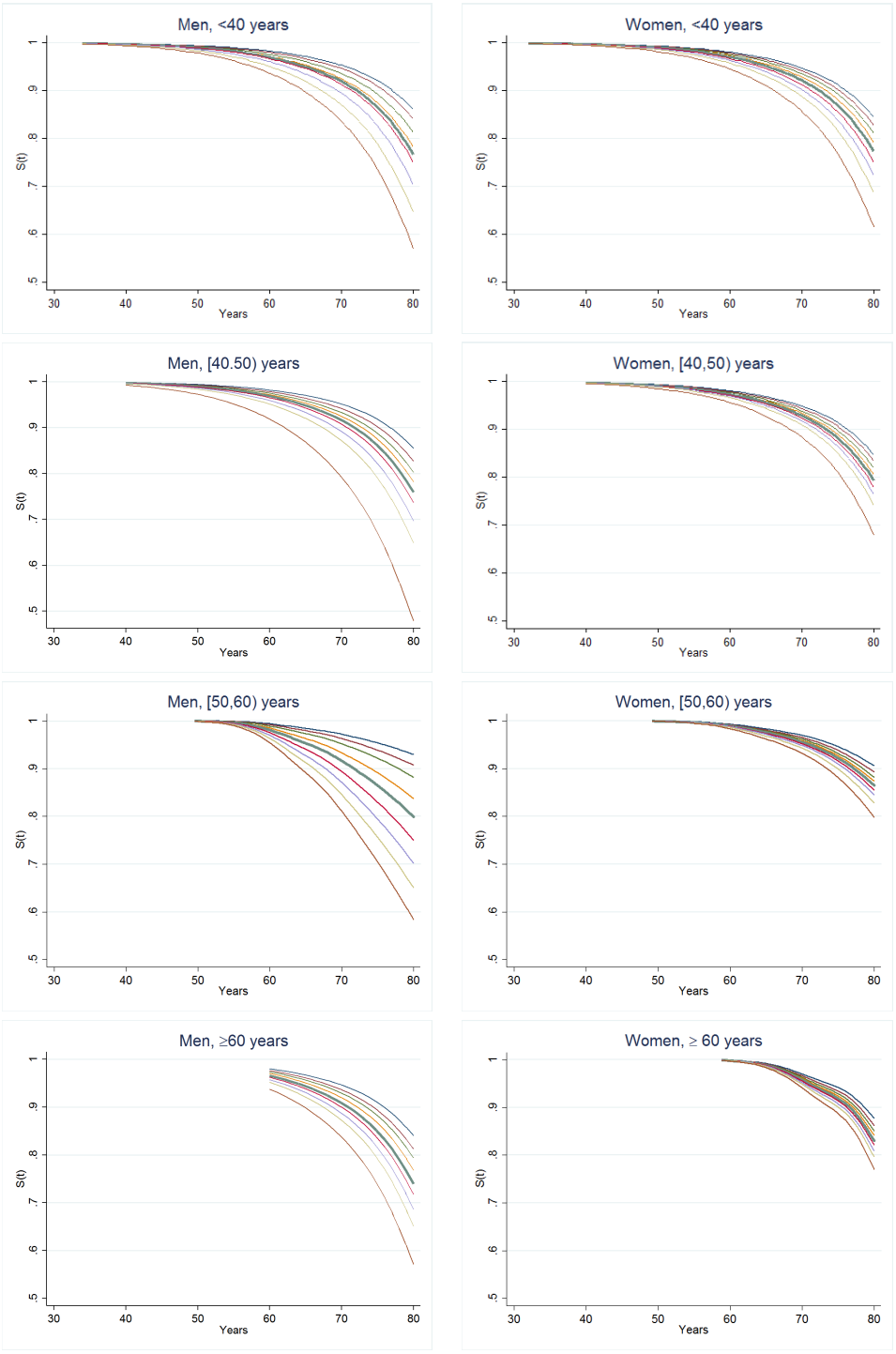

Figure 6.4: Survival probabilities at the 10th, 20th, ..., 90th centiles of the prognostic index. The uppermost line corresponds to the 10th centile of $x\hat{\beta}$ (that is, low risk) and the lowermost line, to the 90th centile (high risk). The bold line represents the 50th centile. This for Granada center, by Sex and Age at recruitment.

Chapter 7

Discussion

In this master thesis, we have analyzed the role of the Adherence to the Mediterranean Diet on global mortality and have performed a comparison of Cox's proportional hazard model and the Flexible Parametric Model proposed by Royston and Parmar.

We conclude that higher Adherence to the Mediterranean Diet reduces the risk of mortality by comparing the third and fourth quartiles of the variable mdscore with the first quartile, which is has a lower adhesion (3rd quartile HR=0.85, 95%CI (0.77; 0.93); the 4th quartile HR=0.78, 95%CI (0.71 ; 0.86)). Regarding the second quartile, when compared to the reference category it showed no significant difference (HR = 0.93, 95%CI (0.85 ; 1.02)).

The point estimates obtained by the Parametric Flexible PH model are very similar to those obtained by the Cox model. In the present study, for example, you can see the estimates for the hazard ratios of the quartiles of the mdscore, for Flexible Parametric Model were: 0.93, 0.85 and 0.78, and for Cox model were 0.93, 0.84 and 0.76, with the same standard error per each one (0.04), so you can see the regression estimates and standard errors are likely to be in close agreement. Moreover, the interpretation of the regression parameters in the Flexible Parametric PH model is exactly the same in the Cox model.

Concerning the survival function, $S(t)$, the Kaplan-Meier plot is an important feature of most survival analyses and is widely presented in publications of applied work. For the Cox model, Stata's `predict` command after `stcox` with the `basesurv()` option provides and estimate of the baseline survival function, $S_0(t) = S(t|x = 0)$. From the baseline survival and the hazard ratios, the survival functions for any combination of covariate values can be calculated. However, all such survival functions are step functions and typically are not particularly smooth. Also, the least precise parts of the curve get the most visual weight, which is a general criticism of Kaplan-Meier survival curves (StataCorp., 2003).

Kaplan-Meier-type estimates of $S(t)$ are composed of a sequence of point estimates of the survival function that are highly serially correlated. Accordingly, Kaplan-Meier plots tend to display "runs" of values that move away from and back towards the general trend, giving and undulating appearance. This may make the curve difficult to interpret and may lead to overemphasis of local features. In this point, splines certainly appear to offer adequate flexibility for approximating, and these functions have been widely used to model continuous variables in medicine and epidemiology (Royston and Lambert, 2011).

Parametric survival models generally provide smooth estimates of the hazard and survival functions for any combination of values. Exceptions are piecewise models, for example, the piecewise exponential models, for which the hazard function is a step function and the survival function has discontinuities in the first derivative.

When working with covariates that are not of direct interest and that do not follow the proportional hazards assumption, the Cox model stratified on them. In the case of the Flexible Parametric PH model, Royston and Lambert (2011) say stratification is the same as including time-dependent effects (See Chapter 7 of Royston and Lambert (2011)).

However, apparently the number of strata with which it works has not been considered; in this study the number of stratum is large (40 strata) and fitting the model in this way (considering these stratum variables (Sex, Center and Age at Recruitment) as time-dependent variables) the point estimates differs from those obtained by the Cox model and the estimated curves do not resemble the Kaplan-Meier curves, so it does not seem a proper way to do it. Moreover, the convergence is not reached in all cases, due to the huge number of strata (forty). So we can not say so lightly that stratification is the same as the effects for a time-dependent variable. Moreover, in this case, it makes no sense to treat stratum variables: Center, Sex and Age at Recruitment as time-dependent variables.

By fitting different Flexible Parametric Models, changing the number of knots for common splines and for stratum variable, some problems with the AIC and BIC presented per each model appears: the degrees of freedom are usually wrong, but this suggests, there are some bugs in the programming of these criteria.

Taking up, very few fitted models achieve the convergence (3 of 16 if we consider the mdscore divided into quartiles, and 3 of 16 as well, considering the mdscore as a continuous variable), this could be due to problems in the initial values (this argument was the response given by Christopher H. Jackson, attempting to start this study using the `R` package called `flexsurv`), however, the collaborative team working with him are trying to fix it. In `Stata`, according to Paul Lambert, whom I also contacted, the problem can be handled by command

`linits`, and by `initstrata`(*varname*), an option not documented yet.

The main disadvantage, from the computational point of view is, at least for this study, how fast the Flexible Parametric model can be compared with Cox model is not seen.

On the other hand, one of the great advantages of this model is not precisely the objective of this work. Here, only a small part of the power of prediction that have these Flexible Parametric models (the hazards proportional, and others not discussed in this study as the probit and the proportional odds) has been shown. You should note that the capacity of prediction can not be taken lightly. It is necessary to validate the predictive model (see further details about this in Chapter 6 of Royston and Lambert (2011)). Also the `predict` command is much more complete for Flexible Parametric model than for the Cox model).

