Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Statistiche

Ciclo XXVIII

Settore Concorsuale di afferenza: 13/ D1

Settore Scientifico disciplinare: SECS-S/ D1

THE ANALYSIS OF SURVIVAL AND LONGITUDINAL DATA FROM LIFE-SPAN CARCINOGENICITY BIOASSAYS ON SPRAGUE-DAWLEY RATS

Presentata da: Daria Sgargi

Coordinatore Dottorato	**Relatore**
Alessandra Luati	Rossella Miglio

Esame finale anno 2018

Abstract

Carcinogenicity bioassay are among the best instruments to strengthen the evidence on which regulatory agencies vase their decision to classify harmful agents as human carcinogens, so they are fundamental to protect public health. The statistical analysis is fundamental to validate the results from carcinogenicity bioassay. This work aims to propose and illustrate some methodologies for the analysis of non-cancer outcomes, in particular for the analysis of time-to-death and of longitudinal measurements of body weights. The data from an old experiment were used for this purpose: 4 experiments aimed at testing the carcinogenic potential of Coca-Cola on Sprague-Dawley rats of different ages (randomized males and females of 7, 30, 39, 55 weeks of age, and their non-randomized offspring, observed since birth) were re-analysed.

Survival analysis aimed to verify the influence of the treatment, controlling for possible differences due to sex, age at beginning of observation and age of the dams at pregnancy. It was performed using Cox proportional hazards models for the rats of second generation, and accelerated failure-times models for those of first generation; the use of frailty terms was evaluated (univariate gamma frailty to account for unobserved heterogeneity applied to data from breeders; shared gamma frailty at the litter level applied to data from offspring).

The analysis of longitudinal body weights of the offspring was aimed at verifying the relevance of treatment, controlling for physiological differences due to sex and age of the dams at gestation. It was performed using linear and nonlinear mixed-effects models to handle the hierarchical structure of the data. Linear models were fitted using log-transformation of time and polynomial terms of order 3; nonlinear models consisted of growth models, in particular the Berkey-Reed model, that is usually used to analyse human growth during infancy, was applied.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Cancer is one of the leading causes of death worldwide, and despite the encouraging progress achieved during the last decades both in prevention and treatment of some types, it remains a major threat to public health. The burden of the disease is substantial and rising, due to increasing incidence rates, the growth and ageing of the population, and the spreading prevalence of risk factors like pollution, smoking, alcohol consumption, obesity and hypertension also in developing countries: these factors reverse the habit to see cancer as a problem regarding economically developed countries, and make it a global issue (Global Burden of Disease Cancer Collaboration, 2015). Given this framework, the importance of prevention is clear.

If we think about prevention in terms of public health, one of the most important issues is to identify the human carcinogens and to regulate their use. Carcinogens are all the biologic or synthetic substances, composites, technologies, occupational exposures and even lifestyles that may have a carcinogenic potential. It is necessary to identify the hazard, that is the capability of causing neoplastic effects under some circumstances; to assess the risk, defined as the actual carcinogenic effect expected if exposed to a cancer hazard, and understand the mechanisms of action.

Experimental and epidemiological studies are the best currently available instruments to test and verify these effects; in particular long term and life-span carcinogenicity bioassays are the main experimental tools when carcinogenicity is under exam. They usually employ rodents, in particular female and male mice and rats, treated with one or more exposure concentration of the tested substance and compared with untreated controls; the main route of administration to choose between, according to the characteristic of the agent, are inhalation, dermal and the most common, oral. The observation usually begins around 7 weeks of age, when the subjects have ended the weaning period and are completely independent, and lasts 104 weeks, or for the whole life; if the human pattern of exposure requests so, perinatal exposures can be evaluated, to verify the effects during the critical developmental phases of gestation and lactation.

The amount of information that a well-designed and well-conducted study can offer is impressive: their primary purpose is to characterise the carcinogenic properties of the agent,verifying if it might increase the age-specific incidence of malignant tumours, reduce its latency, intensify its severity or multiplicity (IARC, 2006) or, less directly, trigger a harmful effect in another agent or act in a combined way with it; they can also establish the existence of a dose-response relationship, identify target organs, and help to set a benchmark dose or a no-observed adverse effect level. Broad, strong and scientifically sound evidence is the base on which the organizations in charge (like the International Agency for Research on Cancer, the European Chemicals Agency, or the Environmental Protection Agency, the Food and Drugs Administration or the National Toxicologic Program for the United States) can classify dangerous agents as human carcinogens. The codification can in turn boost the action from regulatory agencies, that may decide to regulate or ban the use of these substances in order to protect and promote human health.

Since the early 1980s most of the agencies involved in the classification and regulation of dangerous substances collected the most up-to-date and agreed-on options and procedures in experimental design, conduct of the study, reporting and analysis of the results, and formalized them into guidelines and regulations, that have been enriched and reviewed in the light of scientific progress, advance of procedures and consideration for the animal’s welfare (OECD, Test No. 451: Carcinogenicity Studies 2009). The main purposes are to grant and even out the quality of the bioassays they base their knowledge on, and to assure the comparability of results. Even if today almost every agency has specific protocols, a common reference can be recognized in the guidelines drafted by the Environment, Health and Safety Program and the Working Group on Testing from the Organisation for Economic Cooperation and Development. Their *Guidances for testing the chemicals* and the related documents represent the most comprehensive collection of procedures, and among the very few to explicitly treat in detail and depth the statistical analysis of data (Hothorn 2014).

Statistical analysis is universally recognised to be integral part of studies: an appropriate experimental design is the cornerstone to answer the research question, and statistical evaluation of the data is the necessary complement to establish and quantify whether the exposure to the selected agent is associated with adverse effects. Between the different "schools of thought", the classical frequentist approach and the concept of hypothesis testing have been chosen to maintain coherence with most of the work done in toxicology.

One of OECD's guidelines (OECD 2012) is partly devoted to illustrate and explain in depth how to design and conduct the appropriate statistical tests based on the kind of experiment, its objective and the type of data; it also helps to interpret the results and to understand their real meaning and relative importance. The methods are systematically organized into a flowchart: according to the nature of the data, the appropriate "branch" of the tree is chosen, and omnibus tests to highlight overall differences between groups or tests for linear trend are suggested, preceded by several tests to verify the respect of their assumptions. In some cases when these are not met and it is possible to transform the data, a circular path is proposed and the tests repeated.

What stands out is that consolidate and advanced methodologies exists and are routinely used for the analysis of tumor incidence, which is for the direct assessment of carcinogenicity. As it was previously highlighted, nevertheless, a good designed and performed study produces a great amount of additional information such as the body weights, food and beverages consumption, or the time of survival in life-span studies. Their importance to have a picture of the health status of the animals involved and the good progress of the experiment is clear, but this potential at the moment is not fully exploited, since even the most detailed guideline proposes quite superficial analysis like those illustrated in the OECD flowchart.

Today many statistical methodologies exist to treat data of this kind, and their application to carcinogenicity bioassays may allow to fully use all available information and to integrate them, to obtain a richer knowledge of the effects of the tested compound. The idea for this work started from the will to capitalise the potential of the dataapplying different methods to experimental studies, to verify if they can shed a new light on data and contribute to understand and establish the effect of the tested substances on health.

The research was promoted by a local research centre needing help to perform routine statistical analysis and looking for ways to exploit the potetial coming from their experience in carcinogenicity studies. The Ramazzini Institute was founded as a social cooperative in 1987 but is active since the early ‘70s thanks to the work of Prof. Cesare Maltoni, who gathered the link between living and working environment and cancer, and performed life-span bioassays in the laboratories in Bentivoglio (Bo) to understand the mechanisms of action of the disease, and identify and quantify on experimental base the toxic and carinogenic potential of widespread substances. During the decades dozens of experiments contributed to the identifications of many carcinogens.

This work takes up an already published experiment again, the Coca-Cola experiment, that didn’t show ashtonishing results in carcinogenicity tout court, but could present some interesting features in other measures that were not analysed in depth at the time; in the following chapter the experimental design of this bioassay and its results, as were published, will be presented, together with a brief exploratory analysis. The third and the fourth chapters are dedicated to specific topics, and represent the core of the research: survival analysis and the analysis of longitudinal recordings of body weights, respectively. Each will start with a recognition of the literature, then present the data, the models and analysis that were performed, and the results obtained. The research will end with a discussion of the work done, and some conclusion will be drawn.

Chapter 2 The Coca-Cola study

Diet can become one of those life habits to ease the occurrence of cancer in different ways: some food components have the direct ability to induce the disease; some contain potentially dangerous additives to improve their stability, preservation or even just the appearance; or in general, the excessive caloric intake can lead to overweight and obesity, that has been identified as a risk factor for some types of cancer (Calle E.E. 2003) (RAPP 2005). Given the impossibility to study diet as a whole, experimental studies have been used to evaluate the impact of single nutrients or food components.

Coca-Cola is a widespread product among the population of any age, socio-economic status and country, its concentration of sugar and the caloric power are very high: for these reasons, the "Cesare Maltoni" Cancer Research Centre of the the Ramazzini Institute[1] decided to evaluate its possible association with tumour incidence in rodents. Starting from 1986 several sub-studies were performed, each involving rats or different ages, males and females, randomized in two groups, the treated and the controls.

The design and the conduction of the bioassay will be now illustrated along with the results, as they were originally published. In order to have a clearer picture of the available data, the last part of the chapter will contain also a first exploratory analysis.

[1] At the time named European Ramazzini Foundation

2.1 The experimental design[2]

To verify the long-term influence of strong Coca-Cola consumption on the development of tumours, the bioassay planned it to be administered to Sprague-Dawley rats as a substitute of drinking water for the whole life span, until spontaneous death of the animals.

Treatment	Age at start	M	F
Coca-Cola	7 weeks	80	80
Drinking water	7 weeks	100	100
Coca-Cola	55 weeks	70	70
Drinking water	55 weeks	70	70
Coca-Cola	Offspring (55)	28	24
Drinking water	Offspring (55)	32	24
Coca-Cola	30 weeks	55	55
Drinking water	30 weeks	55	55
Coca-Cola	Offspring (30)	74	73
Drinking water	Offspring (30)	110	98
Coca-Cola	39 weeks	110	110
Drinking water	39 weeks	110	110
Coca-Cola	Offspring (39)	67	65
Drinking water	Offspring (39)	49	55

Table 1: Experimental plan of the four bioassays performed for the project

The experimental plan is summarized above in Table 1: it wanted to account for possible differences in the metabolism and the mechanism of action linked to different ages, therefore four different experiments were conducted, involving respectively female and male breeder rats of 30, 39 and 55 weeks of age and all their offspring born in all litters, treated since prenatal life, and female and male non-breeding rats of 7 weeks of age. All animals were bred from the internally grown colony, formed at the beginning of the laboratories' activity in the '70s.

All breeders were identified, separated by sex and assigned to each experimental group, so that no more than one female and one male belonging to

[2] All information and data in this and the following paragraph come from the original publication of Belpoggi F. et al. **Specificata fonte non valida.**.

the same litter were in the same one; they were housed in groups of five in makrolon cages until scheduled time of mating and starting of the treatment, then placed in breeding cages for a week. Females were housed individually for the whole period of the pregnancy and the weaning, and then went back to the regular housing for the rest of the experiment; all pups from every litter after the weaning period were identified, separated by sex and assigned to the same treatment group as their breeders. Cages belonging to both treatments were kept in the same room, under the same living conditions and with the same diet. The treatment consisted in substituting drinking water with ad libitum Coca-Cola, that was supplied every 2 weeks by an Italian retailer and was mechanically shaken before subministration to eliminate CO_2, and it lasted until spontaneous death.

During the observation, the animals were checked daily for physical and behavioural problems; the individual body weight and mean beverages and food consumption for cage were measured weakly for the first 13 weeks of observation, every 2 weeks until week 104, and every 8 weeks later, while complete examinations to check and report everything about the health status were performed weekly then every 2 weeks for the whole life.

After death, complete necropsy was done on every subject, all pathological lesions and all organs and systems[3] underwent histopathology, were preserved in 70% ethyl alcohol (bones were fixed in 10% formalin and decalcified), trimmed and processed as paraffin blocks. 3-5 µm sections of every block were sliced and coloured with haematoxylin-eosin, and finally examined microscopically by a group of pathologists; a senior pathologist supervised and reviewed all tumours and lesions of neoplastic interest.

The statistical analysis was performed on tumour incidences only using a χ^2 test to evaluate the significance of differences between treated and controls.

[3] Skin and subcutaneous tissue, the brain, pituitary gland, Zymbal glands, salivary glands, Harderian glands, cranium (with oral and nasal cavities and external and internal ear ducts, 5 levels), tongue, thyroid and parathyroid, pharynx, larynx, thymus and mediastinal lymph nodes, trachea, lung and mainstem bronchi, heart, diaphragm, liver, spleen, pancreas, kidneys and adrenal glands, oesophagus, stomach (fore and glandular), intestine (4 levels), bladder, prostate, uterus, gonads, interscapular fat pad, subcutaneous and mesenteric lymph nodes, and any other organ or tissue with pathological lesions.

2.2 Results

It was chosen to aggregate data from all breeders, and to consider all offspring together (animals who started treatment since embryonic life and the group observed since the age of 7 weeks).

Fluid consumption differed markedly between treated and control animals, the former drinking twice the amount of the latter; conversely, food consumption was almost 40% minor among the Coca-Cola consumers. Body weight was higher both in breeders and offspring, for both sexes; no significant difference in survival was observed, just a slight decrease in female offspring.

The following differences were observed in tumour occurrence:

1. A slight increase in malignant tumour incidence;
2. A statistically significant ($P < 0.01$) increase in malignant mammary tumours and total mammary tumours, both in breeders and offspring females;
3. A statistically significant increase in adenomas of the exocrine component of pancreas in male ($P < 0.01$) and female ($P < 0.05$) breeders, and in male ($P < 0.01$) and female ($P < 0.01$) offspring; No exocrine carcinomas were observed.
4. An increase in islet cell carcinomas was observed between female breeders and offspring; although it's not statistically significant, it should be noted that looking at the historical controls, only one (0.04%) islet cell carcinoma out of 2274 untreated females was observed.

They finished concluding that the significant rise in the occurrence of mammary gland tumours could confirm a correlation between higher body weight and increased risk of mammary cancer, and that the relatively high number of pancreatic islet cell carcinoma compared to the historical controls shall not be underestimated, even if not significant. Therefore, even if the human consumption is rarely as extreme as the one designed in the experiment, it was confirmed that an abuse of high caloric/ high sugar beverages like soft drinks can ease overweight and obesity, that in turn is a risk factor for human health and cancer.

Chapter 3 The analysis of survival data

When referring to “survival data” in carcinogenicity studies, the narrower interpretation is the correct one, since they literally indicate the time until death. In more common 2-years or chronic experiments, we find under this label the duration of life of the animals deceased before the official time of cessation, that will be those who experienced the “event of interest”, while the surviving ones will be treated as censored; in way less usual life-span studies the meaning is the same, except that there will be no censored observations, since the experiment lasts until the spontaneous death of the last animal.

The duration of life is often under scrutiny in epidemiological studies and in clinical trials, but is rarely the outcome of primary interest in carcinogenicity bioassays, that are aimed to test the incidence of neoplastic lesions under different doses of treatment. Still, these data are collected and analysed in any kind of study since they allow to keep the general trend of the experiment under control during its execution, verifying that the tested compound has not such a high toxic potential to reduce too much the treated groups. The excessive decrease of sample size can threaten the sensitivity of the analysis, and increase the likelihood of obtaining false positives and false negatives (WHO 1987).

When studying cancer these data acquire even more relevance for the nature of the phenomenon, since the probability of developing neoplastic lesions increases with age. Moreover, if the tested substance has the power to affect the survival for toxicity not related to tumour, statistical analysis of incidence may result biased: if the treatment reduces excessively the duration of life, it can underestimate the carcinogenic potential; conversely, if it increases the longevity, the effect on cancer development may be overestimated (OECD 2012).

These considerations led regulatory authorities and journals to request an evaluation of the differences in survival for every study to be considered, and to include indications about how to perform these analyses in the guidelines and the protocols.

3.1 The analysis of time-to-event data according to the guidelines

As it was mentioned in the introduction, guidelines drafted from the most authoritative national and international agencies have become an essential instrument for research centres, enhancing the possibility of producing quality research that can contribute to form the scientific base for the regulation of harmful agents. The documents (OECD 2012) (OECD 2009) from the *Guidelines for testing the Chemicals* Series drafted by Organization for Economic Cooperation and Development will again be the main reference here, since they are the most detailed and comprehensive in the identification of statistical procedures and methodologies.

The first indication they give, is that the differences in survival need to be explored to guarantee the presence of a sufficient number of individuals to preserve the power of the tests to be performed, and to verify whether the analysis of incidence will have to be corrected. Survival data represent times, in this context usually days since the starting of the observation or weeks of age, and have some particular features: their distribution is not symmetric but tends to be positively skewed, and they are often censored - mostly right censored in the context of randomized bioassays. These peculiarities prevent the use of standard methods for continuous data.