Finally, the Flexible Parametric models have others applications, which were not subject of study here. One of them is that using Parametric models, a time-dependent HR can be obtained as a function of the estimated model parameters (the covariates and time). Furthermore, the command `predictnl` in `Stata` implements the delta method using numeric derivatives, to get standard deviations and confidence intervals quite easily. On the other hand, sometimes, a covariate whose effect is nonproportional on the hazards scale may be (much closer to) proportional on another scale, such as the odds or probit (inverse normal probability) scales. There are more applications like prediction out of sample, the use of multiple time scales, etc (for further details, see Royston and Lambert (2011)).

Further Research

According to the implementation, Flexible Parametric PH models can be adjusted appropriately in `Stata`, but not in `R`, when speaking of delayed entry models.

Moreover, in this study, the goodness of fit has been verified for the Cox model, however there is an of goodness of fit methods absence for Flexible Parametric models. Although it is not documented, so far only you can see martingale residuals, to observe if continuous covariates need some kind of transformation. For that reason, this topic is also proposed as a future research.

Bibliography

Agudo, A., L. Cabrera, P. Amiano, et al. (2007). Fruit and vegetable intakes, dietary antioxidant nutrients, and total mortality in Spanish adults: findings from the Spanish cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC- Spain). *Am J Clin Nutr 85*, 1632–1642.

Bingham, S. and E. Riboli (2004). Diet and cancer - the european prospective investigation into cancer and nutrition. *Nat Rev Cancer 4(3)*, 206–215.

Buckland, G., C. González, A. Agudo, et al. (2009). Adherence to the Mediterranean Diet and Risk of Coronary Hear Disease in the Spanish EPIC Cohort Study. *Am J Epidemiol 170(12)*, 1518–1529.

Canchola, A. J., S. L. Stewart, L. Bernstein, et al. (2003). Cox regression using different time-scales. *Western Users of SAS Software. San Francisco, California.*, 1–6.

Clements, M. and X.-R. Liu (2016). *rstpm2: Generalized Survival Models*. R package version 1.3.1.

Consultation, W. E. (2011). Waist circumference and waist-hip ratio.

Corporation, S. (2003). *Survival Analysis and Epidemiological Tables* (1 ed.). Stata Press.

Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society (serie B) 34(2)*, 187–220.

de Boor, C. (2001). *A Practical Guide to Splines*. Revised ed. New York: Springer.

Durrleman, S. and R. Simon (1989). Flexible regresion models with cubic splines. *Statistics in Medicine 8*, 551–561.

EPIC Group of Spain (1997a). Relative validity and reproducibility of a diet history questionnaire in Spain. I. Foods. *Int J Epidemiol 26(Suppl 1)*, s91–s99.

EPIC Group of Spain (1997b). Relative validity and reproducibility of a diet history questionnaire in Spain. II. Nutrients. *Int J Epidemiol 26(Suppl 1)*, s100–s109.

EPIC Spain (2016). Epic study. http://epic.iarc.fr/centers/spain.php. Accessed: March 2016.

EPIC study (2016). Epic study. http://epic.iarc.fr. Accessed: March 2016.

Gómez, G., O. Julià, and K. Langohr (2015). *Análisis de Supervivencia*.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*, 515–526.

Haftenberger, M., A. Schuit, M. Tormo, H. Boeing, N. Wareham, H. Bueno-de Mesquita, M. Kumle, et al. (2002). Physical activity of subjects aged 50-64 years involved in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Public Health Nutr. 5*, 1163–1176.

Harrell Jr., F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Hurley, M. A. (2015). A reference relative time-scale as an alternative to chronological age for cohorts with long follow-up. *Emerging Themes in Epidemiology*, 12–18.

Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software 70*(8), 1–33.

Kaaks, R. and E. Riboli (1997). Validation and Calibration of Dietary Intake Measurements in the EPIC Project: Methodological Considerations. *International Journal of Epidemiology*, S15–S25.

Kalbfleisch, J. and R. Prentice (2002). *Statistical Analysis of Failure Time Data* (2 ed.). John Wiley and Sons, Inc.

Kaplan, E. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association 53, No. 282*, 457–481.

Klein, J. and M. Moeschberger (2003). *Survival Analysis* (2 ed.). Springer-Verlag New York.

Lambert, P., P. Dickman, C. Nelson, and P. Royston (2010). Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statist. Med. 29*, 885–895.

Lambert, P. C., L. Holmberg, F. Sandin, F. Bray, K. Linklater, A. Purushotham, D. Robinson, and H. Møller (2011). Quantifying differences in breast cancer survival between England and Norway. *Cancer Epidemiology 35*, 526–533.

Lambert, P. C. and P. Royston (2009). Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal*, 1–22.

Mitrou, P., V. Kipnis, A. Thiebaut, et al. (2007). Mediterranean dietary pattern and prediction of all-cause mortality in a US population: results from the NIH-AARP Diet and Health Study. *Arch Intern Med 167*, 2461–2468.

Nelson, C., P. Lambert, I. Squire, and D. Jones (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine 26*(30), 5486–5498.

Pencina, M. J., M. G. Larson, and R. B. D'Agostino (2007). Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statistics in Medicine 26*, 1343–1359.

Pérez-López, F., P. Chedraui, J. Haya, et al. (2009). Effects of the Mediterranean diet on longevity and age-related morbid conditions. *Maturitas 64*, 67–79.

Riboli, E., K. Hunt, N. Slimani, et al. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr 5(6B)*, 1113–1124.

Riboli, E. and R. Kaaks (1997). The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol 26(Suppl 1)*, s6–s14.

Roman, B., L. Carta, M. Martinez-Gonzalez, et al. (2008). Effectiveness of the mediterranean diet in the elderly. *Clin Interv Aging 3*, 97–109.

Rooney, J., S. Byrne, M. Heverin, B. Corr, M. Elamin, A. Staines, B. Goldacre, and O. Hardiman (2013). Survival Analysis of Irish Amyotrophic Lateral Sclerosis Patients Diagnosed from 1995-2010. *PLOS ONE 8*, 1–10.

Royston, P. (2001). Flexible parametric alternatives to the Cox model, and more. *The Stata Journal 1*(1), 1–28.

Royston, P. (2004). Flexible parametric alternatives to the Cox model: update. *The Stata Journal 4*(1), 98–101.

Royston, P. (2011). Estimating a smooth baseline hazard function for the Cox model. *Research report No.314, Department of Statistical Science, University College London*, 1–16.

Royston, P. and P. C. Lambert (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model* (1 ed.). Stata Press.

Royston, P. and M. K. B. Parmar (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine 21*, 2175–2197.

Sauerbrei, W., P. Royston, and M. Look (2007). A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal 49*, 453–473.

Slimani, N., G. Deharveng, I. Unwin, et al. (2007). The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr 61*, 1037–1056.

Sofi, F., F. Cesari, R. Abbate, et al. (2008). Adherence to Mediterranean diet and health status: meta-analysis. *BMJ 337*, a1344–a1350.

StataCorp. (2003). *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation.

Trichopoulou, A., T. Costacou, C. Bamia, et al. (2003). Adherence to a Mediterranean diet and survival in a Greek population. *N Engl J Med 348*, 2599–2608.

Trichopoulou, A., A. Kouris-Blazos, M. Wahlqvist, et al. (1995). Diet and overall survival in elderly people. *BMJ 311*, 1457–1460.

Trichopoulou, A. and P. Lagiou (1997). Healthy traditional Mediterranean diet: an expression of culture, history, and lifestyle. *Nutr Rev 55*, 383–389.

Trichopoulou, A., P. Orfanos, T. Norat, et al. (2005). Modified Mediterranean diet and survival: EPIC-elderly prospective cohort study. *BMJ 330*, 991–997.

UNESCO (2016). Intangible heritage: Mediterranean diet. http://www.unesco.org/culture/ich/en/lists?RL=00884. Accessed: March 2016.

Willet, W. C., G. R. Howe, and L. H. Kushi (1997). Adjustment for total energy intake in epidemiologic studies. *Am J Clin Nutr 65(Suppl)*, 1220S–1228S.

Xue, X., M. Y. Kim, and R. E. Shore (2007). Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure. *Lifetime Data Anal 13*, 333–350.

Appendix A

R code for mdscore

- **mdscore**$_{sum}$

```
mdscore_sum <- with(data, vegetables + fruit + legum + fish
              + cereal + olive + wine - meat - dairy)
```

- **mdscore**$_{cdf}$

```
cdf<-function(x){
  Fn<-ecdf(x)
  p<-Fn(x)
  return(p)
}

data$vegetable_cdf    <- cdf(data$vegetable)
data$fruit_cdf        <- cdf(data$fruit)
data$legum_cdf        <- cdf(data$legum)
data$fish_cdf         <- cdf(data$fish)
data$cereal_cdf       <- cdf(data$cereal)
data$olive_cdf        <- cdf(data$olive)
data$wine_cdf         <- cdf(data$wine)
data$meat_cdf         <- cdf(data$meat)
data$dairy_cdf        <- cdf(data$dairy)

mdscore_cdf    <- with(data, vegetable_cdf + fruit_cdf + legum_cdf
              + fish_cdf + cereal_cdf + olive_cdf
              + wine_cdf - meat_cdf - dairy_cdf)

# For categorical index:

q1 <- quantile(mdscore_v2)[[2]]
q2 <- quantile(mdscore_v2)[[3]]
q3 <- quantile(mdscore_v2)[[4]]
```

```
mdscore_cdf_c <-ifelse(mdscore_v2 < q1, 1,
              ifelse(mdscore_v2 < q2, 2,
              ifelse(mdscore_v3 < q3, 3, 4)))
```

- **mdscore**$_{sd}$

```
attach(data)

# Standard deviations from Murcia and Granada

std_f<-function(x){
  m=mean(subset(data,(center=="Granada" | center=="Murcia"))[,x])
  std=sd(subset(data,(center=="Granada" | center=="Murcia"))[,x])
  y=(data[,x]-m)/std
  return(y)
}

fx<-function(x){
       if (x<(-2)) return(-2)
       if (-2<=x & x<(-1)) return(-1)
       if (-1<=x & x<=1) return(0)
       if (1<x & x<=2) return(1)
       if (x>2) return(2)
}

data$vegetable_sd    <- sapply(std_f("vegetable"),fx)
data$fruit_sd        <- sapply(std_f("fruit"),fx)
data$legum_sd        <- sapply(std_f("legum"),fx)
data$fish_sd         <- sapply(std_f("fish"),fx)
data$cereal_sd       <- sapply(std_f("cereal"),fx)
data$olive_sd        <- sapply(std_f("olive"),fx)
data$wine_sd         <- sapply(std_f("wine"),fx)
data$meat_sd         <- sapply(std_f("meat"),fx)
data$dairy_sd        <- sapply(std_f("dairy"),fx)
detach(data)

mdscore_sd    <- with(data, vegetable_sd + fruit_sd + legum_sd
              + fish_sd + cereal_sd + olive_sd + wine_sd
              - meat_sd - dairy_sd)
```

- **mdscore**$_{ter}$

```
vegetable_q   <- quantile(data$vegetable, probs=seq(0,1,1/3))
fruit_q       <- quantile(data$fruit, probs=seq(0,1,1/3))
legum_q       <- quantile(data$legum, probs=seq(0,1,1/3))
fish_q        <- quantile(data$fish, probs=seq(0,1,1/3))
cereal_q      <- quantile(data$cereal, probs=seq(0,1,1/3))
```

```
olive_q       <- quantile(data$olive, probs=seq(0,1,1/3))
wine_q        <- quantile(data$wine, probs=seq(0,1,1/3))
meat_q        <- quantile(data$meat, probs=seq(0,1,1/3))
dairy_q       <- quantile(data$dairy, probs=seq(0,1,1/3))

vegetable_ter <- ifelse(data$vegetable < vegetable_q[[2]],0,
              ifelse(data$veg < vegetable_q[[3]],1,2))
fruit_ter      <- ifelse(data$fruit < fruit_q[[2]],0,
              ifelse(data$fruit < fruit_q[[3]],1,2))
legum_ter      <- ifelse(data$legum < legum_q[[2]],0,
              ifelse(data$legum < legum_q[[3]],1,2))
fish_ter       <- ifelse(data$fish < fish_q[[2]],0,
              ifelse(data$fish < fish_q[[3]],1,2))
cereal_ter     <- ifelse(data$cereal < cereal_q[[2]],0,
              ifelse(data$cereal < cereal_q[[3]],1,2))
olive_ter      <- ifelse(data$olive < olive_q[[2]],0,
              ifelse(data$olive < olive_q[[3]],1,2))
wine_ter       <- ifelse(data$wine < wine_q[[2]],2,
              ifelse(data$wine< wine_q[[3]],1,0))
meat_ter       <- ifelse(data$meat < meat_q[[2]],2,
              ifelse(data$meat < meat_q[[3]],1,0))
dairy_ter      <- ifelse(data$dairy < dairy_q[[2]],2,
              ifelse(data$dairy < dairy_q[[3]],1,0))

mdscore_ter <- with(data, vegetable_ter + fruit_ter + legum_ter
              + fish_ter + cereal_ter + olive_ter + wine_ter
              + meat_ter + dairy_ter)
```

Appendix B

Stata commands

Some key concepts

Before some details of using stset will be explained, it is important to explain some key concepts about meaning of time.

- **Time origin** This defines time 0- that is, when the clock starts and we start recording time. Examples of time 0 include date of diagnosis, date of randomization, and date of birth.
- **Exit time** This defines the time (or date) when a subject stops being at risk, either by experiencing the event or being censored.
- **Failure indicator** This defines whether a subject experiences the event or is censored.
- **Entry time** This defines when the subject starts being at risk. In many cases, this is the same as the time origin- that is, time 0.
- **Analysis time** This is the amount of time the person was at risk- that is, the difference between the exit and the entry times.

stset command

The `stset` command tells Stata about the format of the survival data. Stata only needs to be informed once of the format. All subsequent survival analysis commands (the `st` commands) use this information. The syntax of the `stset` command is given by:

`stset` *timevar* [if][weight][,`failure`(*failvar*[==*numlist*])*other_options*]

- The *timevar* variable is compulsory. It is the survival time (or a date) of the event or the censoring time.

- The `failure`(*failvar*[==*numlist*]) option is optional, but it is good practice to always use ut. If this option is omitted, it is assumed that all subjects experience the event. It is a number list giving the values of *failvar* that indicate a failure, all others values indicate a censoring. In many cases, *failvar* is a single number, but a number list is useful if, for example, different codes are used for different causes of death.
- The `exit()` option gives the latest time at which the subject is at risk. The default is `exit(failure)`; that is the subject is removed from the risk set after the first event, even if there are subsequent records indicating additional failures for the subject.
- The `origin()` option gives the time origin of the time scale. The default is zero.
- The `enter()` option gives the time at which the subject becomes at risk of experiencing the event. It is useful, for example, in period analysis when the survival time is artificially left-truncated.
- The `scale(#)` option transforms the survival time. For example, to transform the time scale from days to years.
- The `id`(*varname*) option specifies an identification (ID) number for each subject.

Variables created by the stset command

The `stset` command usually creates four new variables. It creates five new variables if the `origin()` option is used. These variables contain all the necessary information about the structure of the survival data for the `st` survival analysis commands. The created variables are

- `_t` analysis time when record ends
- `_d` 1 if failure, 0 if censored
- `_t0` analysis time when record begins
- `_st` 1 if the record is included, 0 if excluded
- `_origin` the time origin if the `origin()` option of `stset` is used.

The `stset` command is extremely powerful for setting up survival data and is a feature of Stata that we are particularly fond of compared to the implementation of survival analysis in other packages.

Like most Stata estimation commands, `stpm2` has two parts: parameter estimation (that is, model fitting) and postestimation facilities (prediction). The former is accomplished by `stpm2`, the latter by `predict`.

Model fitting

The syntax of `stpm2` is basically simple:

`stmp2`[*varlist*] [*in*], `scale(hazard`odds—normal) df`(#)`—[`tvc`(*varlist*) `dftvc`(*df-list*)*other_options*]

The covariates are included in *varlist*. There are two keys options: `df()` and `scale()`. The first controls the complexity (degrees of freedom) of the baseline distribution function. The second determines whether the model is to be fit on the `hazard`, `odds`, or `normal` scale.

Models with time- dependent effects require the `tvc` and `dftvc()` options. Some examples are

```
stpm2, scale(hazard) df(3)
stpm2 trt, scale(hazard) df(2) eform
stpm2 trt, scale(odds) df(2)
stpm2 trt, scale(hazard) df(2) tvc(trt)dftvc(1)
```

Postestimation facilities (prediction)

The `predict` command, used after fitting a model with `stpm2`, has many options that provide considerable richness in what we can estimate. The most importante options are probably `survival`, `hazard`, `ci` and `zeros`, followed by `hrnumerator()`, `hrdenominator()`, `hdiff1()`, `hdiff2()`, `sdiff1()`, `sdiff2()`, `at`, and `timevar()`. The `hrnumerator()`and `hrdenominator()` options give HRs (which may vary with time `_t`), irrespective of the `scale()` that we have assumed for covariate effects, The `ci` option generally provides a CI for whatever is being predicted. The `zeros` option predicts with all covariates set to zero, thus giving baseline values.