The suggested steps to identify differences or dose-response relationships in survival are totally based on non-parametric tests like the Mantel-Cox test, the generalized Wilcoxon or Kruskal-Wallis test, and the Tarone trend test. The preferred methodology to compare the duration of lives among the experimental groups consists in estimating the survivor function using the product-limit method, proposed by Kaplan-Meier in 1958. It is based on the construction of a series of time intervals, such that each contains at least one event (that is assumed to occur independently from all others), and the event time is taken as beginning of the interval; the probability of survival is calculated for each interval as the ratio of subjects surviving and subjects at risk. The estimate is the product of the probabilities calculated on all time intervals, and it can be represented graphically as a step function, with the estimate probabilities remaining constant between adjacent death times, and decreasing at each event (Collett 2003). Different curves can be estimated for experimental groups, and differences are to be formally tested using the log-rank test.

The non-parametric approach suggested in the guidelines is very appealing, given its simplicity and the possibility of representing the estimates graphically, that allows to catch many information about the data at a glance. This is useful, and may be sufficient in this kind of experiments, where the subjects under study are randomized and share every condition (type of caging, room, temperature, diet, number and scheduled timing of checking, …) except for treatment. In some situations, nevertheless, we might be interested in obtaining an estimate of the treatment's effect on survival, or in the effect of other factors: here this univariate approach is a useful exploratory tool, but more sophisticated methods are necessary.

Let's take the Coca-Cola study as an example: the consumption of this compound is so widespread that we likely would not expect a strong, direct influence on survival. However, for its characteristic (highly sweetened, highly caloric) we may expect it to have an influence on body weight and mass composition, so that a heavy consumption like the one outlined in this experiment might ease overweight or even obesity on treated rats. These conditions may reasonably affect survival, but we are not able to account for and estimate their effect.

Another feature of this experiments is that they involved adult rats and their offspring, to verify effects of perinatal exposure: all alive pups from all litters were included, each in the same treatment regime as their breeders, so randomization was avoided in the second generation of exposed. This poses a problem of “familiarity”: individuals are more likely to share characteristics and propensity for diseases, so we expect to easily find similar paths among animals from the same litter.

Finally, when plotting the Kaplan-Meier estimates of the survivor or the hazard functions, we may obtain crossing step functions: very close functions with similar paths might mean that relevant differences among groups do not exist, while non-parallel curves of the logarithm of the hazard functions might pose a doubt on the proportionality of the hazard. When there are reasons to question the assumption of proportional hazard, the Wilcoxon test might be more suitable than the log-rank test to assess differences among groups, but a modelling approach could be much more informative and appropriate.

3.2 Other methods for survival analysis:

3.2.1 The Proportional Hazard model

To take advantage of all information and to give thorough answers to the research questions, we need to leave the hypothesis testing approach and choose the modelling approach.

Survival data can be modelled to explore the risk of death at any time after the beginning of the study: the hazard function is the object of interest, representing the instantaneous probability of death at the time, given that the subject has survived until that time. The aim is to determine which conditions affect the form of the hazard function, and to obtain the estimate of the function itself for the individuals. The most commonly used multivariate approach is the proportional hazard model, proposed by Cox (1972), that has lately become quite popular also in carcinogenicity studies thanks to its flexibility and the diffusion of statistical packages of easy use. The Cox model is necessary for the use and understanding of many of the methodologies that will be proposed later, but is not the main topic of this section, so it will be just briefly introduced here.

The proportional hazard model explains the hazard of death at any time t as depending on the values x_1, x_2, …x_p of p explanatory variables recorded at the time origin of the study, X_1, X_2, … X_p:

$$h_i(t) = \exp(\eta_i)\, h_0(t), \qquad \eta_i = \sum_{j=1}^{p} \beta_j x_{ji}$$

where $h_0(t)$ is the baseline hazard function, representing the hazard at time t for a subject for whom all values of the covariates are 0, andη_i is the linear component of the model, containing the explanatory variables $x_p$and the respective coefficients β_p. The estimated effect can be interpreted in terms of hazard ratios ($exp(\beta_i)$): if greater than one, it indicates that the hazard of the event is positively associated with the corresponding covariate. This is one of the most important advantages of the modelling approach: it allows to consider the influence of several risk factors at one time. These are typically categorical, ordinal or continuous covariates, whose values is recorded at the beginning of the observation, but several expansions have been studied in time, so that a better representation of the phenomena under study could be reached.

The essential features of this model are that the baseline hazard is estimated non-parametrically using the maximum likelihood method, so no assumption is made on the distribution of survival times (thus the model is defined *semi-parametric*), and that its formulation assumes the proportionality of the hazards: since the covariates act multiplicatively on the hazard at any time, the hazard in any "group" is a constant multiple of the hazard in any other, and therefore the survival curves should never cross. It is fundamental to assess the adequacy of the model: this includes to verify whether all and only the appropriate explanatory variables have been included in the model, and that the correct functional form has been used; to check for the presence and the nature of extreme values; and of course, to verify the assumption of proportional hazard. Visual inspection of the data is very useful, but must be supported by diagnostic of the residuals.

3.2.2 Accelerated failure-times models

The analysis of graphical representation of the data and the diagnostic of the residuals are fundamental and should never be neglected, to assess the validity of the models built. One of the aspects that should always be evaluated is the tenability of the condition of proportionality of hazards, since it is one of the assumptions on which the Cox model and other parametric models for survival are based. Very often indeed real data do not behave this way: in these cases, alternatives are to be considered.

Models that do not require hazard to remain constant are the accelerated failure time models: the basic idea is that the effect of covariates or risk factors is constant in time and multiplicative on the time scale (instead that on the hazard scale, as in PH models), accelerating or decelerating the event of interest.

In the most generic specification, the survivor function can be written as

$$S(t) = S_0\ (\varphi t)$$

where $S_0\ (t)$ is as usual the baseline hazard function, and φ represents the acceleration factor, dependent on the covariates:

$$\varphi = \exp\left\{\sum\nolimits_{j=1}^{p} \beta_j\ X_{ij}\right\}$$

The model is often expressed in its logarithmic form with respect to time,

$$\log(T_i) = \ \beta_0 \sum\nolimits_{j=1}^{p} \beta_j X_{ij} + \ \sigma\ \varepsilon_i$$

where β_0 is an intercept, β_j are the coefficients of the p covariates X, σ is a scale parameter and ε_i is a random variable that models the deviations of $\log(T_i)$ from the linear part of the model; to the distribution of ε_i corresponds a distribution of the survival times T_i(Bradburn, 2003). The most used specification of the AFT model is indeed parametric, and many distributions can be used to best represent the time to event: exponential, Weibull (that can be used both as parametrization of the PH model and of AFT model), the log-logistic, log-normal or generalized gamma. Non-parametric specifications also exist, the most known being the Buckley and James method (Buckley I. V. 1979), but it is rarely used in real data analysis for its problems in theoretical justification and robustness underlined by Wei (1992) and for the complexity of the computations.

In AFT models, the effect of the covariate is measured on the survival times: this makes the interpretation of the results more immediate. The magnitude of the effect of covariates is not given in terms of hazard ratio, but of time ratio: if we divide the population in groups according to the level of a variable of interest, the time ratio stands for the ratio of the expected survival of a group with respect of the reference group. When >1, the covariate “slows down” the appearance of the event, while if it’s <1, it accelerates death. The model is fitted using the maximum likelihood method.

Accelerated failure time models represent a valuable alternative to the Cox model when the assumption of proportional hazard is not realistic; nevertheless, they are not yet extremely diffused in medical and biological research as they are in the engineering field, where they are widely applied for reliability studies. Some of its characteristic make them appealing: as already mentioned, they are in many contexts more realistic, the parameters are more immediate to interpret, and choosing the wrong distribution affects the estimates less (Lambert, 2004) (Keiding, 1997)

3.2.3 Unobservable heterogeneity and familiarity problem: univariate and shared frailty terms

The methods for survival we have revised so far implicitly assume that individuals in the population are homogeneous; the analysis are performed to verify if differences exist and what determines diverse hazards. It can be the risk factor of interests, such as the tested compound, sex, age at the beginning of the treatment, but it is likely that other factors have an impact on the duration of life.

Unlike in linear regression, it is very important to be sure that all relevant covariates are included in survival models in order to have unbiased coefficients and hazard rate. This is because the hazard rate is time dependent: suppose a characteristic divides the population in a low-risk group and a high-risk group. If the characteristic is observed and modelled, two constant risks functions are estimated, one for each group; but if this variable is omitted, we will have one, common hazard function, decreasing in time since individuals from the high-risk group fail before than the others. When other covariates are included in the model, the effect of the omitted one will therefore alter the estimates of the parameters, because its distribution for each value of the included covariates varies with time. This was demonstrated in, among others, Gail et al. (1984), Schumacher et al. (1987), Bretagnolle and Huber-Carol (1988), Hougaard et al. (1994) Schmoor and Schumacher, 1997; Chastang et al, 1988. Bretagnolle and Huber-Carol also investigated the direction of the bias: their simulations showed that, no matter the number of covariates included in and omitted from the model, the effect of the observed variables will be underestimated, and the asymptotic bias changes the risk for confidence intervals from 5 to 50%.

Very rarely nevertheless it is possible to observe all risk factors that affect survival. Experiments such as carcinogenicity bioassay have less problems than observational studies, thanks to randomization and careful experimental design, but still there are some sources of unobserved heterogeneity: sometimes it can be too expansive in terms of time or resources to measure and report all relevant risk factors with precision and completeness. This is the case with data from the Coca-Cola experiments: basic information on the health status of the animals was observed and registered on a weekly basis on paper supports (we must remember that the experiment was carried out during the 80s): these data are still available, but they were not translated on informatic files since they are not usually used for analysis, so the information regarding the health of animals is essentially lost. This is the reason for the use of mixed effects in survival analysis: the idea is to decompose the variability of survival in two sources: a predictable part, measured by the coefficients of observed risk factors, and an unknown part, not directly observable, described by a *frailty term*. The concept was first elaborated by Greenwood and Yule in 1920, and developed after decades in Clayton (1978) and Vaupel et al (1979): individuals have different not observable characteristics that affect their survival probability, defined frailties, the more frail will survive less than the less frail, creating a kind of selection that can give a distorted picture of the underlying process.

Univariate frailties are random variables whose distribution reflect the nature of the relationship between the subject's and population's survival:

$$h(t, Z) = Z\, h_0(t)$$

They are time independent and act multiplicatively on the baseline hazard function, describing the unobservable heterogeneity among individuals (the variance of the frailty distribution determines the level of heterogeneity in the population: $E(Z) = 1$and $V(Z) = \sigma^2$, if σ^2 is small Z tends to 1 and the population is quite homogeneous).

The survivor function is defined as

$$S(t \mid Z) = \exp\left\{-\int_0^t h\,(s, Z)\; ds\right\} = \exp\left\{-Z\int_0^t h_0(s)\; ds\right\} = \exp\{-Z\, H_0(s)\}$$

Frailties can of course be applied to the classic proportional hazard model containing other covariates, obtaining, for each subject, a hazard function of the form

$$h(t, Z, X) = Z\, h_0(t) \exp\left\{\sum\nolimits_{j=1}^{p} \beta_j x_{ji}\right\}$$

Hougaard (1984, 1986a, 1986b) demonstrated that the survivor and density functions for the whole population, the only observable entities, and the mean and variance of the frailty are characterized using the Laplace transform of the frailty distribution as follows:

$$S(t) = E\, S(t|Z) = E \exp\{-Z\, H_0\,(t)\} = L(H_0(t))$$

$$f(t) = -h_0(t) L'\big(h_0(t)\big)$$

$$E\, Z = -L'(0)$$

$$V(Z) = L''(0) - (L'(0))^2$$

These formulations clarify why it is important to choose a distribution for the frailty that has an explicit Laplace transform, and that maximum likelihood methods can be used for the estimation of regression parameters. The hazard function for the population becomes

$$
\begin{aligned}
h(t) &= E\ (\ h(t,Z) | T > t) \\
&= \int_0^\infty h(t,Z)\ f(z\ |T > t)\ dz = h_0(t) \int_0^\infty z\ f(z\ |T > t)\ dz
\end{aligned}
$$

or, taking into account the presence of covariates,

$$h(t \mid X) = h_0(t) \exp\left\{\sum\nolimits_{j=1}^{p} \beta_j x_{ji}\right\} E\ (Z \mid T \geq t, X)$$

This explains that, since the frailer subjects die earlier, the average frailty of the survivors is not constant, but will decrease in time.

Many types of distribution are possible for univariate frailties, usually parametric, like the Gamma distribution, the most widely diffused thanks to its mathematical properties: it’s always positive and has simple Laplace transform that allows to easily derive in closed forms the measures of interest, and most of all it is flexible, since it can take various shapes. Gamma-distributed frailty terms have been repeatedly used to model univariate and multivariate frailties, also in the analysis of survival in elderly population, so they have been considered for the analysis of these life-span experiments. Other possible parametric specifications are the positive stable, the inverse gaussian, the extended family of power variance function distributions, the lognormal and the compound Poisson, not suitable here since it creates a subgroup with $Z= 0$ that does not experience the event. If no information is available on the trait influencing the hazard among groups, the discrete specification is possible, both binary or finite discrete. The choice should be based on the knowledge of the phenomenon under study, but often simple mathematical convenience guides the selection.

The Coca-Cola experiments have another peculiarity that can be addressed using frailty terms: young and adult rats were randomized using systematic sampling and assigned to treatment or control groups. After one week, they were also assigned for mating, and all the pups from all litters were included in the experiment, continuing the regime they were given during pregnancy and weaning through their dams.

Frailties are extremely useful also in this case, since they can be used to model dependence in cases where the implicit assumption of independence of each subject ad their failure times does not hold. This can happen when analysing recurring events, in studies where one patient act as his own paired control, or in studies where familiar groups are involved. The simplest option to overcome these independence issues are shared frailties: the idea here is that some of the unobservable characteristics that may affect the risk of failure are common among individuals belonging to the same cluster, like siblings, and they can influence the effect of observed covariates on survival. Here the random variable $Z$is constant in time and is associated to the group, instead of individuals, accounting for dependence among the subjects.

The formalization of shared frailty models is due to Clayton (1978), and can be found also in Therneau and Grambish (2000) or Hougaard (2000). The survival times are independent conditional on the frailties, and the hazard takes the form

$$h(t) = Z_i h_{0j}(t) \exp\left\{\sum\nolimits_{j=1}^{p} \beta_j x_{ji}\right\}$$

where $h_{0j}(t)$ is as usual the baseline hazard, that can be derived semi-parametrically or parametrically, $\beta_j$are the regression parameters linked to the fixed effect of observable covariates and the frailties are defined by the parameters of their distribution, are independent between clusters and identically distributed, and shared by the individuals belonging to the same cluster. This way, the survivals are conditionally independent with respect to the frailties, and dependent inside the group, thanks to the frailty term. Supposing we had just two clusters, we would get

$$\begin{aligned} S(t_1, t_2|Z) = {} & S_1(t_1)^2 S_2(t_2)^2 \\ & = \exp\{-Z\, H_{01}t_1\} \exp\{-Z\, H_{02}t_2\} \ = \exp\left\{-Z\ \sum\nolimits_{i=1}^{2} H_{0i}t_i\right\} \end{aligned}$$

where $H_{0i}(t) = \int_0^t h_{0i}(s)\, ds$. Again, the Laplace transform is extremely relevant to obtain the marginal survivor function:

$$S(t_1, t_2) = L(H_{01}(t_1) + H_{02}(t_2))$$

As anticipated before, the standard distribution assumed for shared frailties is the Gamma, with mean 1 and variance σ^2, so the marginal survivor is denoted as

$$\begin{aligned} S(t_1, t_2) &= L(H_{01}(t_1) + H_{02}(t_2)) \\ &= \left(1 + \sigma^2 \left(H_{01}(t_1) + H_{02}(t_2)\right)\right)^{-\frac{1}{\sigma^2}} \\ &= \left(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1\right)^{-\frac{1}{\sigma^2}} \end{aligned}$$

The extension to multivariate case, with more than two clusters, is attributable to Cook and Johnson (1981), that defined the survivor function as

$$S(t_1, \dots, t_p = \left(\sum\nolimits_{i=1}^{p} S_i\,(t_i)^{-\sigma^2} - p + 1\right)^{-\frac{1}{\sigma^2}}$$

with equal correlation among the survival of the individual belonging to the same group. This specification seems appropriate in such a case, where multiple litters are included, each composed by more than two siblings.

An estimate of the frailty term within each litter can be found (Nielsen et al. 1992) as

$$\widehat{z_i} = \frac{1/\sigma^2 + \sum_{j=1}^{n_i} \delta_{ij}}{1/\sigma^2 + \sum_{j=1}^{n_i} \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}\, H(\,t_{ij})}$$

The properties of shared frailties survival models were demonstrated, both adding other covariates or not, in several works: Murphy (1994) showed its consistency and asymptotical normality (1995), Giddens (1999) tested for gamma distributions with the semiparametric survival function, and Cui and Sun (2004) demonstrated graphically and with numerical methods the adequacy of the Gamma distribution.

Of course, the shared frailty approach has its limitations, and is not sufficiently flexible in several contexts: for example, it assumes that unobservable risk factors are equal among all individuals in each group, and that the association among subjects is positive in most of the cases (Xue and Bookmeyer, 1996). Since the characteristics of rats belonging to the same litter are extremely similar, there will be no need to consider more complex specifications for the multivariate frailty, like for example the correlated frailty model (where within each group a frailty term is associated to each individual: these random variables are positively or negatively associated jointly distributed, and different distribution can be chosen to model different situations, like again the gamma, the log-normal or the compound Poisson).