Some examples:

```
stpm2 trt, scale(hazard) df(2)
predict basesurv, survival zeros
predict surv1, at(trt 1)
hazarddiff, hdiff1(trt 1) ci
predict survdiff, sdiff1(trt 1) ci
stpm2 trt, scale(hazard) df(2) tvc(trt) dftvc(1)
predict hr, hrnumerator(trt 1) hrdenominator(trt 0) ci
```

Appendix C

Outputs related to the Cox Model

C.1 Residual Analysis for Cox model selected

In this section, the residual analysis for Cox model selected (using mdscore_{cdf} as a categorical variable is presented.

Table C.1: Residual Analysis: Proportional Hazards Assumption

	rho	**chisq**	**p-value**
mdscore			
2nd quartile	-0.009	0.30	0.582
3rd quartile	-0.001	0.01	0.928
4th quartile	-0.028	2.92	0.087
Smoke Status			
Current 1-15 c/d	0.020	1.48	0.224
Current 16-25 c/d	-0.022	1.81	0.179
Current $\geq$26 c/d	-0.035	4.35	0.037
Former $<$ 10 years	-0.010	0.32	0.570
Former 11-20 years	-0.009	0.29	0.589
Former $>$20 years	-0.002	0.01	0.925
Cigar/Pipe, occasional	-0.015	0.78	0.376
Current/Former,missing	-0.031	3.59	0.058
Unknown	0.023	1.82	0.177
BMI (kg/m^2)			
$[25, 30)$	0.019	1.29	0.256
$\geq$30	0.032	3.71	0.054

Table C.1: (continued)

	rho	**chisq**	**p-value**
Physical Activity			
Inactive	0.014	0.71	0.399
Moderately inactive	-0.035	4.64	0.031
Active	-0.028	2.85	0.091
Energy Intake (kcal/day)	-0.019	1.50	0.221
Waist Circumference			
Abdominal Obesity	-0.003	0.03	0.854
Educational Level			
None	-0.013	0.57	0.450
Technical/Professional	0.007	0.17	0.679
Secondary	-0.021	1.62	0.202
University	0.012	0.50	0.478
Unknown	0.023	1.91	0.167
GLOBAL		41.40	0.015

c/d: cigarettes per day

As you can see the null hypothesis for proportionals risk is no rejected for every category of mdscore_{cdf}, body mass index, energy and waist circumference.

Also we can see this with the plot for **Schoenfeld residuals** for the main variable:

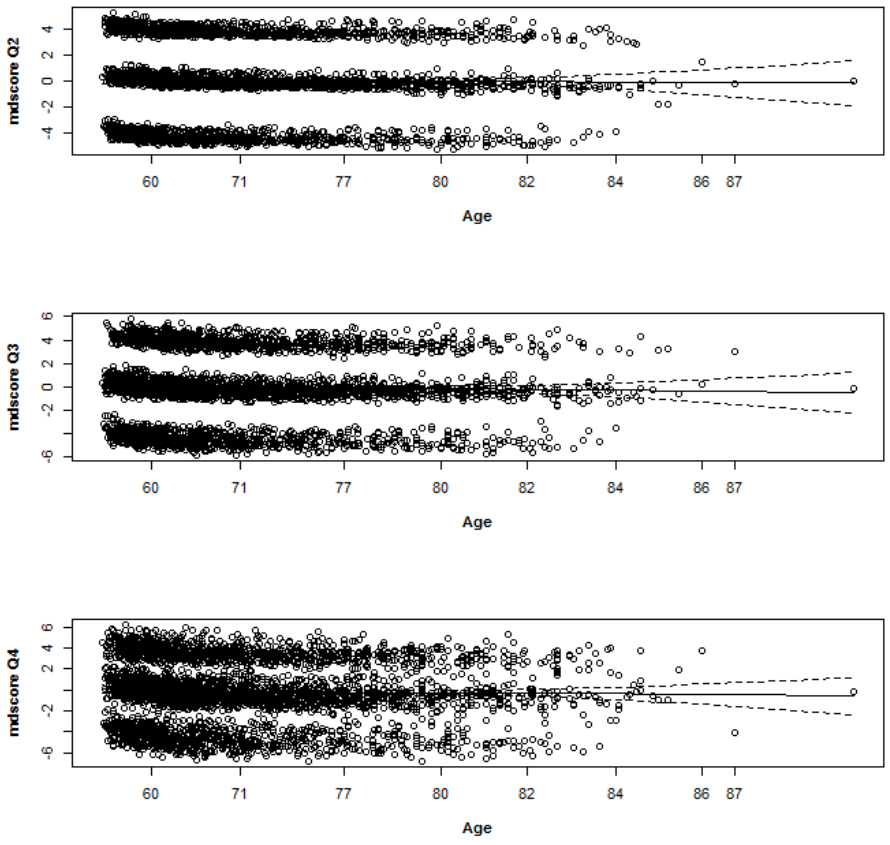

Figure C.1: Residual Analysis: Schoenfeld Residuals for categories of mdscore

Influence of each individual (dfbeta residuals)

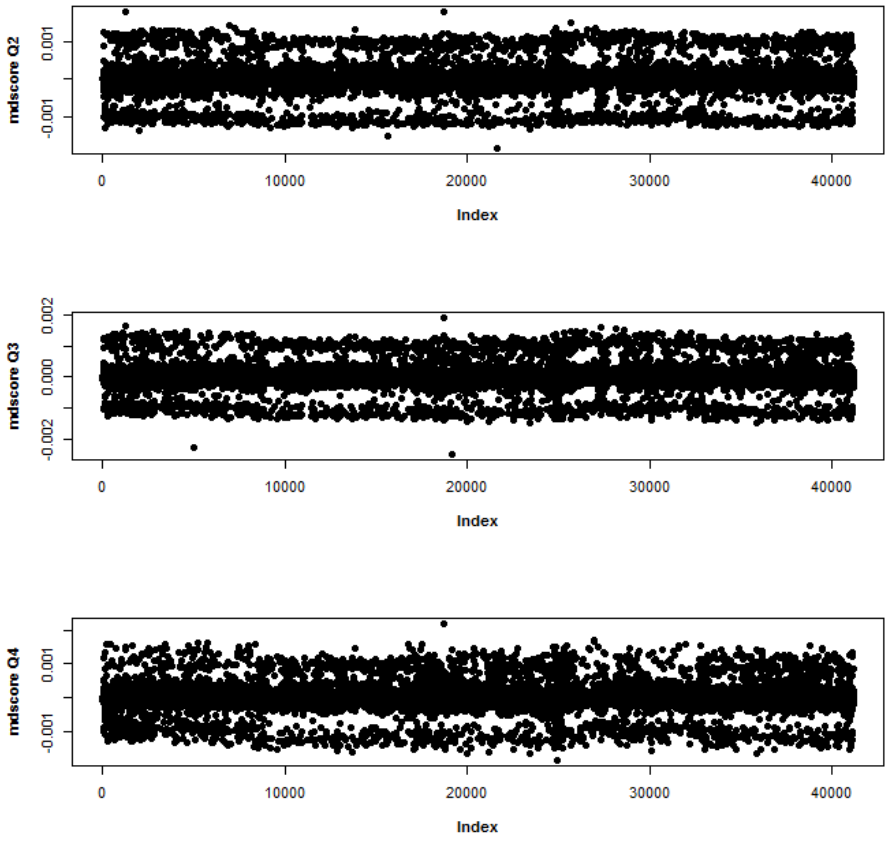

Figure C.2: Residual Analysis: Dfbeta Residuals for categories of mdscore

No influential individuals are observed.

Global Fit (deviance residuals)

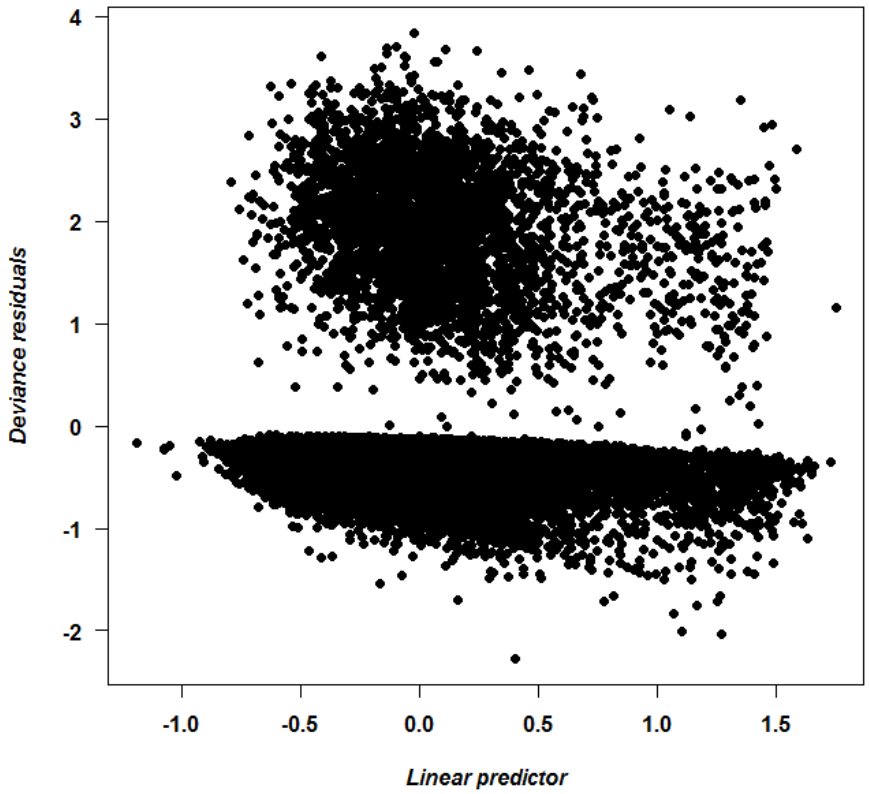

Figure C.3: Residual Analysis: Deviance Residuals for model considering md-score divided into quartiles.

Through the analysis of the deviance residuals it can be seen that the overall fit is appropriate.

Also remember, Cox models has been adjusted for others versions of the md-score, and these are presented below.

C.2 Cox Model for other versions of mdscore

Model with mdscore as a continuous variable

Next model is with mdscore_{cdf} variable, as continuous one:

Table C.2: Cox Model considering mdscore as continuous variable.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore	-0.11	0.90	0.02	-5.78	<0.001	0.86 − 0.93
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.94	<0.001	1.37 − 1.76
Current 16-25 c/d	0.84	2.31	0.16	12.43	<0.001	2.02 − 2.63
Current $\geq$ 26 c/d	1.42	4.14	0.33	17.94	<0.001	3.54 − 4.83
Former < 10 years	0.42	1.52	0.09	6.86	<0.001	1.35 − 1.71
Former 11-20 years	0.23	1.25	0.10	2.81	0.005	1.07 − 1.47
Former > 20 years	0.06	1.07	0.11	0.60	0.547	0.86 − 1.32
Cigar/Pipe, occasional	0.43	1.53	0.10	6.59	<0.001	1.35 − 1.74
Current/Former, missing	0.37	1.45	0.39	1.38	0.167	0.86 − 2.46
Unknown	-0.42	0.65	0.66	-0.42	0.673	0.09 − 4.66
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.42	0.001	0.75 − 0.93
$\geq$ 30	0.03	1.03	0.07	0.52	0.604	0.91 − 1.17
Physical Activity						
Inactive	0.05	1.05	0.06	0.95	0.341	0.95 − 1.17
Moderately inactive	0.07	1.07	0.05	1.58	0.115	0.98 − 1.16
Active	-0.03	0.97	0.06	-0.49	0.627	0.85 − 1.01
Energy Intake (kcal/day)	0.00	1.00	0.00	-2.76	0.006	1.00 − 1.00
Waist Circumference						
Abdominal obesity	0.11	1.11	0.05	2.36	0.018	1.02 − 1.22
Educational Level						
None	0.02	1.02	0.04	0.47	0.639	0.94 − 1.10
Technical/Professional	-0.13	0.88	0.06	-1.79	0.073	0.76 − 1.01
Secondary	0.01	1.01	0.08	0.18	0.858	0.88 − 1.17
University	-0.17	0.84	0.06	-2.63	0.009	0.74 − 0.96
Unknown	0.19	1.21	0.21	1.13	0.260	0.87 − 1.70

c/d: cigarettes per day

Also for this model the residual analysis was checked.

Proportional Hazards

Table C.3: Residual Analysis: Proportional Hazards Assumption

	rho	**chisq**	**p-value**
mdscore	-0.021	1.68	0.195
Smoke Status			
Current 1-15 c/d	0.020	1.46	0.228
Current 16-25 c/d	-0.022	1.80	0.180
Current $\geq$26 c/d	-0.034	4.11	0.043
Former $<$ 10 years	-0.010	0.34	0.562
Former 11-20 years	-0.010	0.33	0.565
Former $>$20 years	-0.001	0.01	0.930
Cigar/Pipe, occasional	-0.015	0.82	0.365
Current/Former,missing	-0.031	3.61	0.058
Unknown	0.023	1.87	0.171
BMI (kg/m^2)			
$[25, 30)$	0.019	1.29	0.256
$\geq$30	0.032	3.69	0.054
Physical Activity			
Inactive	0.014	0.73	0.394
Moderately inactive	-0.035	4.66	0.031
Active	-0.028	2.90	0.089
Energy Intake (kcal/day)	-0.018	1.39	0.238
Waist Circumference			
Abdominal obesity	-0.003	0.03	0.855
Educational Level			
None	-0.012	0.53	0.468
Technical/Professional	0.006	0.14	0.704
Secondary	-0.021	1.60	0.206
University	0.012	0.53	0.465
Unknown	0.023	1.99	0.158
GLOBAL		39.12	0.014

c/d: cigarettes per day

As you can see the null hypothesis for proportionals risk is no rejected for the main variable of interest mdscore$_{cdf}$, smoke intensity, body mass index, energy, waist circumference and education level.

Also this can be seen with the plot for Schoenfeld residuals:

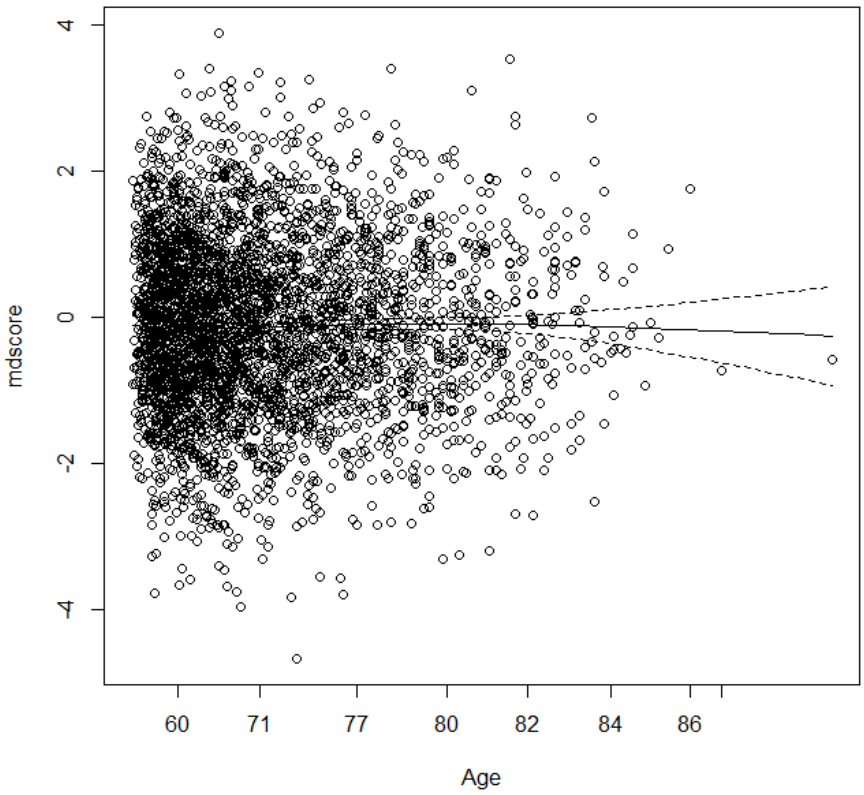

Figure C.4: Residual Analysis: Schoenfeld Residuals for mdscore as continuous variable.