3.3 Analysis of time-to-event data from the Coca-Cola study

As it was presented in the previous chapter, the Coca-Cola study involved randomized Sprague-Dawley rats of different ages, as well as their offspring, that was included in the observation in toto and without randomization. For all subjects the event of interest is spontaneous death, since these are life-span carcinogenicity studies and no terminal sacrifice was planned; subsequently, no censored observation exists, all subject experience the event. The outcome, then, is time to death, measured in weeks of age of each subject. All information about the identification of individuals were recorded at the beginning of the study: they are presented in the following table, and they represent the variables of interests, and the main other possible risk factors that we want to control:

Variables	Description	Values
Entry	*Weeks of age at the beginning of the observation*	*7 weeks; 30 weeks; 39 weeks; 55 weeks*
Age	*Duration of life (in weeks)*	
Event	*Spontaneous death*	*1 deceased*
Treatment	Experimental regime	0 drinking water (control); 1 Coca- Cola (treated)
Sex	Sex	0 female; 1 male
Momstart	Age of the dam at pregnancy (only available for offspring)	30 weeks; 39 weeks; 55 weeks
Famid	Identifier for litters, common for all siblings born from the same breeders	1-98

Table 2: Variables for survival analysis; "setting" variables in italic, experimental variables in normal font.

3.3.1 Descriptive analysis

Here a brief summary of the data and some really basic descriptive statistics are reported:

Treatment	Age at start	M	F
Coca-Cola	7 weeks	80	80
Drinking water	7 weeks	100	100
Coca-Cola	55 weeks	70	70
Drinking water	55 weeks	70	70
Coca-Cola	Offspring (55)	28	24
Drinking water	Offspring (55)	32	24
Coca-Cola	30 weeks	55	55
Drinking water	30 weeks	55	55
Coca-Cola	Offspring (30)	74	73
Drinking water	Offspring (30)	110	98
Coca-Cola	39 weeks	110	110
Drinking water	39 weeks	110	110
Coca-Cola	Offspring (39)	67	65
Drinking water	Offspring (39)	49	55

Table 3: Composition of experimental groups; breeders and their offspring are listed sequentially.

The experimental groups of breeders were formed to be balanced in terms of sex and treatment regime, but not in terms of ages groups. The offspring, since had not been randomized, present slightly less homogeneous characteristics, as it might be expected the pups born form older dams are fewer. A marginal analysis was used to highlight associations between the outcome and explanatory variables (Box-and-Whiskers plot and density plots are reported in Figure1).

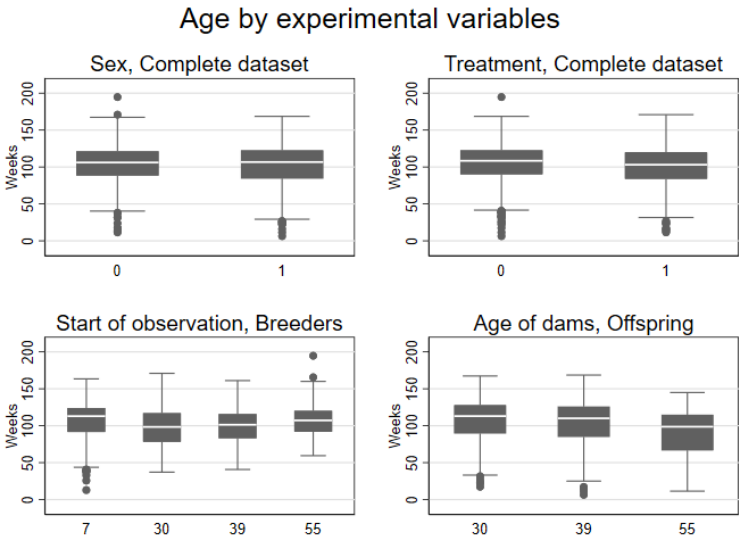


Figure 1: Age by experimental variables, Box-and-Whiskers plots

The Box-and-Whisker plots don’t show significant alterations in the survival depending on sex, and the exposition to treatment seems to have a negligible effect, too; the age at beginning of the observation and the age of the dams at pregnancy, on the other hand, determine more differences; these are confirmed when estimating the Kaplan Meier survival curves. Log-rank and Wilcoxon test were performed to verify the equality of survivor functions; they are not reported here for brevity, but they confirm the statistical significance of the differences among the curves calculated for animals entering in the experiment at different ages, for different ages of the dams, and surprisingly between rats belonging to treated and control group.

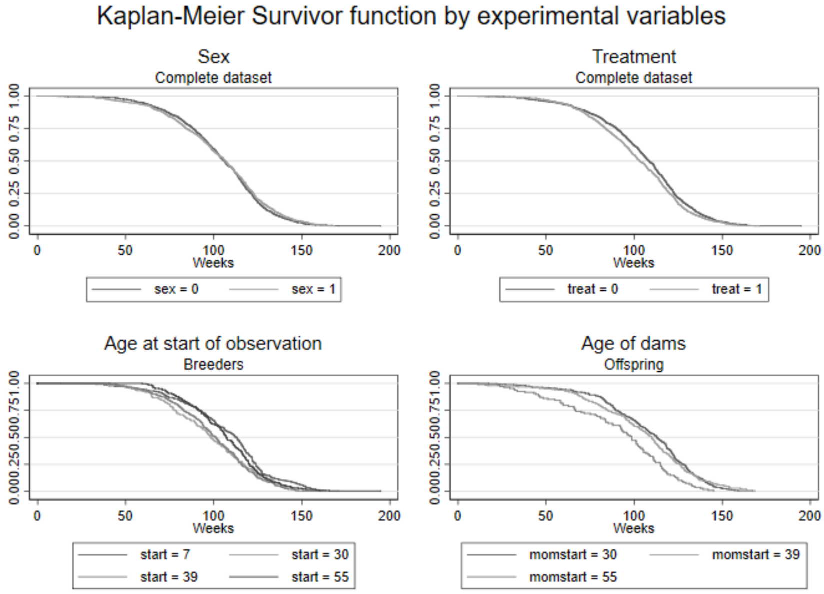


Figure 2: Kaplan-Meier survival estimates for experimental groups

According to the guidelines, the statistical analysis of these data could have stopped here. However, applying the methods that have been presented so far might give interesting insights on the problem, and more appropriate, robust and accurate results.

3.3.2 Analysis of time-to-event for the first-generation's rats

We are dealing with two generations of animals, that, as we said, differ mainly because the first was randomized, while the second was not, so entire litters are present among the offspring. For this reason, it was chosen to separate the analysis and to apply to each generation the most suitable methodologies.

The analysis of survival times for the breeders started with the classic semi-parametric Proportional hazard model, that was used to evaluate the influence of sex, treatment and their interaction. The variables were tested in univariate and multivariate models, and, coherently with the results of explorative analysis, treatment resulted the only experimental condition to represent an additional risk factor, increasing the probability of the event of about 11% for the subjects from the treated groups. As the plot of the estimated survivor function shows, the differences are not really marked, anyway.

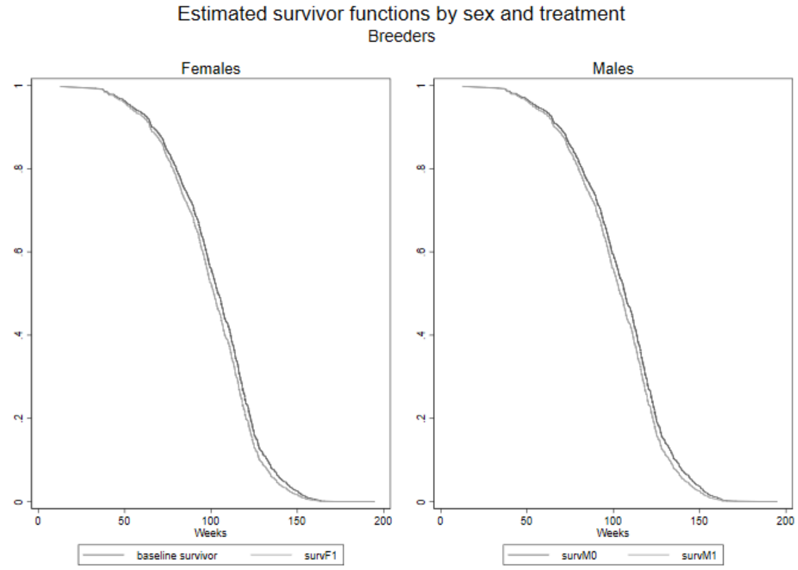


Figure 3: Estimated survivor function from the PH model, breeders

The exploration of the residuals showed that the overall fit of the model is very good, except for the last part, where the decreasing number of observation causes expectable issues; the plot of deviance and martingale residuals showed that some individuals could be too influential in determining the estimates, and the df betas and loglikelihood displacement helped individuate them. A careful examination of data confirmed that they don’t correspond to measurement or recording errors, and they fall in the range of plausible values, so they shall not be considered as outliers and excluded from the dataset. Schoenfeld residuals, “log-log” plots and tests made to verify that the log hazard-ratio function is constant over time showed that the survival functions are not so different among the groups we are considering, but that the assumption of proportionality of hazard is not tenable with respect to variable “sex”.

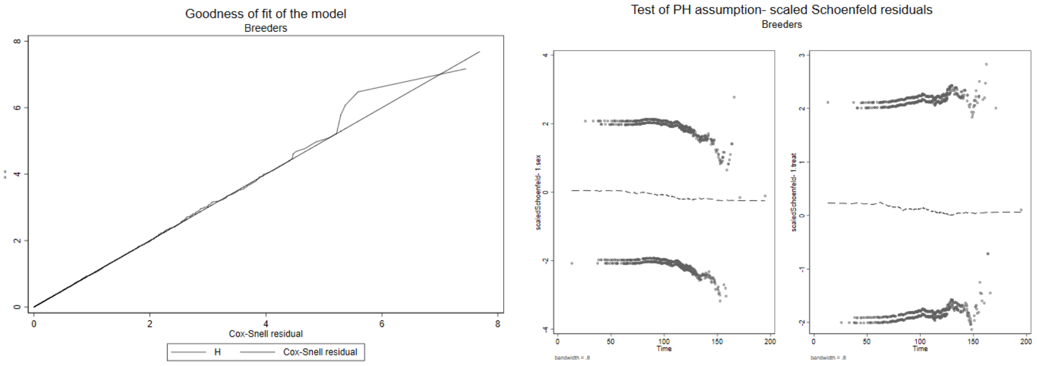


Figure 4: Residuals to evaluate the assumptions underlying PH model, breeders

Since the fundamental assumption of this approach is in doubt, it was decided for the use of the accelerated failure-time metric, and the evaluation of the effect of a univariate frailty has been included in the parametric AFT models only. The lognormal, loglogistic and generalized gamma distributions were tested in order to find the best parametrization for the model:

Variables	Lognormal	Loglogistic	Generalized gamma
1.sex	0.005	0.003	0.021*
1.treat	-0.029*	-0.034**	-0.025**
Constant	4.610***	4.640***	4.729***
Ancillary	-1.283***	-1.902***	-1.533***
Kappa			1.093***
Log-likelihood	-157.843	-112.815	-20.483
AIC	323.686	233.630	50.966
BIC	344.3665	254.311	76.817

**** p<0.01, ** p<0.05, * p<0.1*

Table 4: Estimates and goodness-of-fit measures for different distributions of AFT models

The comparison of the log-likelihood and AIC of each fit, together with the graphic representation of Cox-Snell residuals, showed that the generalized gamma is the best choices for the data.

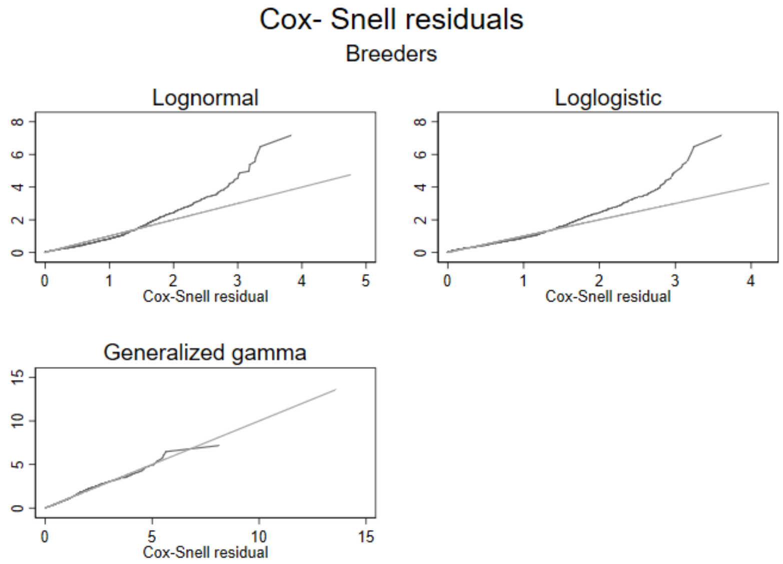


Figure 5: Cox-Snell residuals for the evaluation of the best distribution for AFT models.

No univariate frailty could be introduced in the Generalized gamma AFT model for computational problems of the software; it was tried therefore to fit a model with a univariate frailty term using the “second best” option, the Loglogistic distribution, that still showed a reasonable capacity to model the data. This compromise proved useless in the end, since the term to evaluate unobserved heterogeneity at the individual level was far from significant, and did not improve the fit of the model in any way.

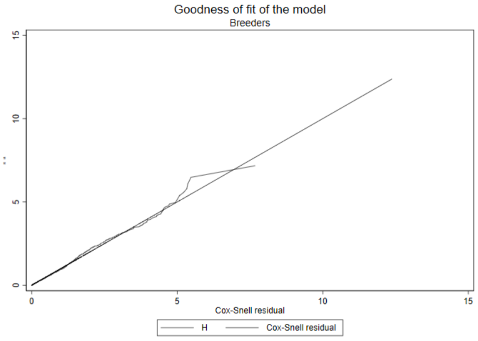
We can conclude that the survival experience of the animals belonging to the “first generation” group is best analysed using an accelerate failure-time model, and the most suitable distribution to represent their hazard function is a generalized gamma.

	PH	AFT	AFT
		Generalized gamma	Loglogistic
Variables	Coef	Coef	Coef
1.sex	-0.106*	0.021*	0.003
	(0.056)	(0.012)	(0.014)
1.treat	0.109**	-0.025**	-0.033**
	(0.056)	(0.012)	(0.014)
Constant		4.729***	4.642***
		(0.012)	(0.012)
Ancillary		.216***	-1.915***
		(.005)	(.0232)
Kappa		1.093	
		(.074)	
Theta (gamma frailty)			5.42e-09
Log-likelihood	-8009.919	-20.48326	-117.862
AIC	16023.84	50.96653	245.724
BIC	16034.18	76.81712	271.5746

Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table 5: Results of PH and AFT models for breeders, and measures of goodness of fit.

The estimate coefficients confirm that the differences in survival among sexes are not relevant, while the treatment is responsible for a small but significant acceleration of the risk of the event (or, to be more straightforward looking at the negative coefficient, that the treatment causes the failure to happen more rapidly, so the expected time-to-event decreases). Residuals have been analysed mostly graphically, and they showed that the model is overall a reasonable representation of the data, and that there shouldn’t be major issues regarding the respect of the assumptions.



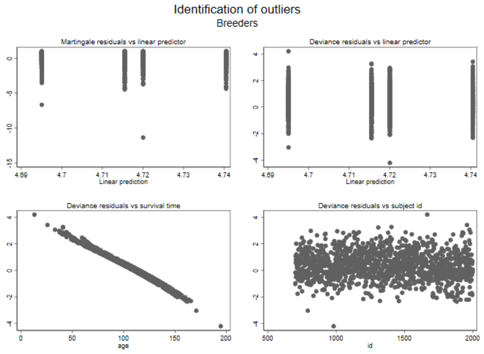


Figure 6:Residuals to evaluate the assumptions underlying AFT model, breeders

3.3.3 Analysis of time-to-event for the second-generation's rats

The procedure that was followed to analyse the group of the offspring was essentially the same followed before: it started with the fitting and interpretation of a Cox model, whose assumptions were evaluated in order to assure its validity.

The models were fitted to assess the effect of treatment on the risk of death, controlling for a possible effect of sex and of one more variable, the age of the dams at the beginning of gestation, that can take the values of 30, 39 or 55 weeks of age (in the dataset it is identified as "momstart").

Model 1- Proportional Hazard					
	Coef	Std. Err.	z	P>|z|	[95% Conf. Interval]
Observations 699					
1.sex	0.044	(0.076)	0.571	0.568	-0.106 - 0.193
1.treat	0.182	(0.077)	2.373	0.018	0.032 - 0.332
39.momstart	0.026	(0.085)	0.307	0.759	-0.140 - 0.192
55.momstart	0.553	(0.112)	4.953	0.000	0.334 - 0.772
Log-likelihood	-3870				
AIC	7748.019				
BIC	7766.218				

Table 6: Proportional hazard model, offspring

As before, sex was not found to increase the risk of the event in a statistically significant way, while treatment has a relevant effect, so we can say that there is sufficient evidence to link a heavy consumption of Coca-Cola since prenatal life with an increase in the hazard of death; as expected, the later a dam undergoes mating, the higher becomes the risk of death for the pups: the risk for the pups of dams of 39 weeks is similar to those of 30, while for those bred by the oldest dams the risk increases sensibly.