Influence of each individual (dfbeta residuals)

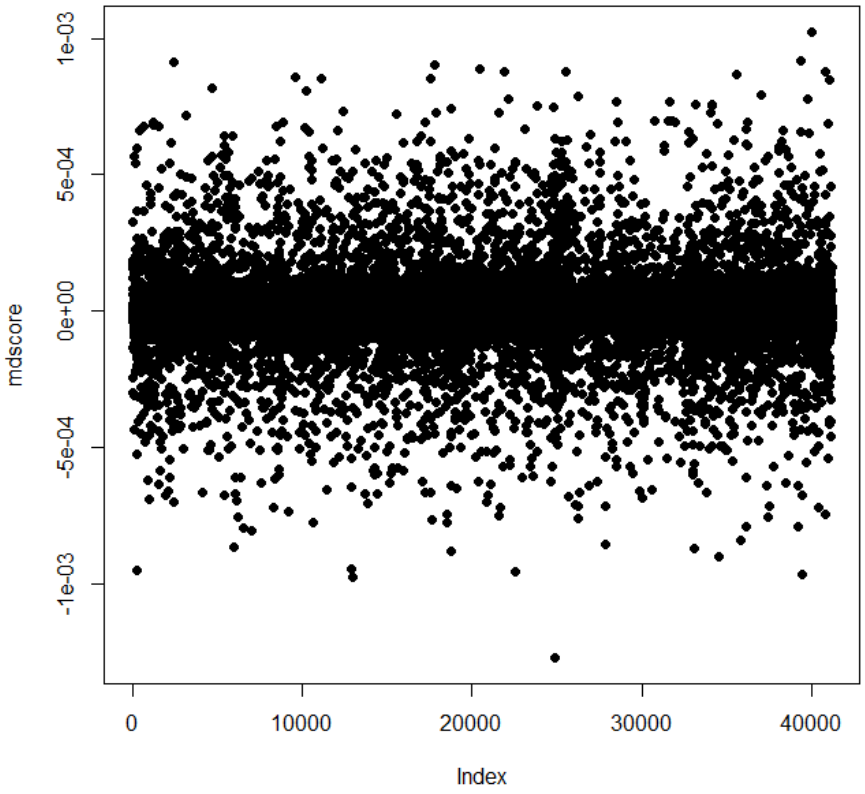

Figure C.5: Residual Analysis: Dfbeta residuals for mdscore as continuous variable.

No presence of highly influential observations is observed.

Global Fit (deviance residuals)

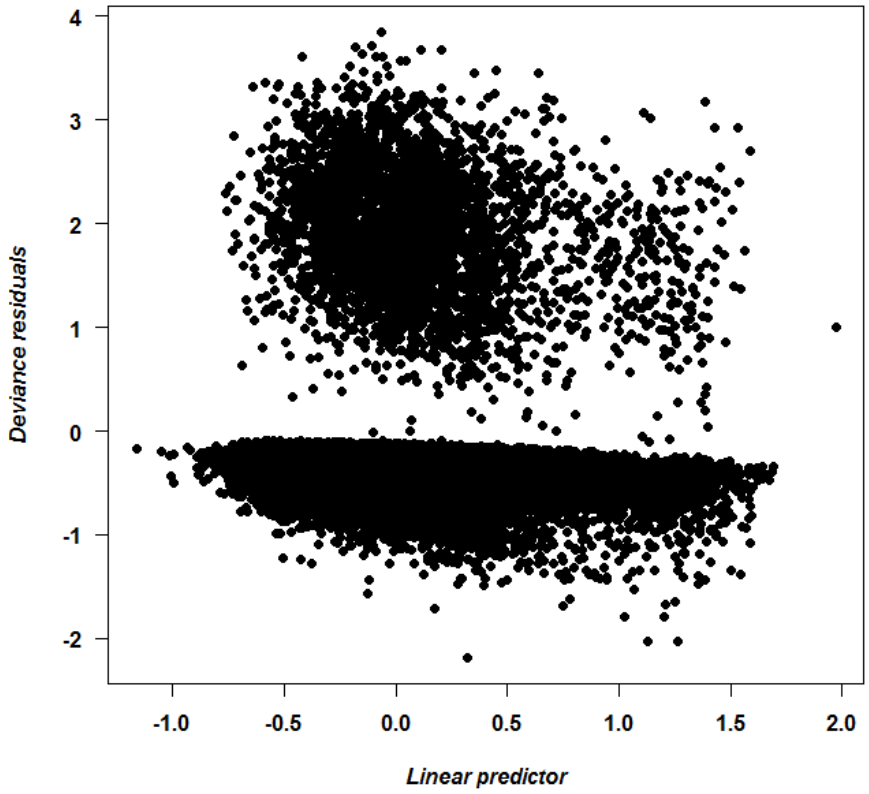

Figure C.6: Residual Analysis: Deviance residuals for model considering md-score as a continuous variable.

The overall fit seems adequate.

Initial model

This model is called initial model, only have the adjust variables:

```
> m0<-coxph(Surv(Age_Recr,Age_Exit,cens) ~ Smoke_Status +BMI
+           + Physical_Activity + Energy + Waist_C + L_School
+           + strata(Center,Sex,Age), data)
```

Table C.4: Initial model fitted: only with others covariates.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
Smoke Status						
Current 1-15 c/d	0.45	1.56	0.10	7.03	<0.001	1.38 − 1.77
Current 16-25 c/d	0.85	2.35	0.16	12.70	<0.001	2.06 − 2.68
Current $\geq$ 26 c/d	1.45	4.27	0.34	18.37	<0.001	3.66 − 4.98
Former < 10 years	0.41	1.51	0.09	6.81	<0.001	1.34 − 1.71
Former 11-20 years	0.22	1.25	0.10	2.71	0.007	1.06 − 1.46
Former > 20 years	0.06	1.06	0.11	0.53	0.595	0.86 − 1.31
Cigar/Pipe, occasional	0.43	1.54	0.10	6.64	<0.001	1.35 − 1.75
Current/Former,missing	0.38	1.46	0.39	1.41	0.159	0.86 − 2.48
Unknown	-0.42	0.66	0.66	-0.42	0.676	0.09 − 4.69
BMI (kg/m^2)						
$[25, 30)$	-0.19	0.83	0.04	-3.56	<0.001	0.75 − 0.92
$\geq$ 30	0.03	1.03	0.07	0.38	0.701	0.90 − 1.16
Physical Activity						
Inactive	0.06	1.06	0.06	1.16	0.247	0.96 − 1.18
Moderately inactive	0.07	1.08	0.05	1.66	0.096	0.99 − 1.17
Active	-0.04	0.96	0.06	-0.55	0.581	0.85 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-1.95	0.051	1.00 − 1.00
Waist Circumference						
Abdominal obesity	0.11	1.12	0.05	2.44	0.015	1.02 − 1.22
Educational Level						
None	0.02	1.02	0.04	0.43	0.666	0.94 − 1.10
Technical/Professional	-0.14	0.87	0.06	-1.95	0.051	0.76 − 1.00
Secondary	0.02	1.01	0.08	0.21	0.837	0.88 − 1.18
University	-0.17	0.84	0.06	-2.64	0.008	0.74 − 0.96
Unknown	0.18	1.20	0.21	1.06	0.291	0.86 − 1.68

c/d: cigarettes per day

Model with mdscore$_{sum}$ as the main variable.

This model include the mdscore$_{sum}$ index, obtained by adding the components.

```
> m.sum<-coxph(Surv(Age_Recr,Age_Exit,cens) ~ mdscore_sum +
+              Smoke_Status + BMI + Physical_Activity + Energy +
+              Waist_C+ L_School + strata(Center,Sex, Age), data)
```

Model with mdscore$_{sd}$ as the main variable.

This model include the mdscore$_{sd}$ index, obstained by standard deviations.

```
> m.sd<-coxph(Surv(Age_Recruitment,Age_Exit,cens) ~ mdscore_sd +
+             Smoke_Status + BMI + Physical_Activity + Energy +
+             Waist_C + L_School + strata(Center,Sex,Age), data)
```

Model with mdscore$_{ter}$ as the main variable.

This model include the mdscore$_{ter}$:

```
> m.ter <-coxph(Surv(Age_Recr,Age_Exit,esvital) ~ mdscore_ter +
+               Smoke_Status + BMI + Physical_Activity + Energy +
+               Waist_C + L_School + strata(Center,Sex,Age),data)
```

Table C.5: Cox Model with mdscore_{sum} as the main variable.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore$_{sum}$	-0.09	0.92	0.02	-3.49	<0.001	0.87 − 0.96
Smoke Status						
Current 1-15 c/d	0.45	1.57	0.10	7.08	<0.001	1.38 − 1.78
Current 16-25 c/d	0.85	2.34	0.16	12.67	<0.001	2.05 − 2.67
Current $\geq$ 26 c/d	1.44	4.21	0.33	18.20	<0.001	3.61 − 4.92
Former < 10 years	0.42	1.52	0.09	6.88	<0.001	1.35 − 1.71
Former 11-20 years	0.22	1.25	0.10	2.76	0.006	1.07 − 1.46
Former > 20 years	0.06	1.06	0.11	0.55	0.581	0.86 − 1.31
Cigar/Pipe, occasional	0.43	1.54	0.10	6.70	<0.001	1.36 − 1.75
Current/Former,missing	0.39	1.48	0.40	1.45	0.146	0.87 − 2.51
Unknown	-0.41	0.67	0.67	-0.41	0.684	0.93 − 4.74
BMI (kg/m^2)						
$[25, 30)$	-0.19	0.83	0.04	-3.51	<0.001	0.75 − 0.92
$\geq$ 30	0.03	1.03	0.07	0.43	0.666	0.91 − 1.17
Physical Activity						
Inactive	0.06	1.06	0.05	1.13	0.258	0.96 − 1.18
Moderately inactive	0.07	1.07	0.05	1.65	0.099	0.99 − 1.17
Active	-0.04	0.96	0.06	-0.54	0.588	0.84 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-1.67	0.096	1.00 − 1.00
Waist Circumference						
Abdominal Obesity	0.11	1.12	0.05	2.44	0.015	1.02 − 1.22
Educational Level						
None	0.019	1.02	0.04	0.46	0.642	0.94 − 1.10
Technical/Professional	-0.133	0.88	0.06	-1.87	0.062	0.76 − 1.01
Secondary	0.014	1.01	0.08	0.18	0.855	0.88 − 1.17
University	-0.173	0.84	0.06	-2.62	0.009	0.74 − 0.96
Unknown	0.182	1.20	0.21	1.06	0.291	0.86 − 1.68

c/d: cigarettes per day

Table C.6: Cox Model with mdscore_{sd} as the main variable.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore$_{sd}$	-0.05	0.90	0.02	-5.78	<0.001	0.86 − 0.93
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.94	<0.001	1.37 − 1.76
Current 16-25 c/d	0.84	2.31	0.16	12.43	<0.001	2.02 − 2.63
Current $\geq$ 26 c/d	1.43	4.14	0.33	17.94	<0.001	3.54 − 4.83
Former < 10 years	0.42	1.52	0.09	6.86	<0.001	1.35 − 1.71
Former 11-20 years	0.23	1.25	0.10	2.81	0.005	1.07 − 1.47
Former > 20 years	0.07	1.07	0.11	0.60	0.547	0.86 − 1.32
Cigar/Pipe, occasional	0.43	1.53	0.10	6.59	<0.001	1.35 − 1.74
Current/Former, missing	0.37	1.45	0.39	1.38	0.167	0.86 − 2.46
Unknown	-0.41	0.65	0.66	-0.42	0.673	0.09 − 4.66
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.42	0.001	0.75 − 0.93
$\geq$ 30	0.03	1.03	0.07	0.52	0.604	0.91 − 1.17
Physical Activity						
Inactive	0.05	1.05	0.06	0.98	0.327	0.95 − 1.17
Moderately inactive	0.07	1.07	0.05	1.60	0.110	0.98 − 1.17
Active	-0.03	0.97	0.06	-0.47	0.640	0.85 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-2.76	0.006	1.00 − 1.00
Waist Circumference						
Abdominal obesity	0.11	1.11	0.05	2.36	0.018	1.02 − 1.22
Educational Level						
None	0.02	1.01	0.04	0.47	0.639	0.94 − 1.10
Technical/Professional	-0.13	0.88	0.06	-1.79	0.073	0.76 − 1.01
Secondary	0.01	1.01	0.08	0.17	0.865	0.88 − 1.17
University	-0.17	0.84	0.06	-2.63	0.008	0.74 − 0.96
Unknown	0.20	1.22	0.21	1.14	0.253	0.87 − 1.71

c/d: cigarettes per day

Table C.7: Cox Model with mdscore_{ter} as the main variable.

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore$_{ter}$	-0.03	0.97	0.01	-4.74	<0.001	0.96 − 0.98
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.95	<0.001	1.37 − 1.76
Current 16-25 c/d	0.84	2.32	0.16	12.49	<0.001	2.03 − 2.64
Current $\geq$ 26 c/d	1.43	4.19	0.33	18.11	<0.001	3.58 − 4.89
Former < 10 years	0.41	1.51	0.09	6.81	<0.001	1.34 − 1.71
Former 11-20 years	0.22	1.25	0.10	2.78	0.005	1.07 − 1.47
Former > 20 years	0.06	1.06	0.11	0.58	0.563	0.86 − 1.31
Cigar/Pipe, occasional	0.43	1.54	0.10	6.65	<0.001	1.35 − 1.75
Current/Former,missing	0.37	1.45	0.39	1.38	0.168	0.85 − 2.46
Unknown	-0.44	0.65	0.65	-0.44	0.663	0.09 − 4.60
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.42	0.001	0.75 − 0.93
$\geq$ 30	0.03	1.03	0.07	0.51	0.612	0.91 − 1.17
Physical Activity						
Inactive	0.05	1.05	0.06	1.00	0.317	0.95 − 1.17
Moderately inactive	0.07	1.07	0.05	1.60	0.111	0.98 − 1.17
Active	-0.03	0.97	0.06	-0.50	0.615	0.85 − 1.10
Energy Intake (kcal/day)	0.00	1.00	0.00	-2.72	0.007	1.00 − 1.00
Waist Circumference						
Abdominal Obesity	0.11	1.11	0.05	2.37	0.018	1.02 − 1.22
Educational Level						
None	0.02	1.02	0.04	0.45	0.656	0.94 − 1.10
Technical/Professional	-0.13	0.88	0.06	-1.85	0.064	0.76 − 1.00
Secondary	0.02	1.01	0.08	0.21	0.837	0.88 − 1.18
University	-0.17	0.84	0.06	-2.63	0.009	0.74 − 0.96
Unknown	0.19	1.21	0.21	1.09	0.277	0.86 − 1.69

Appendix D

Output of Flexible Parametric Model

Next model is with mdscore_{cdf} as a continuous variable.

Table D.1: Flexible Parametric Model with a mdscore cdf as continuous variable

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
mdscore	-0.10	0.91	0.02	-5.29	<0.001	0.87 − 0.94
Smoke Status						
Current 1-15 c/d	0.44	1.55	0.10	6.95	<0.001	1.37 − 1.76
Current 16-25 c/d	0.83	2.29	0.15	12.45	<0.001	2.01 − 2.61
Current ≥ 26 c/d	1.43	4.17	0.33	18.23	<0.001	3.57 − 4.86
Former < 10 years	0.41	1.51	0.09	6.77	<0.001	1.34 − 1.70
Former 11-20 years	0.22	1.25	0.10	2.77	0.006	1.07 − 1.46
Former > 20 years	0.05	1.05	0.11	0.42	0.672	0.85 − 1.29
Cigar/Pipe, occasional	0.48	1.61	0.10	7.51	<0.001	1.42 − 1.83
Current/Former, missing	0.32	1.38	0.37	1.20	0.230	0.82 − 2.34
Unknown	-0.48	0.62	0.62	-0.47	0.635	0.09 − 4.42
BMI (kg/m^2)						
$[25, 30)$	-0.18	0.83	0.04	-3.45	0.001	0.75 − 0.92
≥ 30	0.03	1.03	0.07	0.52	0.602	0.91 − 1.17
Physical Activity						
Inactive	0.07	1.07	0.06	1.30	0.195	0.97 − 1.19
Moderately inactive	0.07	1.07	0.05	1.53	0.125	0.98 − 1.17
Active	-0.03	0.97	0.06	-0.47	0.638	0.85 − 1.10

Table D.1: (continued)

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
Energy Intake (kcal/day)	-0.00	1.00	0.00	-2.53	0.011	1.00 − 1.00
Waist Circumference						
Abdominal Obesity	0.10	1.10	0.05	2.15	0.032	1.01 − 1.20
Educational Level						
None	0.01	1.01	0.04	0.20	0.845	0.93 − 1.09
Technical/Professional	-0.11	0.90	0.06	-1.52	0.128	0.78 − 1.03
Secondary	0.01	1.01	0.08	0.20	0.841	0.88 − 1.17
University	-0.19	0.83	0.05	-2.90	0.004	0.73 − 0.94
Unknown	0.14	1.15	0.20	0.81	0.419	0.82 − 1.61
rcs						
rcs_1	0.69	1.98	0.21	6.37	<0.001	1.61 − 2.45
rcs_2	0.00	1.00	0.03	-0.07	0.941	0.93 − 1.07
rcs_3	-0.02	0.98	0.01	-1.84	0.065	0.96 − 1.00
rcs_4	-0.02	0.98	0.01	-3.01	0.003	0.97 − 0.99
rcs Center						
rcs Granada_1	0.06	1.07	0.03	2.26	0.024	1.01 − 1.12
rcs Granada_2	-0.02	0.98	0.01	-1.94	0.053	0.97 − 1.00
rcs Murcia_1	0.09	1.09	0.03	3.17	0.002	1.03 − 1.15
rcs Murcia_2	-0.02	0.98	0.01	-3.18	0.001	0.96 − 0.99
rcs Navarra_1	0.13	1.14	0.03	4.31	<0.001	1.07 − 1.21
rcs Navarra_2	-0.03	0.97	0.01	-3.12	0.002	0.96 − 0.99
rcs Gipuzkoa_1	-0.01	0.99	0.03	-0.28	0.778	0.94 − 1.05
rcs Gipuzkoa_2	0.00	1.00	0.01	0.09	0.926	0.98 − 1.02
rcs Sex						
rcs Female_1	-0.34	0.71	0.02	-11.55	<0.001	0.67 − 0.76
rcs Female_2	-0.07	0.93	0.02	-3.89	<0.001	0.89 − 0.96
rcs Age						
rcs $[40, 50)_1$	0.08	1.08	0.11	0.74	0.458	0.88 − 1.32
rcs $[40, 50)_2$	-0.02	0.98	0.02	-0.82	0.412	0.95 − 1.02
rcs $[50, 60)_1$	0.12	1.13	0.13	1.09	0.277	0.91 − 1.41
rcs $[50, 60)_2$	-0.05	0.95	0.03	-1.78	0.074	0.90 − 1.01
rcs $\geq 60_1$	0.08	1.09	0.12	0.73	0.468	0.87 − 1.36
rcs $\geq 60_2$	-0.10	0.90	0.04	-2.25	0.025	0.82 − 0.99