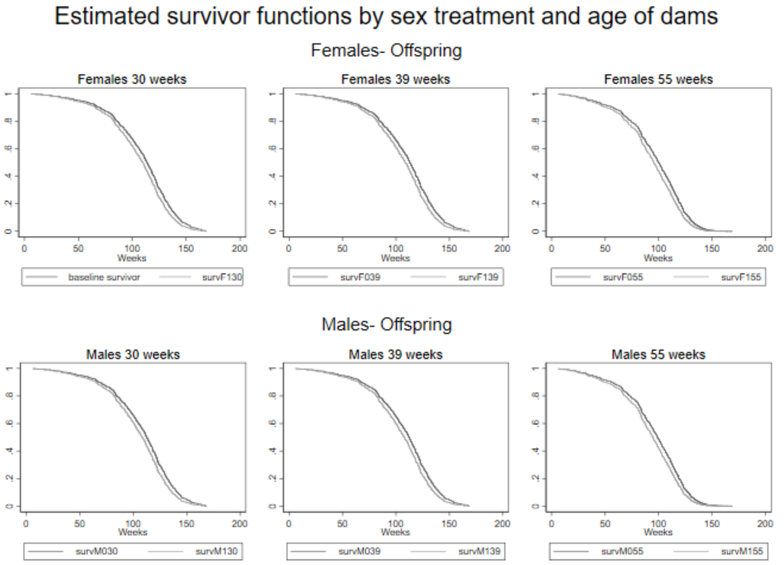


Figure 7: Estimated survivor functions by sex, treat and age of dams, offspring

The results showed that for this dataset the assumption of proportionality of hazard is realistic: even if Schoenfeld and scaled Schoenfeld residuals show some irregular trend, a formal test for the equality of log hazard-ratio over time excluded important violations. No observations were found to have an excessive influence on the estimates or on likelihood of the model, and globally it performs quite good for these data.

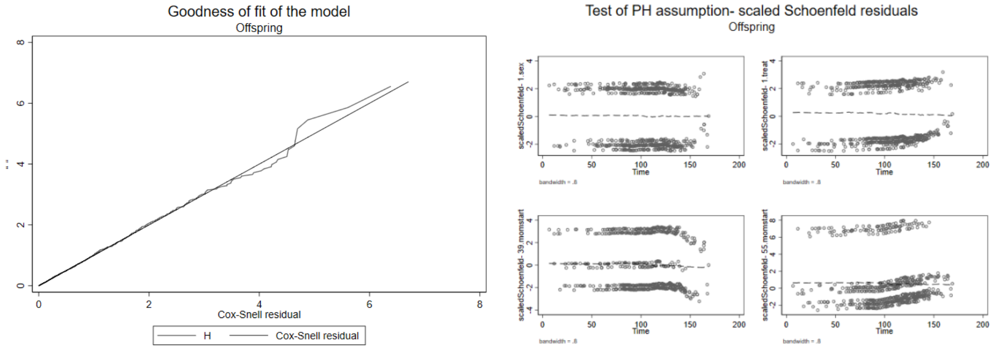


Figure 8: Residuals for the evaluation of the assumptions underlying PH model.

A frailty term was than introduced in the model, this time representing a proper shared latent random effect, considering each litter a cluster: its effect represents the common characteristics that individuals belonging to the same cluster verisimilarly share. It accounts for the possible association existing among them, that would make unreasonable the assumption of independence of the subjects. Frailty terms are assumed here to be gamma-distributed with mean 1 and variance θ, and they affect the hazard multiplicatively; a significant shared frailty would confirm that correlation among members of the same litter exists (and the degree of the correlation is measured by θ), and is relevant in explaining survival patterns.

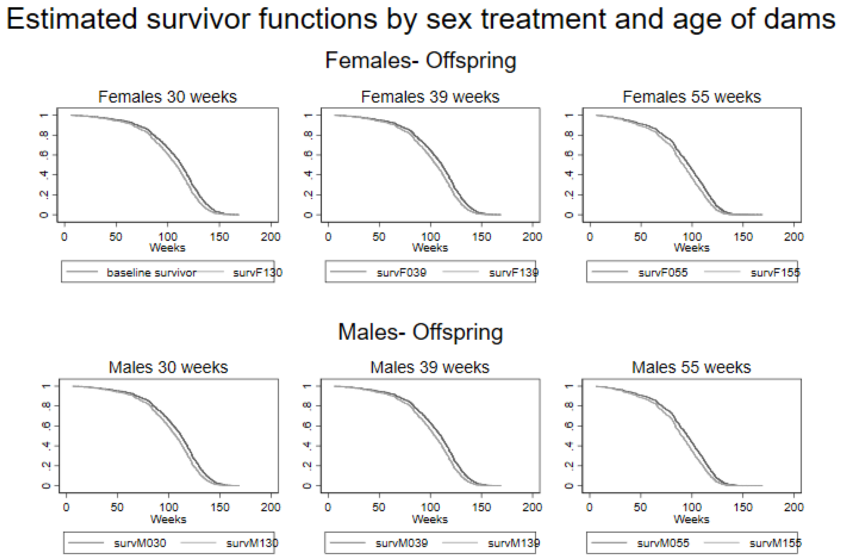


Figure 9: Estimated survivor functions from PH model with a shared frailty term

The term proved significant for the dataset of offspring: the effect of shared characteristics on survival is relevant and has to be considered. On the other side, it must be noticed that treatment loose of importance if this new variable is considered.

Model 2- Proportional hazard + shared Gamma frailty					
	Coef	Std. Err.	z	P>|z|	[95% Conf. Interval]
Observations	699				
Number of groups	98				
1.sex	0.048	(0.081)	0.587	0.557	-0.112 - 0.207
1.treat	0.229	(0.125)	1.830	0.067	-0.016 - 0.474
39.momstart	0.095	(0.138)	0.684	0.494	-0.176 - 0.365
55.momstart	0.688	(0.172)	4.004	0.000	0.351 - 1.024
theta	.166	(.046)		0.000	
Log-likelihood	-3850				
AIC	7708.57				
BIC	7726.768				

Table 7: Estimates and goodness of fit of PH model with a shared Gamma frailty

The post-estimation analysis for model validity and goodness of fit do not show major problems or violations in the assumptions underlying the model, neither regarding the presence of outliers, nor the proportionality of hazards, so for the second-generation group it was not necessary to look for a better specification using different metrics.

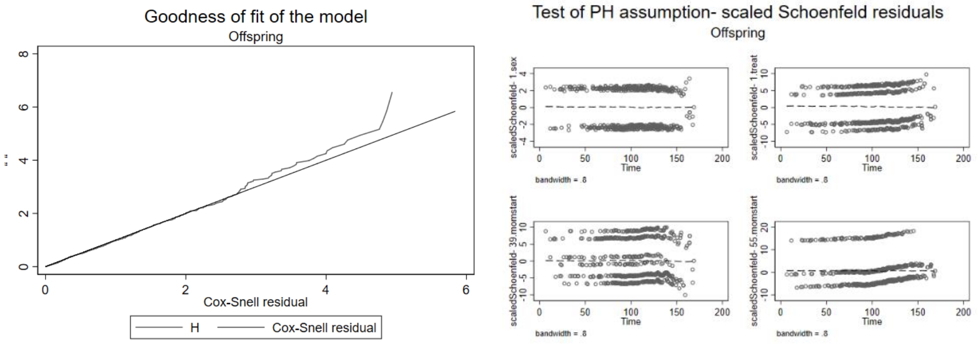


Figure 10: Residuals for the evaluation of the assumptions underlying PH model with Gamma frailty

Regarding the analysis of survival times, we can draw some conclusions: first, it is important to go beyond the plain comparison of nonparametric estimates. Indeed, mostly in case of complex data, or atypical ones, like those that were analysed here, where we had different generations and members of the same litters in experimental groups that were not created using randomization, the hazard of death could be attributed to factors that are not of major relevance in reality. A good example comes from the survival experience of the second-generation rats: had we not included the shared frailty term, that accounts for unobservable common features of individuals (it can be a genetic predisposition for some kind of disease or tumours, the fact of belonging to a very numerous litter, where individuals are necessarily smaller and therefore possibly weaker, and so on), we might have attributed the observed differences in the times to death entirely to the treatment.

Another caveat is to always bear in mind the importance of verifying that the methods that are chosen to analyse data are suitable for them, and that it is neither granted nor banal that all assumptions on which models are based, are always met. Here, the hypothesis of proportionality of hazard did not hold in the breeders' dataset, and that emerged only after the examination of the residuals; the importance of model checking is clear for statisticians, but the same can't be assumed for all researchers from different fields that follow the whole experiment, included data analysis.

Chapter 4 The analysis of longitudinal data

Data collected measuring one or more variables of interest in repeated occasions over the same set of subjects, that we assume to constitute a random subsample of the population of interest, are often referred to as repeated measures data. When repeated measures are ordered in time or space, and we cannot assume that the observations within each individual has been assigned randomly, we call these data *longitudinal*. When they are available, individual patterns of change can be observed: longitudinal data can thus provide richer information, but request particular attention, since the drop out of the individuals under study represent in some applications an important issue (clearly not in our case), and correlation rises among observations registered for the same subject.

Long-term and life-span carcinogenicity experiments often request several types of longitudinal data: measurements of body weights and mean consumption of food and beverages for cage are routinely collected with quite dense schedule, since they are an important track of the animal well-being and of the good course of the experiment itself (OECD 2002). They should be monitored jointly, since changes in body weight can be related to alterations caused by interactions between nutrients and the tested substance, or by a different palatability of foods. Sensitive changes (in particular losses) may be important signals of health problems, and should be always regarded with attention. These indicators also gain importance *per se* because they are heavily related to metabolic, hormonal and homeostatic functions, growth and sexual maturation. In this study, where the tested compound is a highly caloric and sugary beverage, body weight in particular should be of primary interest, in association with classic tumour incidence, since it is well established that overweight and obesity are positively associated with the increase of risk of many types of cancer. Based on a systematic review of the published scientific literature, IARC assessed that the absence of excess body fat has a preventive effect in humans for cancers of the colon and rectus, oesophagus, kidneys, breast and endometrium, and, with sufficient evidence, for cancers of the gastric cardia, liver, gallbladder, pancreas, ovaries, thyroid, meninges, and on multiple myeloma (Lauby-Secretan B 2016).

4.1 The analysis of longitudinal data according to the guidelines

In usual carcinogenicity studies, nevertheless, body weight variations and food consumption do not directly represent the primary outcome of interest, so their analysis is, again, not particularly thorough. Guidelines (OECD 2012) suggest to represent graphically group means, to promptly identify unexpected trends; this is most important in during the early to middle part of life, while after approximately 80 weeks of age the rodents enter the geriatric phase, and are more prone to weight losses due to ageing, diseases or tumours; also, it was noted that lighter rats tend to live longer than heavier ones, so means might be biased downwards. Formal analysis of data always refers to body weights and food consumptions by timepoint of interest and averaged on groups, since no methodology for repeated measures is explicitly advised. It should start with a test to identify the presence of outliers, like the Dixon and Massey test, followed by a test to evaluate the assumption of normality (either the Kolmogorov-Smirnov or the Shapiro-Wilk test). In case of not-normally distributed data, a logarithmic transformation is suggested, and the test repeated. If data fall back into normality assumption, another test for outliers such as the Extreme Studentized Deviate statistics could be carried out; then the F-test, or the Levene's or Bartlett's tests can be performed to evaluate the homogeneity of variance, respectively if the experiments has two or more groups. If the F-test highlights that variances are homogeneous, a Student's t-test can be used to evaluate differences between groups, while if they are heterogeneous the comparison is performed with the modified t-test, using Satterthwaite's method. In presence of more than two groups having homogeneous variances, the comparison between all groups can be performed using one-way ANOVA followed by Duncan's multiple range test or Tukey's Honest Significant Difference test, and by Dunnett's test for pair-wise comparisons between the control and the dosed groups. If the assumption of normality is not tenable even after transformation and analysis of outliers, or if the Levene's test shows that variances are heterogeneous, some alternative methods are suggested: the Kolmogorov-Smirnov's test or the Wilcoxon rank sum test can be used to compare two groups, while global differences among more experimental groups can be tested with a Kruskal-Wallis ANOVA by ranks, or a Jonckheere's test. If significant differences are found, multiple comparisons are possible using distribution-free methods like Dunn's or Shirley's tests. A concise representation of the suggested analysis can be seen in Figure 1.

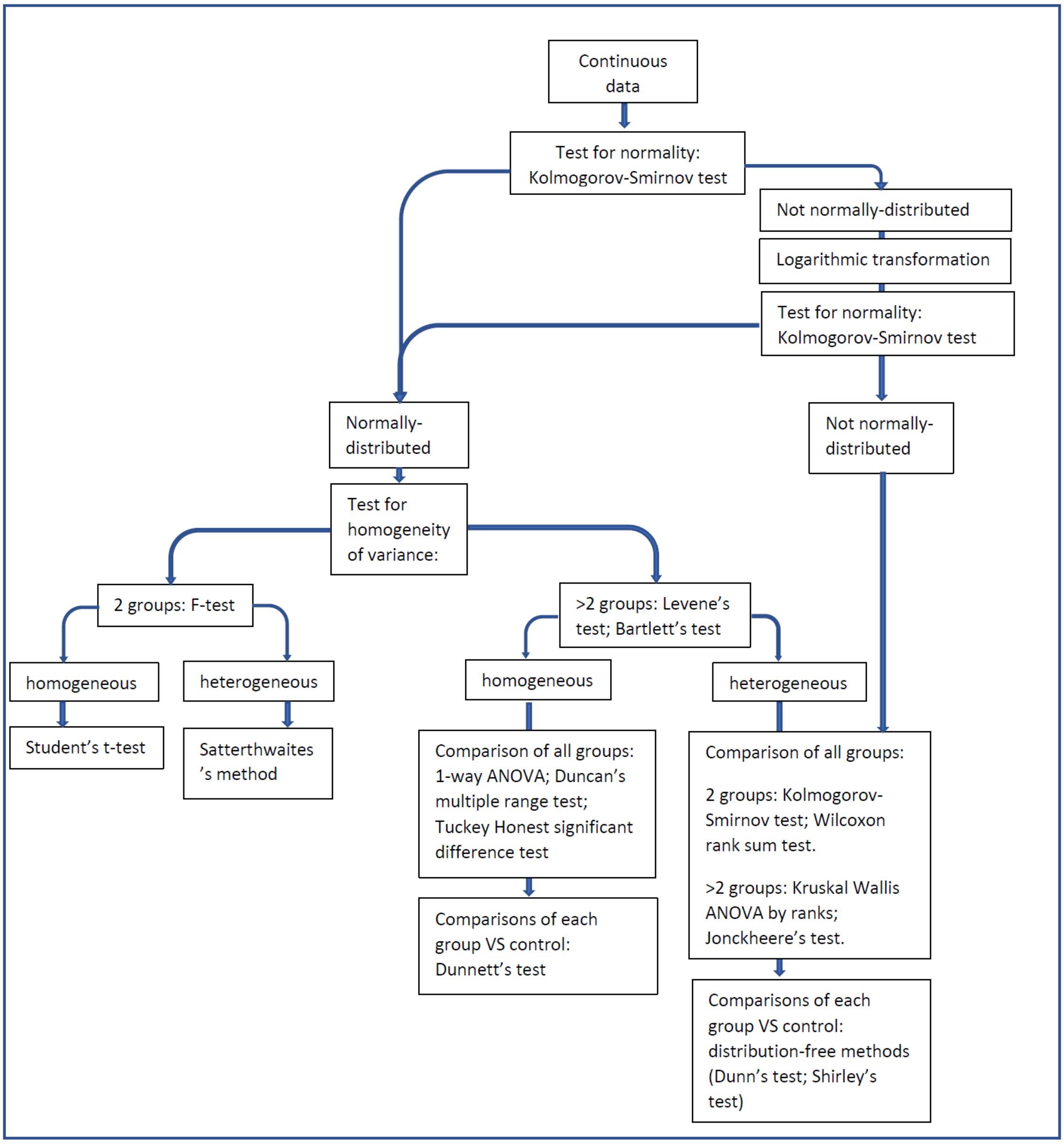

Continuous data
Test for normality: Kolmogorov-Smirnov test
Not normally-distributed
Logarithmic transformation
Test for normality: Kolmogorov-Smirnov test
Normally-distributed
Not normally-distributed
Test for homogeneity of variance:
2 groups: F-test
>2 groups: Levene's test; Bartlett's test
homogeneous
heterogeneous
homogeneous
heterogeneous
Student's t-test
Satterthwaites 's method
Comparison of all groups: 1-way ANOVA; Duncan's multiple range test; Tuckey Honest significant difference test
Comparison of all groups:
2 groups: Kolmogorov-Smirnov test; Wilcoxon rank sum test.
>2 groups: Kruskal Wallis ANOVA by ranks; Jonckheere's test.
Comparisons of each group VS control: Dunnett's test
Comparisons of each group VS control: distribution-free methods (Dunn's test; Shirley's test)

Figure 11- OECD statistical decision tree summarizing procedures for the analysis of continuous data

The approach suggested in the guidelines is one of the many that have been developed based on the analysis of variance since the beginning of XX century: the foundation goes back to the work of the astronomer G. Biddel Airy, and was later formalized by R. A. Fisher (Fitzmaurice G 2008).Several approaches took the moves from the ANOVA paradigm, and the one just described is probably one of the simplest, since it consists in summarizing the collection of measurements for each individual in a single or a set of values, that are than compared using the

ANOVA method. Means are a very immediate measure, and the area under the curve (AUC) is also quite common. Appealing for its simplicity, this approach has clear drawbacks: the loss of information, and the fact that completely different series of data can produce the same summary measure; the impossibility to evaluate the effect of time-varying covariates; the violation of the assumption of homogeneity of variances at the basis of ANOVA, that is very likely when measurements are irregular, not equally spaced or have missing data.

Another possible method, one of the earliest, was the mixed-effect ANOVA, also referred to as the univariate repeated-measures ANOVA: since the structure of longitudinal data shows some similarities to that of data from randomized block designs, for which ANOVA had been developed, the methodology was applied to repeated measures data, regarding the individuals as the blocks. The model can be written as

$$Y_{ij} = X'_{ij}\beta + b_i + e_{ij}$$

with $i = 1, \ldots, N$ individuals and $j = 1, \ldots n$ measurements, where X'_{ij} is a design vector, β a vector of regression parameters, $e_{ij} \sim N(0, \sigma_e^2)$and $b_i \sim N(0, \sigma_b^2)$ doesn't stands as the proper block effect anymore, but as a random subject effect representing the unobserved or unmeasured characteristics that cause differences in the outcome in each subject. This factor accounts for positive correlation among subsequent measurements within-subject, but is forced to follow a quite restrictive structure of the covariance, maintaining constant variance and covariance ($Var(Y_{ij}) = \sigma_b^2 + \sigma_e^2$, and $Cov(Y_{ij}Y_{ik}) = \sigma_b^2$). The assumed compound symmetric structure of the covariance matrix is not the best for longitudinal data since correlation is likely to decrease as separation in time increase, and sometimes the variance does not hold constant in time. Some shortcomings of the model were addressed in time: Greenhouse and Gassier (1959) proposed an adjustment to handle more general covariance structure, while Henderson (1963) developed a method for unbalanced data. Anyway, the idea of allowing for random differences among subjects is the basis of various subsequent regression methods, so the univariate repeated-measures ANOVA can be considered as a precursor of the regression methods that will be addressed later.

A slightly more complex method was also used for the analysis of longitudinal data, the repeated-measures analysis by MANOVA, or repeated-measures ANOVA, a special case of MANOVA that handle multivariate, correlate data, and allows less stringent assumptions on the structure of the covariance matrix (covariances are only assumed homogeneous across subjects). This method, too, has some limitations, since it can't handle time-unbalanced designs and missing data, leading to inefficiencies and possibly biased results.

These methods can be used for longitudinal data after verifying their suitability, but their limitations must be kept in mind; more sophisticated and flexible methods are available. They will be exposed and applied to the Coca-Cola experiment data, and their usefulness evaluated.

4.2 Other methods for longitudinal data

4.2.1 Linear mixed-effects models

Mixed effects models are likely the most widely used tool for continuous outcomes whose residuals distribute normally but are not independent or homoscedastic: these are characteristics of "grouped" data, where grouping may arise from clustering (measuring outcome on pups from the same litter for example), or from repeated measurements or longitudinal studies, where individuals are assessed repeatedly over time or under different experimental conditions. The correlation between measurements arising from the feature of the data is conveniently addressed in these models thanks to the possibility to choose between several types of parsimonious covariance structures.

Furthermore, the basic idea of having a common functional form for all individuals, with parameters that vary among subjects is quite intuitive and can be appropriate in numerous situations. Indeed, these models are called "mixed" since they can incorporate explicative variables as fixed or random effects: fixed effects are associated to continuous or discrete covariates, and represent unknown parameters that are constant in the population, for each level or value of the associated independent variable. When the levels of a categorical variable can be considered as random realizations of a sample space, and they are not of particular interest *per se*, they can be modelled as random effects.

Its foundations lie in the ANOVA paradigm, as can be seen in the works of Scheffé (1959) and Harville (1977), but also in the idea of allowing for random differences across individuals when analysing growth curves, as in Wishart (1938), Box (1950), Rao (1958), Potthoff and Roy (1964), and Grizzle and Allen (1969). Another contribution in the development of mixed-effect model was the two-stage approach, adopted and made popular by the US National Institute of Health. According to this approach, the distribution of the repeatedly measured outcome is the same for all subjects and is characterized in the first stage, but the parameters can vary randomly over the units (so they also can be referred to as random effects), and their distribution constitutes the second stage of the model.

Laird and Ware (1982) were the first to propose a flexible class of mixed models for longitudinal data: it includes bot growth and repeated-measures models as special cases, and introduces population parameters, individual effects and within-subject variation, as well as between-subject variation.

In their representation, the $n * 1$ vector of responsesfor the ith subject can be modelled as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \qquad\qquad i = 1, \dots N$$

where

- $\boldsymbol{X}_{ij}$ is a $n_i * p$ design matrix of explanatory variables or fixed factors;
- $\boldsymbol{\beta}$ is a $p * 1$ vector of unknown population parameters, or fixed effects coefficients, describing the relationships between the outcome and the explanatory variables for group defined by levels of a fixed factor (for example, describing the contrast between males and females);
- $\boldsymbol{Z}_i$ is a $n_i * q$ design matrix of variables orrandom factors;
- $\boldsymbol{b}_i$ isa$q * 1$vector of unknown random effects, specifically referred to a given level of a random factor, usually representingthe deviations from the relationships described by fixed effects. Random effects can be set as

random intercepts (random deviations for an individual or cluster from the overall fixed intercept), or as random coefficients (random deviations for an individual or cluster from the overall fixed effects). They are assumed to follow a multivariate normal distribution ($\sim N(\mathbf{0},\ \boldsymbol{D})$), with$D$being a $q * q$symmetric, positive definite variance-covariance matrix;

- $\boldsymbol{\varepsilon}_i$ is a$n_i * 1$vector of errorsfor the ith subject for each measurement occasion, whose terms do not need to be independent but can be correlated within individual. The residuals for each subject follow, again, a multivariate normal distribution ($\sim N(\mathbf{0},\ \boldsymbol{R_i})$),with 0 mean and a positive definite $q * q$ variance-covariance matrix,$\boldsymbol{R}_i$.

Several covariance structures can be specified both for $\boldsymbol{D}$ and for $\boldsymbol{R_i}$[4]. For $\boldsymbol{D}$, a very common definition is the unstructured, where no additional restriction is assumed on the value of its elements apart from positive- definiteness and symmetry, so the variance components to be estimated are $q * (q - 1)/2$ (the variance σ_b^2 for each of the q random effects and the covariance $\sigma_{bk,bl}$ for each couple). Other more parsimonious covariance structures are possible, but they require more constraints: an example isthe diagonal matrix, where only the variances are estimated, while the covariances are set to 0. The simplest specification for $\boldsymbol{R_i}$ is the diagonal structure, that require the estimation of one single parameter for the variance component, σ_b^2, since the residuals are assumed to be uncorrelated within individual, and to have common variance. The covariance matrix can alternatively take a compound symmetry structure, that assumes equal variances and equal covariances among observations within subject, and is suitable for example for repeated assessments under the same experimental conditions, when equal correlation of the residuals is plausible. Another quite common structure is the 1st order autoregressive or AR(1), under which the variance components vector only contains two parameters, the variance σ_b^2, that is assumed constant and always positive, and a correlation parameter ρ, whose values go from -1 to +1; the covariance is calculated as $\sigma_b^2\rho^w$, where

[4] The following presentation of Linear Mixed effect models is based on the work of West et al. (2010)

w represents the lag between observations within subjects (so for adjacent observations is higher, and tends to 0 for very far observations): this makes it suitable for equally spaced experimental designs. Other structures, with less constraints, are possible, but are less parsimonious, and require the estimation of more parameters of the variance components. Mixed models even allow to assume that subjects characterized by different levels of variables may share the same structure of the covariance matrix, but have different values for the parameters of the vector of variance components of $\boldsymbol{D}$ and $\boldsymbol{R_i}$.

An alternative specification of the mixed effect model, referred to all individuals together, is given as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}\,\boldsymbol{u} + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{u} \sim N\,(\boldsymbol{0}, \boldsymbol{G}) \text{ and } \boldsymbol{\varepsilon} \sim N\,(\boldsymbol{0}, \boldsymbol{R})$$

Here $\mathbf{Y}$, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are vectors obtained from “stacking” respectively the $\boldsymbol{Y}_i$, $\boldsymbol{u}_i$ and $\boldsymbol{\varepsilon}_i$ vectors for all subjects vertically, the $n \times p$ design matrix $\boldsymbol{X}$ is obtained by stacking all $\boldsymbol{X}_i$ matrices vertically and the $\boldsymbol{Z}$ matrix is a block-diagonal matrix, with blocks on the diagonal defined by the $\boldsymbol{Z}_i$ matrices. The $\boldsymbol{G}$ matrix is therefore a block-diagonal matrix containing the variance-covariance matrix for all random effects, with blocks on the diagonal defined by the $\boldsymbol{D}$ matrices, while the $n \times n$ matrix $\boldsymbol{R}$ is a block-diagonal matrix of variance-covariance for all residuals, where the $\boldsymbol{R}_i$ individual matrices define the blocks on the diagonal.

Since it is assumed that the random effects and the error terms follow a normal distribution, the whole model can be written marginally as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

where$\varepsilon_i \sim N(0,\ V_i\ )$and $V_i\ = R_i + Z_i\, D\, Z_i'$, with the covariances of the observations (the off-diagonal elements) allowed to be correlated, so different from 0. Consequently, the marginal distribution of the vector of responses is defined as $y_i \sim N(X_i\,\beta,\ R_i + Z_i\, D\, Z_i{}'\ )$. This representation is worth being reminded since it’s in this framework that are estimated the fixed effects and the variance components in most statistical software, and lies at the basis of the likelihood approach for estimation.

Inference on the parameters estimates can be based on least squares and maximum likelihood methods, or, formulating the model the appropriate way, using an empirical Bayesian method. Under the classical frequentist approach, β and θ (where θ is a $q * 1$ vector containing the variances and covariances from R_i and D) can be obtained maximizing the likelihood function, as usual: the likelihood is set as a function of the parameters of the hypothesized model, taking all assumptions into account: for the mixed effect model, the marginal specification of the model is used. The values of the parameters that, given all assumptions, make the observed values of the outcome most likely, are the maximum likelihood estimates for the parameters. The likelihood function $L(\beta, \theta)$ is given by the product of all individual likelihood functions, and the log-likelihood is found as usual, as the natural logarithm of the likelihood function, resulting in

$$l_{ML}(\beta, \theta \mid y) = -\frac{1}{2}\left[n \cdot \log(2\pi) + \log(\boldsymbol{V}_i) + (\, y_i - \, X_i\, \beta)\boldsymbol{V}_i^{\,-1}(\, y_i - \, X_i\, \beta)\right]$$

The estimates of the covariance parameters are obtained using iterative procedures, until the reaching of convergence; once the ML estimates of θ are found, they are used to compute (directly) the estimated value of $\boldsymbol{V}_i$, and finally to calculate the generalized least square for the regression parameters $\boldsymbol{\beta}$. Since an estimate of the values in $\boldsymbol{V}_i$are used, the *βs* are also called the *empirical best linear unbiased parameters.* The main problem with maximum likelihood estimates of the variance components, is that they do not account for the degrees of freedom used to estimate the parameters of $\boldsymbol{\beta}$, so they are biased, as discussed in Verbeke & Molenberghs (2000). To overcome this issue, the Reduced Maximum Likelihood estimates are more often used. REML was proposed initially to overcome the problem of the estimation of variance components from unbalanced or incomplete block design, and was then adopted as an alternative tool and even preferred to classical ML method since the estimates are not biased, because the degrees of freedom used for the estimation of the fixed effects are considered. This allows to obtain estimates of the values in the $\boldsymbol{V}_i$ matrix; then the parameters of the $\boldsymbol{\beta}$ can be found using the methods from ML approach, using generalized least squares. The estimates obtained with the REML and ML method are different: for $\boldsymbol{\beta}$, ML produces biased results, while REML does not, as already mentioned; for the variances of $\boldsymbol{\beta}$, both methods give results that are biased downwards, because neither compensates the uncertainty coming from the use of empirical estimates of $\boldsymbol{V}_i$. Usually, this problem is overcome trying to find the best possible estimates of the $\boldsymbol{V}_i$, using REML to fit models with alternative structures for $\boldsymbol{D}$ and $\boldsymbol{R_i}$.

The estimation of the covariance parameters (that is, the maximization of the log-likelihood function under the assumption of positive-definite matrices $\boldsymbol{D}$ and $\boldsymbol{R_i}$)is usually perform using the Expectation-maximization algorithm, the Newton-Raphson procedure or the Fisher scoring algorithm. In most statistical software, first appropriate starting values for the parameters are estimated with the EM algorithm; these are then used for the following estimations, using the preferred algorithm. The most common method is the Newton-Raphson, both for ML and REML.

Mixed effects models have also proved to be robust in the analysis of unbalanced data when compared to the General Linear Model framework (Pinheiro and Bates, 2000): individuals with incomplete observations can still be included in the analysis, and with complete data, too, mixed effects models provide advantages over the GLM.

So far, the response variable for individuals was assumed to follow a linear and continuous trend, but often it can present discontinuities, or show a nonlinear path, making reasonable the assumption that the likelihood function may depend on the parameters in a non-linear way. It is the case of several physical phenomena, and growth is the main example. Several adaptations may be adopted to cope with these situations, the most common being the splitting of the time of analysis into subperiods, so that the linearity assumption underlying the model is reasonable within each one. The drawback of this simple solution is that it prevents from analysing the phenomenon under study in its richness, treating discontinuities and nonlinearities as problems to overcome to fit the data to the model, rather than as the characteristics of the process itself.

Two approaches can be tried in this situation (Singer J. D. 2003): the first is more empirical, based on the observation of individual trajectories and the manipulation of data. One possibility is to identify a suitable transformation of the outcome or the time scale that can lead to a linearization of the trajectories for each individual, so that the assumption of linearity holds, and the simpler model illustrated so far can be fitted. Many ways to transform data of course exist, and the most suitable must be found every time through several tries[5].Another possibility is to introduce in the model different predictors that all together characterise time, that is, to represent time as a polynomial function. So, a second-order polynomial may be fitted to account for trajectories that resemble quadratic change with one stationary point; a third-order polynomial will account for cubic change, and so on. The main issues with this device is the increased complexity of the model and the interpretability of the regression parameters. The second approach is more theory-based, and requests to identify a reasonable functional form underlying the trend of subject's data, that will be used in the model to describe the relationship between the response and explicative variables. As we will show in the next paragraphs, both strategies have been applied to the body weights data of the second generation of rats from the Coca -Cola study.

4.2.2 Nonlinear mixed effects models

The general model presented in the previous paragraph and all the expedients that allow to represent growth trajectories of many shapes share a characteristic: they imply an association linear in the parameters between the outcome and the explicative variables, because the model is composed by individual growth parameters, that are linked in a linear way. Often however, the likelihood function depends on the parameters in a non-linear way: it is the case of several physical phenomena, and growth is the main example. In such cases, the use of nonlinear model is justified by the possibility to obtain a more interpretable model, and to

[5] A useful tool to have an idea of the possibilities is given in Mosteller and Tuckey (1977), where the ladder of transformation is associated to the so-called *rule of the bulge* (the idea is to associate the approximate shape of the plots to a suitable transformation of the outcome or time, to be applied to all individuals)

use a smaller number of parameters, mostly if we compare them with high-order polynomial models.

Lindstrom and Bates (1990)were the first to present a general, nonlinear mixed effects model for data in which the assumption of the normality of residual holds, but the expectation function is nonlinear. The model can be written as

$$y_{ij} = f\left(x_{ij}, \beta, u_i\right) + \boldsymbol{\varepsilon}_{ij}, \qquad i = 1, \ldots, N$$

where$f$ is a real valued function, $x_{ij}$is a vector of covariates containing both within- and between-subjects covariates, β is a $q * 1$ vector of unknown parameters of fixed effects, u_i is a vector of unobservable subjective random parameters following a multivariate normal distribution with 0 mean and variance-covariance matrix Σ, andε_i is the usual error vector of dimension $n_i * 1$, following a multivariate normal distribution with 0 mean and variance-covariance matrix $\sigma^2 \Lambda$.

It is useful to adopt the two-stages representation of the model[6], since it helps clarifying how the non-linear function is used to express the individual trajectory of change at level 1

$$y_{ij} = m\left(x_{ij}^{w}, \varphi_i\right) + \varepsilon_{ij},$$

where m describes the behaviour of the individual growth as depending on individual-specific parameters φ_i and the vector of within-subject covariates x_{ij}^{w}, while the inter-individual variability can be expressed using a regular linear relationship at level 2:

$$\varphi_i = d\ (x_{ij}^{b}, \beta, u_i)$$

where d is a vector function that explains the variation of individual-specific parameters between subjects, and incorporates β, the vector of parameters for the population, and x_{ij}^{b}, the set of between-subjects covariates.

The assumptions underlying the non-linear mixed effects model are that the random effects $u_i$and the error terms ε_i are independent between each other and across individuals, that $\sigma^2 > 0$ and that matrix Σ is definite nonnegative.

[6]The following presentation of nonlinear mixed effect models is mainly based on the work of Demidenko (2013)

An important part of the modelling process regards the choice of which parameters to consider random, so individual-specific, and which can be regarded as population-averaged, so fixed.