Table D.1: (continued)

	coef	**HR**	**S.E.**	**Z**	**p-value**	**95%CI**
constant	-2.52	0.08	0.01	-28.07	<0.001	0.07 − 0.10

c/d: cigarettes per day
rcs: restricted cubic spline

Appendix E

Stata code

E.1 Cox models fitted

The Cox model fitted in R project was also done in Stata, by the following code

```
use "C:\data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
generate msdcore_cent= mdscore_cdf - r(mean)
summarize msdcore_cent

* m0: With all the covariates (except mdscore)
stcox i.Smoke_Status i.BMI i.Physical_Activity Energy i.Waist_C i.
    L_School, strata(Center Sex Age) nolog nohr
stcox i.Smoke_Status i.BMI i.Physical_Activity Energy i.Waist_C i.
    L_School, strata(Center Sex Age) nolog
estat ic
estat phtest, log detail

* m.cdf: with mdscore cdf into quartiles
stcox i.cdf_c i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C i.
    L_School, strata(Center Sex Age) nolog nohr
stcox i.cdf_c i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C i.
    L_School, strata(Center Sex Age) nolog
estat ic
estat phtest, log detail
```

```
* m1b: with mdscore_cdf as continuous variable
stcox mdscore_cdf i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog nohr
stcox mdscore_cdf i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog
estat ic

* m2: with mdscore_sd variable
stcox mdscore_sd i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
    i.L_School, strata(Center Sex Age) nolog nohr
stcox mdscore_sd i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
    i.L_School, strata(Center Sex Age) nolog
estat ic

* m3: with mdscore_sum variable
stcox mdscore_sum i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog nohr
stcox mdscore_sum i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog
estat ic

* m4: with mdscore_ter variable
stcox mdscore_ter i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog nohr
stcox mdscore_ter i.Smoke_Status i.BMI i.Physical_Activity Energy Waist_C
     i.L_School, strata(Center Sex Age) nolog
estat ic
```

E.2 Flexible Parametric PH model fitted

```
use "C:\data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
summarize mdscore_cdf
```

```
generate cdf_cent = mdscore_cdf - r(mean)
summarize cdf_cent

* Selected model with mdscore_cdf as a continuous covariate
xi: stpm2 cdf_cent i.Smoke_Status i.BMI i.Physical_Activity Energy_cent
    Waist_C i.L_School,tvc(i.Center i.Sex i.Age) scale(hazard) df(4)
    dftvc(2) nolog eform
estat ic

* Selected model with mdscore_cdf as a categorical covariate
xi: stpm2 i.mdscore_c i.Smoke_Status i.BMI i.Physical_Activity
    Energy_cent Waist_C i.L_School, tvc(i.Center i.Sex i.Age) scale(
    hazard) df(4) dftvc(2) nolog eform
estat ic
```

E.3 Graphics

E.3.1 Baseline graphics

```
use "C:\data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
summarize mdscore_cdfw
generate cdf_cent = mdscore_cdf - r(mean)
summarize cdf_cent

* Creating stratum variables
egen strata=group(Center Sex Age)

* Kaplan-Meier curves
sts gen s0=s, by(strata)
separate s0, by(strata)

* Survival curves by Cox
stcox cdf_cent i.Smoke_Status i.BMI i.Physical_Activity Energy_cent
    Waist_C i.L_School, strata(Center Sex Age) nolog basesurv(surv0)
separate surv0, by(strata)
```

```
* Survival curves by Flexible Parametric PH model
quietly xi: stpm2 cdf_cent i.Smoke_Status i.BMI i.Physical_Activity
    Energy_cent Waist_C i.L_School, tvc(i.Center i.Sex i.Age) scale(
    hazards) df(4) dftvc(2)
predict sfp0, survival zeros
separate sfp0, by(strata)

* Graphic
line s016 surv016 sfp016 _t, sort
```

E.3.2 Kaplan-Meier and mean survival curves

```
use "C:\data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
summarize mdscore_cdf
generate cdf_cent = mdscore_cdf - r(mean)
summarize cdf_cent
quietly xi: stpm2 cdf_cent i.Smoke_Status i.BMI i.Physical_Activity
    Energy_cent Waist_C i.L_School,tvc(i.Center i.Sex i.Age) scale(
    hazards) df(4) dftvc(2)
predict xb, xbnobaseline
quietly stpm2 xb, scale(hazards) df(4)
keep if Center==2 & Sex==2 & Age==4
centile xb, centile(25 50 75)
generate cutpoints = .
forvalues j=1/3{
quietly replace cutpoints = r(c_‘j’) in ‘j’
}
xtile xbc4= xb, cutpoints(cutpoints)
forvalues j=1/4{
predict s‘j’ if xbc4==‘j’, meansurv
sts gen km‘j’ = s if xbc4==‘j’
}
line s1 s2 s3 s4 km1 km2 km3 km4 _t, sort connect(l l l l J J J J)
    lpattern(- - - - l l l l) xtitle("Years") ytitle("S(t)")
```

E.3.3 Conditional survival probabilities

```
use "C:\data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
summarize mdscore_cdf
generate cdf_cent = mdscore_cdf - r(mean)
summarize cdf_cent
quietly xi: stpm2 cdf_cent i.Smoke_Status i.BMI i.Physical_Activity
    Energy_cent Waist_C i.L_School,tvc(i.Center i.Sex i.Age) scale(
    hazards) df(4) dftvc(2)
generate t70 = 70
generate t75 = 75
predict s70, timevar(t70) survival
predict s75, timevar(t75) survival
keep if _t>=70

* Frequency table first
by Sex, sort : tabulate Center Age

* For Granada, the graphics of
*Granada, men, [50,60)
*Granada, men, >=60
*Granada, women, [50,60)
*Granada, women, >=60
*can be obtained.

* For example
keep if Center==2 & Sex==2 & Age==4
generate s7570 = s75/s70
histogram s7570
centile s7570, centile(50)
```

E.3.4 Survival probabilities across the risk spectrum

```
use "C:data.dta", clear
stset Age_Exit, failure(cens==1) id(epic_id) enter(Age_Recr)
summarize Energy
generate Energy_cent = Energy - r(mean)
summarize Energy_cent
summarize mdscore_cdfw
generate cdfw_cent = mdscore_cdfw - r(mean)
summarize cdfw_cent
quietly xi: stpm2 cdfw_cent i.Smoke_Status i.BMI i.Physical_Activity
    Energy_cent Waist_C i.L_School,tvc(i.Center i.Sex i.Age) scale(
    hazards) df(4) dftvc(2)
predict xb, xbnobaseline
quietly stpm2 xb, scale(hazards) df(4)
keep if Center==2 & Sex==1 & Age==1
summarize _t0
generate timevar= _n/1 in 34 / 80
forvalues j = 1 /9 {
local centile = ‘j’ * 10
quietly centile xb, centile(‘centile’)
local cxb = r(c_1)
predict s‘j’, at (xb ‘cxb’) survival timevar(timevar)
}
line s1 s2 s3 s4 s5 s6 s7 s8 s9 timevar, sort legend(off) lpattern(l ..)
    lwidth(medthin medthin medthin medthin thick medthin ..) xtitle("
    Years") ytitle("S(t)")

generate timevar= _n / 1 in 40 / 80
forvalues j = 1 /9 {
local centile = ‘j’ * 10
quietly centile xb, centile(‘centile’)
local cxb = r(c_1)
predict s‘j’, at (xb ‘cxb’) survival timevar(timevar)
}
line s1 s2 s3 s4 s5 s6 s7 s8 s9 timevar, sort legend(off) lpattern(l ..)
    lwidth(medthin medthin medthin medthin thick medthin ..) xtitle("
    Years") ytitle("S(t)")
```