A specific feature of NLME models is that the random parameters φ_i appear in a nonlinear function, implying that the expected value of the response can't be expressed in closed terms, in terms of population averaged parameters. This makes it is very difficult to establish the statistical properties of the model in small samples; indeed, even the condition of a number of subjects *N* tending to infinity while the number of observations for each individual $n_i$remainis finite, that is a sufficient condition for linear mixed effect models to be consistent, asymptotically normal and efficient, is enough only for maximum likelihood estimation. Maximum likelihood estimation, yet, requires the integration of the unobservable individual-specific parameters, and this leads to the presence of a multidimensional integral. To avoid this problem, various approximation methods for the estimation of the model exist. They can be grouped into two categories, the two-stage methods, and those that approximate the NLME function to a linear function, to reduce the model to a nonlinear marginal mixed effects model. In the former methods, nonlinear least squares are used to estimate individually the subject-specific parameters from the first stage; then these are used as observations for the second stage model. These methods can be proficiently used to when σ^2 is relatively small, and the number of observations for each cluster relatively large. The latter methods allow to obtain good approximations when an estimate of the random effects from the penalized nonlinear least square estimator is used. All methods, anyway, become equivalent when the number of clusters and of observations for each cluster tend to infinity. As already mentioned, the ML estimator is consistent when the number of subjects tends to infinity and the number of observations for each individual remains finite, but this characteristic can be lost if the distribution of the random effects is mis-specified (however, if also the number of observations per cluster tend to infinity, we can consider it consistent even if the distribution of the random effects is not appropriate). The ML estimator also returns correct standard error for all estimates.

Choosing the correct functional form for the model is very important: the choice of a not suitable model may prevent convergence, produce negative variances for the random effects, computational difficulties and so on. The observation of individual growth paths can ease the choice, but a theory-driven approach is possible, as it will be illustrated in the next paragraph. The growth and maturation processes have been long studied in humans, and several models have been proposed to formalize the growth patterns during infancy, childhood and adolescence. Since some similarities can be recognized in the growing process of very young humans and rats, some of these models have been "borrowed" and applied here.

4.2.3 Growth models

The analysis of the progress of growth and maturation in humans has a long history. It covers the study of variations of stature and weight, and of the velocity of growth standardized relative to the passage of time, as well as of the acquisition of secondary sexual characteristics, to investigate the progressive achievement of adult status. Describing the normal progress of childhood growth is important per se, and even more interesting is the study of any variation from the regular pattern: differences in size, velocity and timing of maturation can give indications on health problems at different levels. In normal conditions, indeed, growth is determined by the personal genetic complement and ruled by several hormonal systems, but many and various environmental factors, first of all nutrition, represent a fundamental constraint.

Auxological studies always looked at mathematics to formalize and model the normal growth phases[7], starting from the observation and measurement of individuals under study. They usually are of two types, with different goals: cross-sectional studies usually aim at obtaining age-dependent reference curves, while longitudinal studies have the purpose to understand the growth process over some period.

[7] The illustration of growth models is mainly based on the work of Hauspie et al, 2011.

Longitudinal data and the will to understand the growth process and highlight possible differences due to the treatment regime is what we have in the context of this study on rats, too, so these methods were applied to this unusual field.

It is well established that the pattern of growth of body dimensions of the "general type" (to be distinguished from those of lymphoid, neural and genital type) is increasing and S-shaped, from birth to the adult age, since it progresses rapidly in the first years, then slows down, and accelerates again around the so-called pubertal spurt, to become a plateau when the approximate adult size is reached. Simple linear regression is therefore clearly not the best tool to manage this kind of data. As already mentioned in the previous sections, polynomial models are sometimes used to represent longitudinal growth data, since they are computationally convenient, but they are preferred for small periods, and their major drawback is the difficult interpretation of the higher-order terms.

Another interesting alternative might be the use of smoothing methods, that can help highlighting the shape of growth form "noisy" data, estimating a curve without an *a priori* fixed parametric model. The most used techniques are smoothing splines, kernel estimators, or local polynomials. They may represent a good alternative in this case, where a lot of observations were registered at quite close measurements, so data might be prone to measurements errors and short terms variations.

Regression models based on an adequate parametric function would be the best alternative, since they are estimated using fewer parameters that have some biological interpretation, but they are quite insidious: the choice of the appropriate functional form is difficult, it is necessary to consider all the facets of the growth process to reduce the risk of obtaining biased results, and their estimation is computationally quite demanding.

Several models have been developed in time to represent phases of growth: most structural models are monotonously increasing functions, that were designed to describe the growth of dimensions for which only rising values are possible, and only during the initial part of life (mostly infancy, more rarely adolescence), so their appropriateness to model body weights during the whole lifespan must be carefully evaluated; here only a small selection will be presented and applied to data.

All will be illustrated as if they were referred to the individual: all functions, indeed, can be substituted to the generic function at the level 1 of the nonlinear multilevel mixed effect model that was illustrated before, since they describe the behaviour of each individual.

The first parametric model was elaborated already in 1937 by Jenss and Bayley: it was developed to describe data from birth to approximately 8 years using 4 parameters combined in a function with a linear and an exponential part, accounting for growth and its decreasing rate:

$$y = a + b\,t - \,e^{c+d\,t}$$

Another option is the Count model proposed in 1942, that uses only 3 parameters combined in a linear way

$$y = a + b\,t + c\,\ln(t+1)$$

This model proved to perform slightly worse than the Jenss and Bayley, but both share the quality of remaining robust relative to the choice of starting values for the parameters. The Count model was slightly modified in 1973 by Berkey and Reed in a way that maintained the simple, linear structure but added one or two parameters

$$1^{st}\ order\!:\ y = a + b\,t + c\,\ln(t+1) + \frac{d}{t}$$

$$2^{nd}\ order\!:\ y = a + b\,t + c\,\ln(t+1) + \frac{d_1}{t} + \frac{d_2}{t^2}$$

accommodating for one or two additional inflection points (and so allowing to consider periods of acceleration), and leading to a better fit compared to the previous alternatives.

Adolescence growth was initially analysed using the logistic function

$$y = p + \frac{k}{1+\,e^{a-b\,t}}$$

and the Gompertz function

$$y = p + k\,e^{-e^{a-b\,t}}$$

because they well reflected its main features of a sigmoid trend, starting from a lower "asymptote", experiencing a sharp increase in velocity and, after an inflection point, a decline in the growth rate until the approximation to the upper asymptote, the adult size. Both models proved to perform quite good, but they required the lower age bound (the cut-off between childhood and adolescence) to be determined arbitrarily, so their use became limited in time, in favour of alternatives, such as the Preece-Baines model. Several models were also proposed for the analysis of growth from birth or early childhood until adult age form the '80s, all sharing the fundamental idea of combining different (at least two or three) functions, so that each component could describe the type of growth typical of a smaller period of life. The first were the double- and triple- logistic functions, composed by the sum of separate logistic functions, each to model infancy, mid-childhood and adolescent phases. These were modified by Bock and colleagues in 1994 replacing each logistic with a generalized logistic function; more options were provided, between other, by Shohoji- Sasaki who proposed the Count-Gompertz function in 1987, or by Jolicoeur and colleagues, that created the JPPS, the JPA-1 and JPA-2 models. Despite all these methods would be extremely interesting to analyse more in depth, they will not be further treated here, since they were created specifically to model height/length instead of weight, and they were formulated quite specifically to render the human progress from infancy, trough childhood, the so-called take off, the whole adolescence spurt, until adulthood. Our data hardly share these characteristics, so simpler models will be preferred.

4.3 The analysis of body weights

The analysis of body weights in the first generation of rats, those originally randomized and involved in the experiments, was not performed here, but it was chosen to study only the rats belonging to the second generation, since their characteristic make them more interesting, challenging, and overall more worth a thorough analysis.

We can say that their treatment started in the pre-natal phase, because their dams underwent mating and started gestation one week after the beginning of the experiment, when they were already exposed to the treatment or control regime; moreover, they continued the exposure during the whole period of the pregnancy and the weaning. The observation of the second generation of rats started at 8 weeks of age[8], when all pups have certainly reached a complete independence from their dams.

Age of dams	Males		Females	
	Treated	Control	Treated	Control
30 weeks	74	110	73	98
39 weeks	67	49	65	55
55 weeks	28	32	24	24
Total	169	191	162	177

Table 8: Rats by sex, treatment and age of dams at start of gestation

Male and females from treated and control groups are quite balanced, while there is an important decrease as the age of the dams at the beginning of gestation increases, as we would physiologically expect.

Measurement	Missing	
	n	%
Week 9 (1 of observation)	0	0
Week 30	13	0,018
Week 70	78	0,111
Week 112	348	0,498
Week 114	367	0,525
Week 122	639	0,914
Week 130	688	0,984
Week 146	698	0,998

Table 9: Lost to follow-up in absolute number and percentage at selected measurement occasions

The inspection of individual and mean weights per experimental group and age of the dams at pregnancy showed some important features that are to be considered in the following analysis:

[8] The “time” variable will be centered in the analysis, so that it will not represent the weeks of age but the weeks since the beginning of observation and treatment. The equivalence between these two measures is extremely straightforward (ctime= time-8).

- observations for all animals are present only for the first occasions, then they start to experience mortality; by week 114, when rats are in their elderly age (and measurements of body weight start being performed every 8 weeks) half of the original population remains; since later assessments happens more sporadically, they involve fewer and fewer subjects, as from table 3;

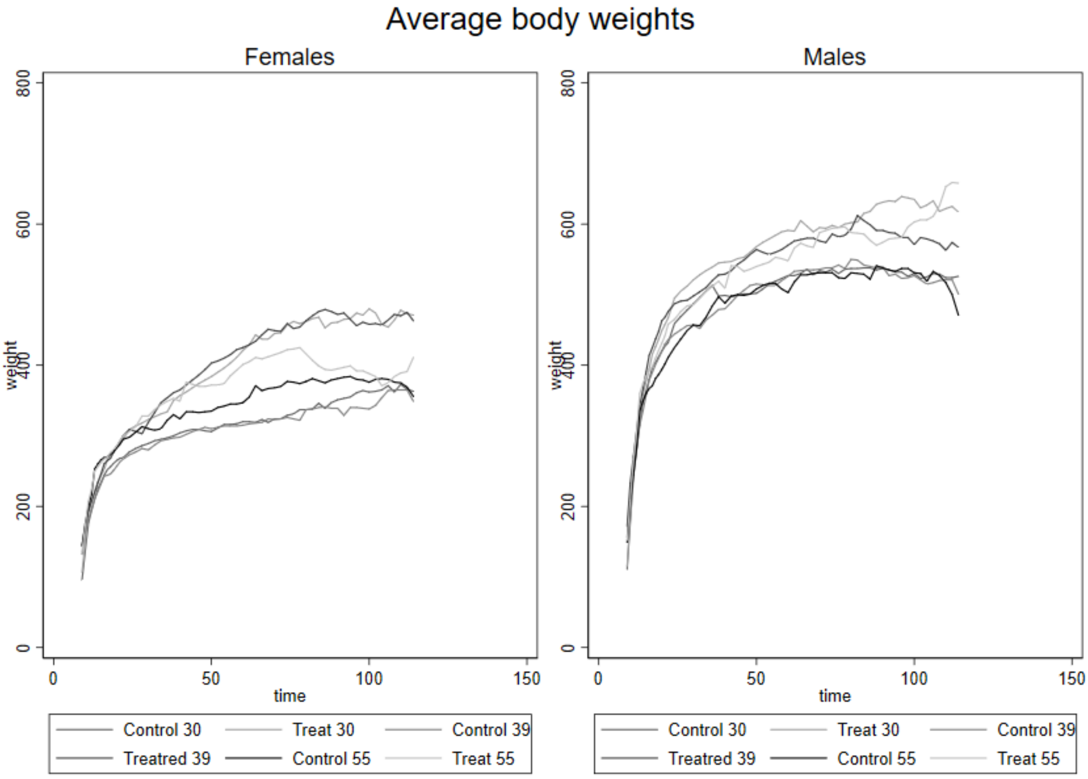


Figure 12: Average weight for females and males by treatment and age of dams at start of gestation

- as expected, males are heavier than females, the treatment proportionally affects body weight, and the age of dams at pregnancy doesn't seem relevant for rats from the control groups, while it corresponds to sensibly different paths for the treated;
- the weights sharply increase during the first weeks of observation, until around the age of 13 weeks, then the growth stops or continue at a slower pace once the adult size is reached;

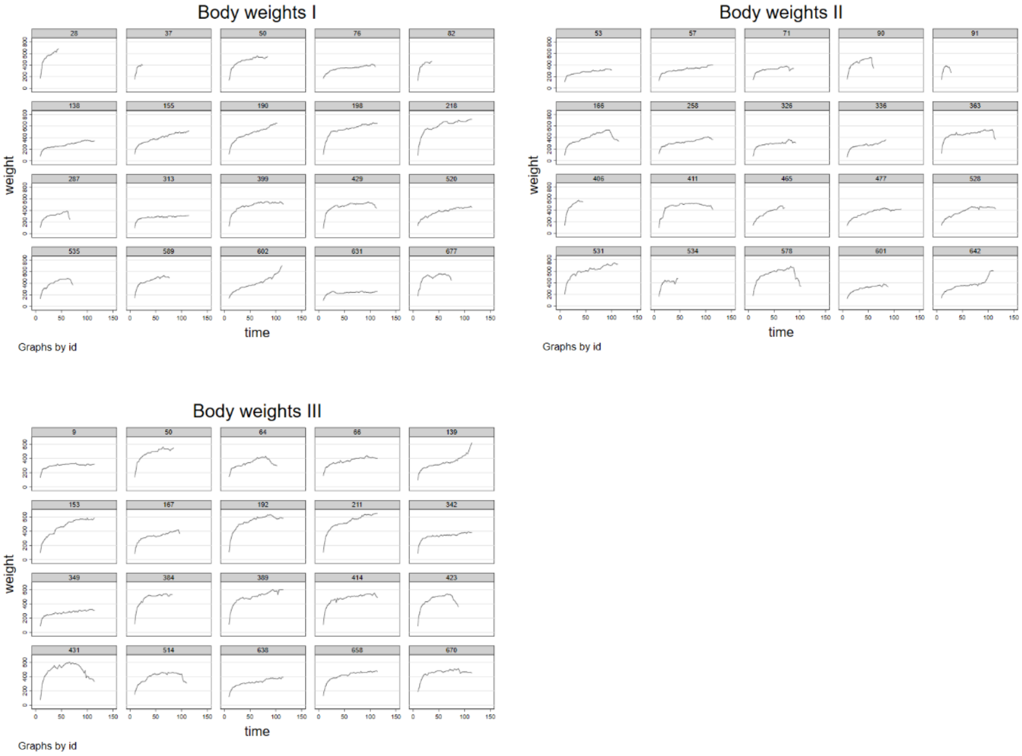


Figure 13: Growth trajectory of body weights from a random sample of II generation rats.

- variability is very important, both within subjects, and among them. Most growth lines present indeed a lot of erraticism that is usually attributable to small variations in the health status of the subjects, that are not relevant but, since the measurements are very frequent, are registered. All trends, nevertheless, are quite similar in shape during the first period of growth, while during the adult/elderly period peculiar patterns appear: very often, weights decrease in the very last part of life, because of diseases, or the physiologic ageing process. Some animals, in contrary, experience an extreme (in magnitude and velocity) increase of weight that is likely due to the presence of neoplastic masses.

The conclusions that can be drown from this exploration of data is that models will probably suffer from this pronounced variability, that it is important to consider all the variables that characterize the data (sex, treatment and the age of dams at gestation), and that linear models can’t be directly fitted to data, so several expedients will be applied, and their use and suitability evaluated. These can be resumed in (I) mathematically manipulating variables so that the linear assumption appears reasonable, (II) using polynomial functions to represent time, so that a non-linear trend can be modelled using a linear function and, finally, (III) fitting nonlinear “human” growth models. Furthermore, since the observations in the last occasions of measurement are drastically reduced and represent very old rats, that likely experience weight variations according to their health status more than to the treatment regime, the observations registered after week 114 will not be considered.

The first step to analyse body weight was fitting linear mixed effects models; to do so, some mathematical manipulation of data was necessary, so that the assumption of linearity remained reasonable; after several tries[9],the log-transformation of the time variable was chosen.

It was chosen to compare the estimates and performance of two models that could be justifiable and meaningful for these data: one only allows for random intercept and slopes for each litter, since, at least in the first part of life, the weights of each pup within the same litter tend to be quite similar, but the differences among litters can be remarkable; they tend to level-up with time, but a great variability at individual level remains. The alternative possibility was to include a random intercept only at the litter level, and to include the subject level in the analysis, allowing every subject within each litter to have its own random intercept and slope. Information criteria and likelihood-ratio tests favoured this last option, since such a setting would probably help explaining the peculiar growth paths in adult/elderly life we highlighted before, refining the fit of the models and reducing the unexplained variability; on the other hand, it probably unnecessary weights the model down.

Fixed effects were assessed in a very straightforward way, since they consist of experimental conditions: the relevance of sex, treatment received, and the age of the dams at the beginning of gestation was evaluated, and the first two were found to have a relevant influence on body weights. Variables and model selection were performed according to statistical significance of the estimates, to the results of

[9] Weight raised at the power of 2, 2.5, 3, 3.5, 4; natural logarithm of time, time$^{1/2}$, 1/time, 1/time^{2}.

likelihood ratio tests in case of nested models, or considering information criteria as AIC and BIC, and, as usual, pairing all this with the knowledge of the phenomenon under study and biological meaning. The last consideration regards the fact that this type of repeated measurements can't be independent between subsequent assessments, but each value highly depends on the previous ones, with correlation decreasing the more the observations become far apart in time. To account for this, the possibility of shaping the structure of the residual errors at the *individual* level was evaluated. The variance-covariance structure was built to have serially correlated residuals using an exponential specification, where

$$cov(\varepsilon_{it}, \varepsilon_{it'}) = \sigma_{\varepsilon}^{2}\ exp\{-\gamma(t' - t)\}$$

which tends to the variance the closer are the measurements, and exponentially decreases to zero as they become more distant in time.

Observations	36,606	Model 1		Model 2	
Number of groups	98	Coef	s.e.	Coef	s.e.
Fixed part					
Time (ln(centered time))		77.688***	(1.980)	76.405***	(0.914)
Treatment		12.170**	(5.290)	17.858***	(4.837)
Sex		158.158***	(0.543)	66.573***	(3.144)
Constant		67.011***	(4.513)	100.786***	(3.810)
Random part					
Litter: Unstructured (1)	sd(logct)	19.401	(1.430)		
Identity (2)	sd(_cons)	35.886	(3.193)	17.480	(2.325)
	corr(logct,_cons)	-.707	(.061)		
Subject: Independent (2)	sd(logct)			18.329	(.782)
	sd(_cons)			1.49e-07	.
Residual: Exponential (2)	rho(e)			.395	(.015)
	sd(Residual)	47.552	(.176)	51.360	(1.007)
Goodness of fit					
Log restricted-likelihood		-193723.6		-155124.1	
AIC		387463.1		310264.3	
BIC		387531.2		310332.4	

*** p<0.01, ** p<0.05, * p<0.1

Table 10: Multilevel mixed effects models using transformed variables; results of regression of weight on ln(time, centered) and experimental variables.

The models give quite similar results in terms of significance and type of effect of each variable. If we consider only random differences at the litter level, we find out that female rats weight approximately 70 g at their first assessment at 8 weeks of age while males are sensibly heavier, of around 160 g; treatment is responsible of a smaller, but still relevant, increase in weight, of around 12 g. The physiological growth that all animals experience in time is around 77 g for females of the control group, per unit of time (which is on logarithmic scale).

The graphical representation of the fixed effects in figure 4 and6, and of the fixed and random effects for some selected litters in figure 5may help clarify these findings.

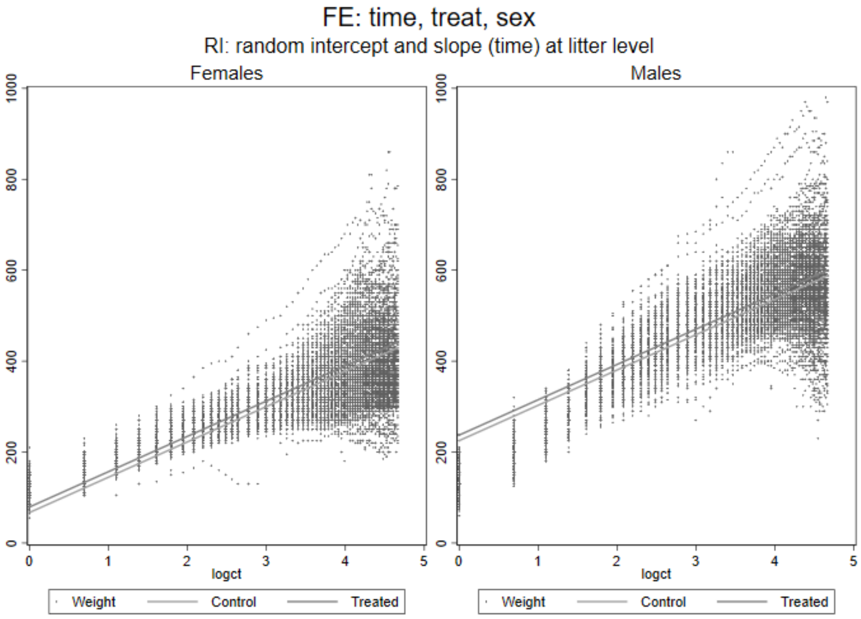


Figure 14: Plot of data and estimated fixed effects, model 1

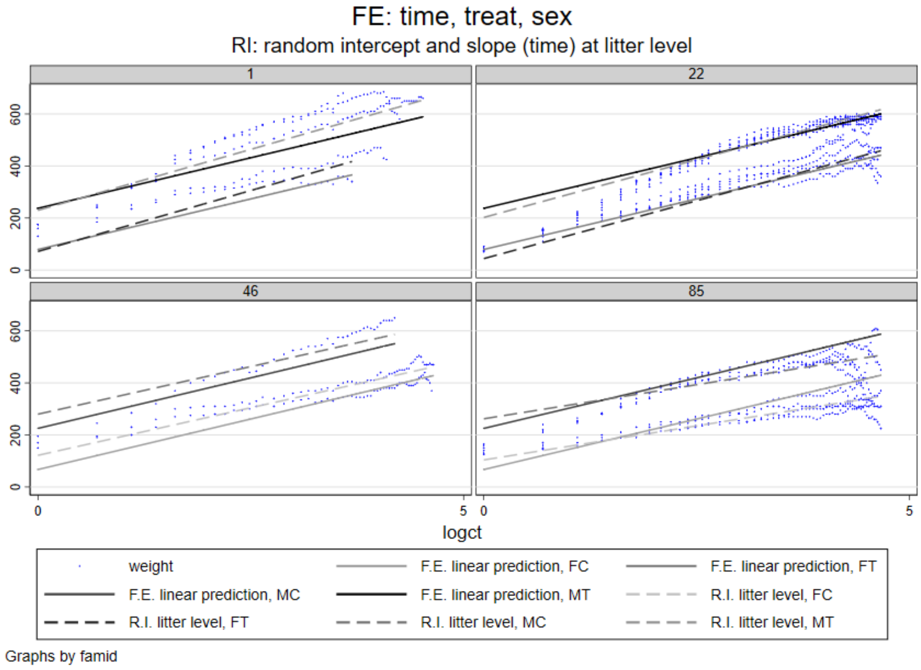


Figure 15: Data and linear prediction with fixed and random portion in four selected litters, model 1.

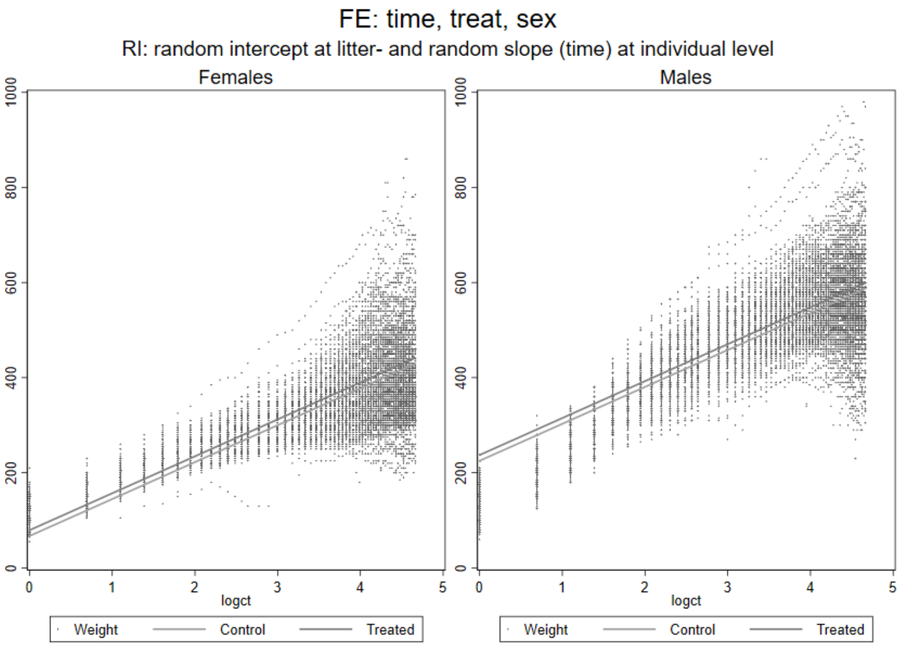


Figure 16: Plot of data and estimated fixed effects, model 2.

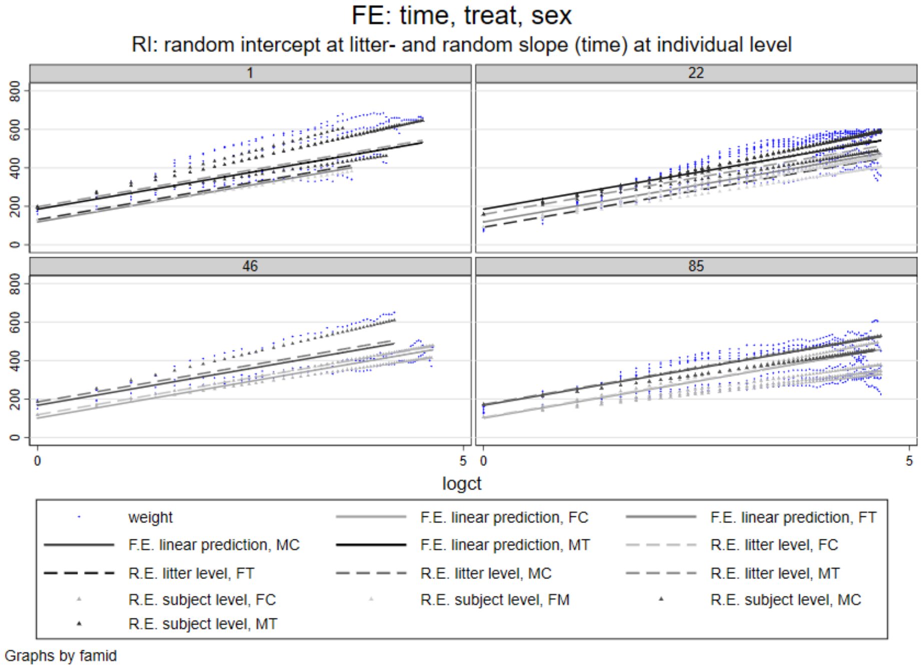


Figure 17: Data and linear prediction with fixed and random portion in four selected litters, model 2.

Residuals were analysed to verify whether the underlying basic assumptions (linearity of the relationship between the outcome and the regressors, normality and homoscedasticity of residuals) were met, and they highlighted several problems.

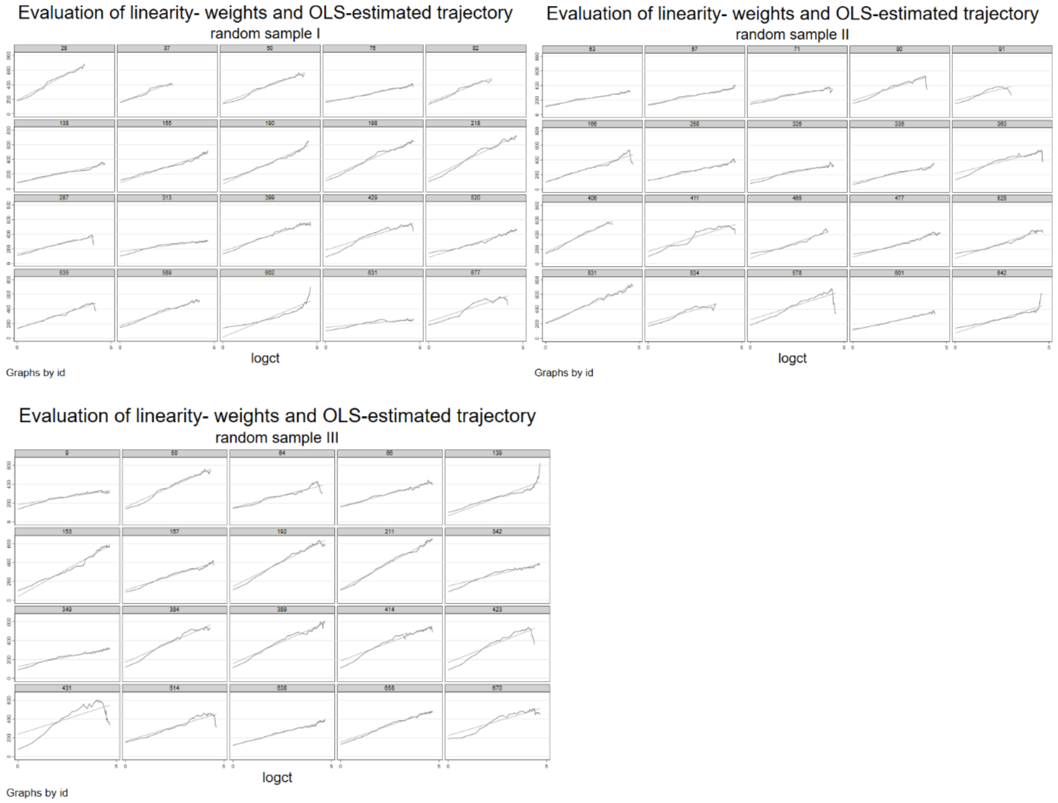


Figure 18: Evaluation of linearity: comparison between growth trajectory and linear regression in three random samples

As it was clear from the beginning, linearity is hardly respected even when a transformation of the variables is used: the main problem clearly arises in the relation between growth of body weights and time, since even if a linear trend can approximate reasonably well the growth in the first period of life, the unpredictable variability in the last phase is often too high.

The representation of normal probability plots and standardized residuals estimated for each level of the models shows more issues that can probably be attributed, again, to the extreme variability of body weights among rats: too many observations present extreme values both in the upper and in the lower tail of the distribution.

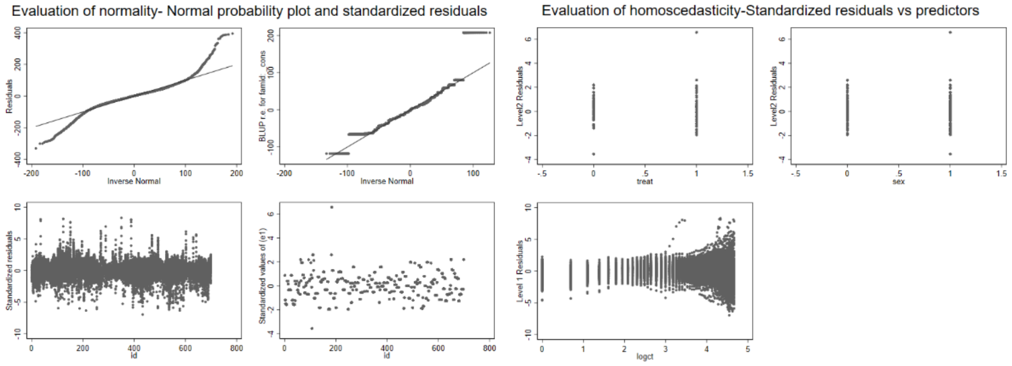


Figure 19: Evaluation of normality and homoscedasticity of residuals using Normal probability plots and standardized residuals.

Even the assumption of homoscedasticity raises some doubts: coherently with what was highlighted so far, body weights can grow to quite extreme values, or fall sharply in relatively short time, mostly in adult and aged rats. The fact that standardized residuals tend to increase with time is therefore not so surprising.

These analyses show that indeed linear mixed effect models are a powerful tool to represent the multilevel nature of data, to account for the fact that they are repeated measurements of the same physical trait, and to handle the lack of randomization of this particular dataset, too. The great variability in the possible growth paths and its non-linear trend in time doesn’t allow, nevertheless, to consider those illustrated above as the optimal models for this type of analysis, and to take the estimates as definitive. Given these drawbacks, a different approach was tried: the same models were fitted using polynomial terms to represent time, so that the trajectory of growth could be better represented.

The models were built considering again sex, experimental treatment and age of each dams at the beginning of gestation as covariates to include in the structural part; it was chosen to include a random effect depending on the variable “time” at the familiar level.

Observations	36,606	Quadratic		Cubic	
Number of groups	98	Coeff	s.e.	Coeff	s.e.
Fixed Part		.		.	
Time		6.906***	(0.178)	11.871***	(0.266)
Time ^{2}		-0.048***	(0.002)	-0.174***	(0.006)
Time ^{3}				.0008***	(.00003)
Treatment		9.039*	(5.224)	11.270**	(5.221)
Sex		158.225***	(0.613)	158.184***	(0.569)
Constant		162.674***	(3.790)	124.929***	(3.975)
Random Part					
Litter:					
Unstructured	sd(ctime)	1.715	'(.102)	2.456	.
	sd(ctime2)	.015	(.001)	.049	.
	sd(ctime3)			.0003	.
	sd(_cons)	27.184	(2.188)	29.202	.
	corr(ctime,ctime2)	-.901	.	-.883	.
	corr(ctime,ctime3)			.752	.
	corr(ctime,_cons)	-.209	.	-.401	.
	corr(ctime2,ctime3)			-.962	.
	corr(ctime2,_cons)	.342	(.023)	.438	.
	corr(ctime3,_cons)			-.368	.
	sd(Residual)	53.517	(.198)	49.662	.
Goodness of Fit					
Log restricted likelihood		-198178.5		-195553.3	
AIC		396377		391118.7	
BIC		396462.1		391169.7	

*** p<0.01, ** p<0.05, * p<0.1

Table 11 : Multilevel mixed effects models using polynomial terms; results of regression of body weights on second and third degrees polynomial terms for time and experimental variables

Quadratic and cubic models were evaluated adding the second and the third power of variable “time” in the model equation, and the cubic was found to be the better for these data. The results are not so dissimilar to those obtained before, most of the differences among body weights can be explained by sex, but the consumption of Coca Cola ad libitum is a relevant factor, as well.

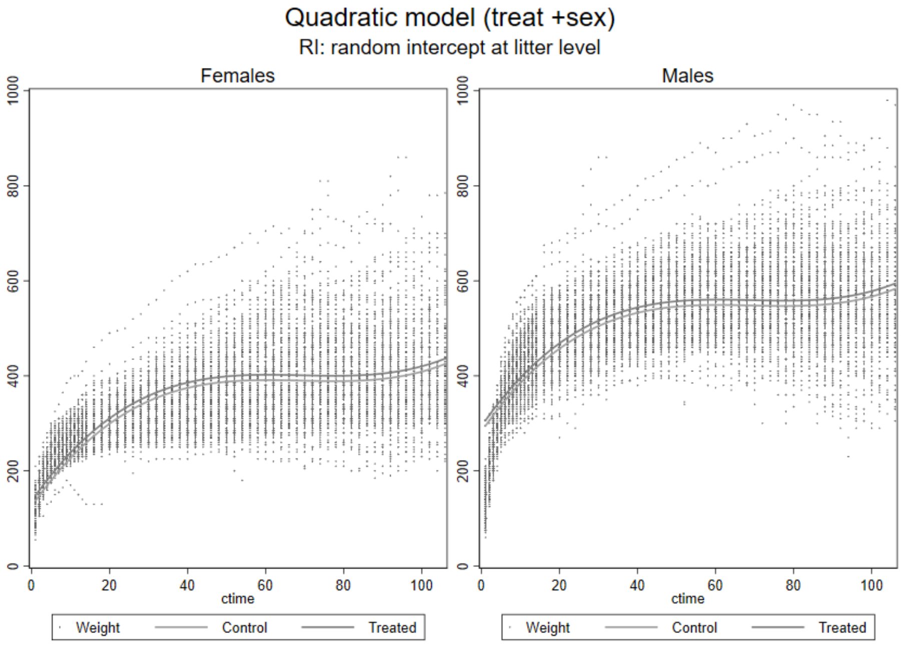


Figure 20: Plot of data and estimated average predictors, cubic model.

Differences among litters are relevant, too, so it’s worthy adding random effects at this level.

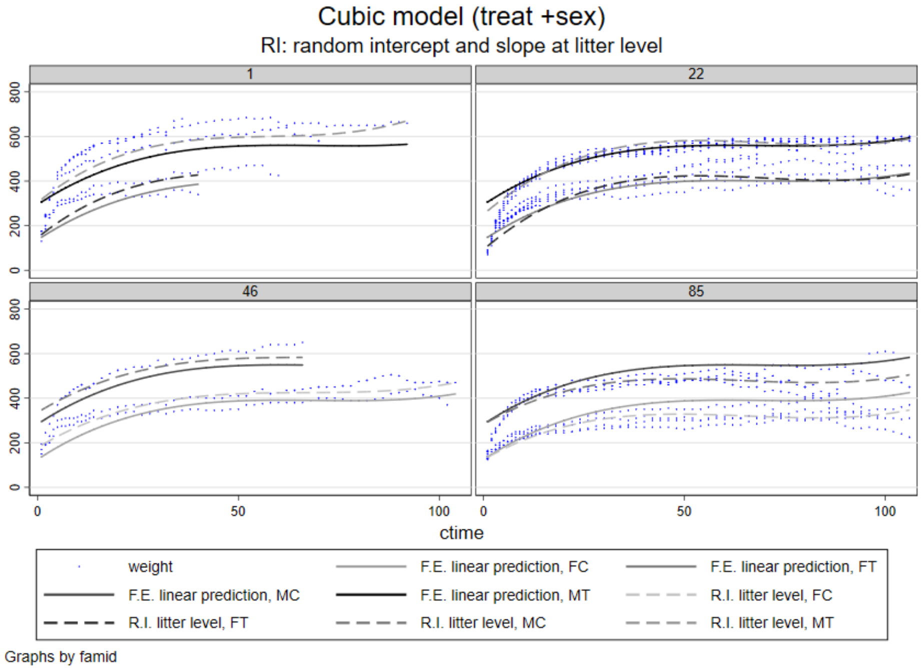


Figure 21: Data and linear prediction with fixed and random effects in four selected litters, cubic model.

The display of residuals highlights that the problems arising from the extremes behaviour of weights of elderly or ill rats remains quite evident: residuals can't really be said to distribute normally, and appear quite heteroscedastic in relation to time, too, as can be seen in figures 12.

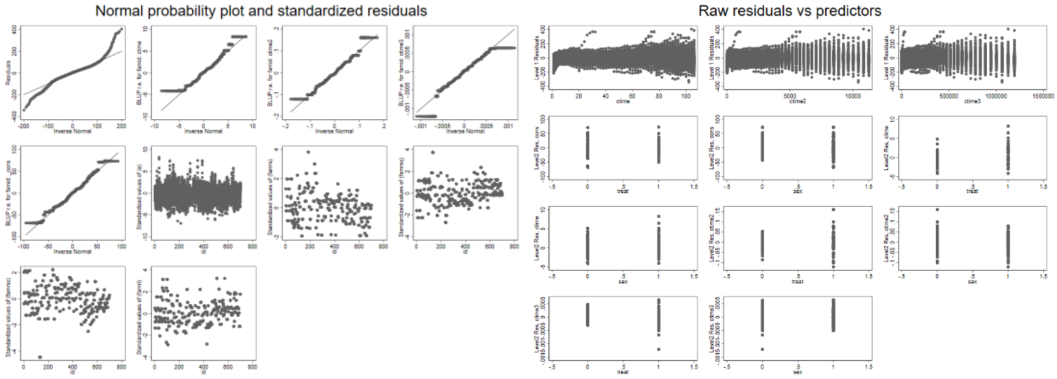


Figure 22: Evaluation of the assumption of normality and homoscedasticity of residuals, cubic model.

Polynomial terms have proven to be a worthy expedient to be able to model curve trajectories remaining in the framework of linear mixed effects models, but the noisy nature of data continues to create issues that this model can hardly handle and that may prevent to consider the results reliable.

A third option was attempted, just in an explorative way: a generalized additive mixed effect model was fitted, where time was not included directly, but trough a smoothed function. It was then added to the other regressors, always keeping a multilevel structure with random effects at litter level. This attempt proved quite effective in taking into account time and model its irregular trend in quite a simple way. It may be proficiently used to explore data, and eventually to model them, but the thorough evaluation of this method is left for the next future.

In this occasion it was chosen to explore more in depth the models for human growth, and to try to apply them to animal growth. The choice of which to use was based on the possibility to adapt them to the available data in a relatively easy way, that means that the models had not to be too strictly bonded to characteristics, rhythms and timings typical of human growth only. It was indeed showed (Sengupta 2013) that rat and human life are correlated, taking the whole life span into account, so that one human year equals 13.2 rat days; it is better, nevertheless, to consider different phases separately, since while humans develop very slowly, rats have a very accelerated pace, and reach sexual maturity around 6 weeks of age: according to this equivalence, the rats that we are observing could be considered adolescents, but as becomes clear from Table 5, it is impossible to draw an exact and constant equivalence between rats' age and humans' age.

Period	Correspondence to 1 human year
Entire life span	13.2 rat days
Weaning period	42.4 rat days
Pre-pubertal period	3.3 rat days
Adolescent period	10.5 rat days
Adulthood	11.8 rat days
Aged phase	17.1 rat days
Average	16.4 rat days

Table 12: Correspondence of one human year with rat days at different phases of life, from Sengupta (2013)

This is the reason behind the choice of not considering those parametric, nonlinear models that were designed to account for the specific features and mechanisms of human growth, but to focus on those that could be used to describe a path, similar for intensity and velocity to that of humans during young age. Some of the models that might answer to these needs are the Jenns and Bayley, the Count and the 1st order Berkey and Reed model, that are usually employed to analyse growth during adolescence, and are therefore built to describe a steep growth during the first part of the curve, that slows down at the reaching of the approximate adult size. They were fitted to data to verify which could be the best, and the choice fell on the Berkey and Reed model, basing the decision on the graphical examination of the curves and on the comparison of the measures of goodness of fit. The model uses four parameters to describe the specific functional form of the individual growth curve

$$1^{st}\ order\colon\ y = a + b\,t + c\ \ln(t+1) + \frac{d}{t}$$

that may represent the starting point, growth rate, acceleration and deceleration of growth, so that the trajectory is allowed to be curvilinear and to have one or more inflection points. Here, the function was built so that each of the parameters would be dependent on age[10], sex and treatment received; a random effect at the litter level was introduced and evaluated for all of them, so that each litter is not constrained to have the same intercept, or, for example inflection point.

		Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]
a	sex	565.789	(33.214)	17.035	(0.000)	500.691 - 630.887
	treat	-380.528	(33.304)	-11.426	(0.000)	-445.803 - -315.252
	_cons	655.279	(28.321)	23.137	(0.000)	599.770 - 710.788
b	sex	-0.150	(0.122)	-1.224	(0.221)	-0.390 - 0.090
	treat	-0.846	(0.196)	-4.310	(0.000)	-1.230 - -0.461
	_cons	1.567	(0.147)	10.639	(0.000)	1.279 - 1.856
c	sex	78.149	(9.415)	8.301	(0.000)	59.696 - 96.601
	treat	-118.262	(9.444)	-12.522	(0.000)	-136.773 - -99.751
	_cons	89.059	(8.016)	11.110	(0.000)	73.347 - 104.770
d	sex	-3,381.295	(137.819)	-24.534	(0.000)	-3,651.416 - -3,111.175
	treat	1,330.012	(138.127)	9.629	(0.000)	1,059.288 - 1,600.736
	_cons	-3,125.875	(117.836)	-26.527	(0.000)	-3,356.830 - - 2,894.921
Litter: Identity	var(U0)	'.567	'.081			.427 - .752
var(Residual)		1966.77	14.557			1938.44 – 1995.51

Figure 23: Estimates from Berkey- Reed model, offspring

The estimated parameters are not easy to interpret; what is relevant is that the treatment is responsible for significant differences in all parameters. This becomes very clear if we represent the curves graphically: the plot for some randomly selected litters shows that the model fits most of the growth trajectories quite good, while the average curves for treated and controls, separated by sex, show that both these covariates affect sensibly body weights.

[10] This time was considered without centering or transformation, since this model requires time to be expressed as “age”.

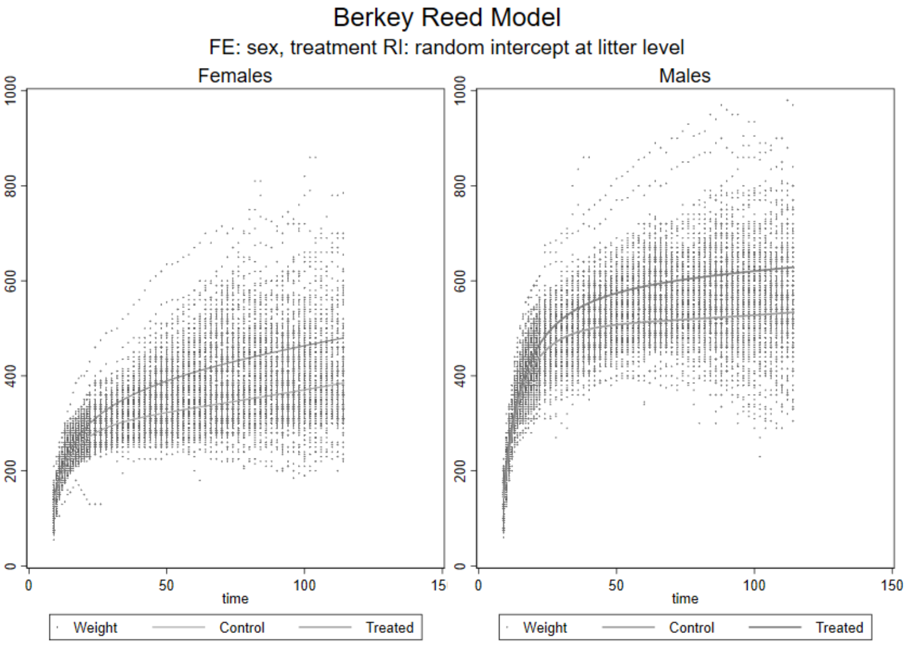


Figure 24:Plot of data and estimated average predictions, Berkey-Reed growth model

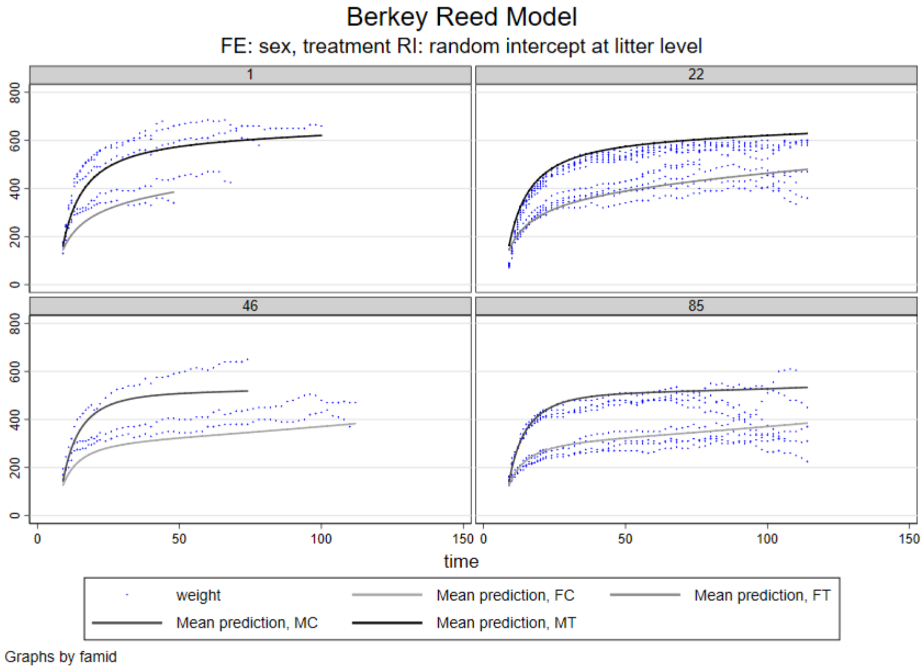


Figure 25: Plot of data and linear predictions with fixed and random part in four randomly selected litters, Berkey-Reed growth model

The examination of residuals points out the same problems that emerged with all other analysis: the residuals can hardly be considered to respect all the assumptions that underlie the model.

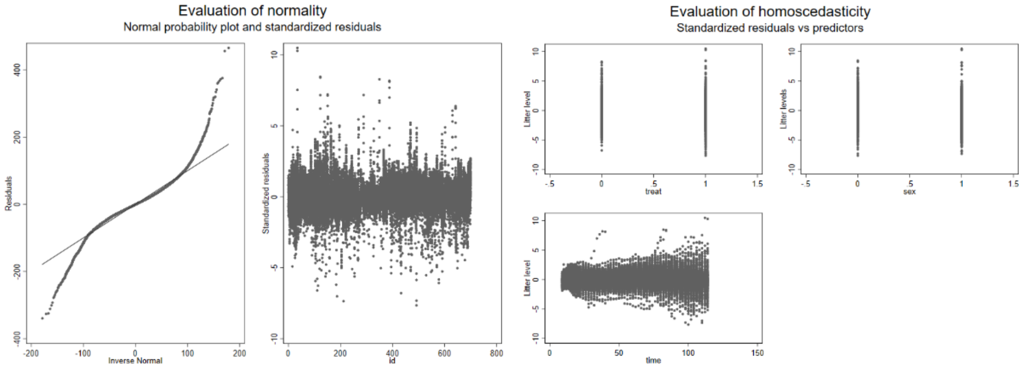


Figure 26: Evaluation of the assumptions of normality and homoscedasticity of the model using residuals.

At this point it is useful to draw some conclusions: these data have some very peculiar characteristics, the most relevant for their analysis are non-linearity and the fact that they can take unexpected, extreme turns upwards or downwards, mostly when rats are close to the end of life, reflecting the presence of important neoplastic masses or a worsening of the health conditions. These features should discourage the use of methods based on the comparison of measures of synthesis like the group means, because they could be heavily affected by the atypical recordings, giving an unrealistic picture of the situation and possibly preventing to detect differences caused by experimental factors.

The use of multilevel mixed effects models is surely to be encouraged, since it allows to analyse directly the recording of each subject, without concerning about the differences in the duration of the recordings. They are also a precious tool in case of clustered data, and in this rather peculiar experimental design, when no randomization was performed on a whole cohort of rats. The most straightforward specifications of such models using a proper linear function is not the best option, because it requires a transformation of the variables, so that the advantage of a simple functional forms is counterbalanced by the difficult interpretation of transformed variables. Even the introduction of polynomial terms, that allow to represent a curve trajectory remaining in the frame of a linear function, slightly reduces the ease of the interpretation, but it can still be acceptable if it allows a more faithful representation of the growth trajectories; as their plot showed, nevertheless, it’s not always the case. The methods that we feel to choose as preferred are the nonlinear growth models that were “borrowed” from human studies: a wide variety exists, so one can select the most appropriate every time, according to the characteristics of data. In this case, too, the model parameters are not easy to interpret, but a clear indication is provided about the statistical significance of each covariate and its ability to affect the outcome.

Chapter 5 Conclusions

The research centre that asked for this collaboration, that concretized in this research, is active since decades in testing the carcinogenic potential of substances and compounds that the population is frequently exposed to, for occupational or lifestyle reasons, and was in several occasions the first to outline the health risks connected to chemicals that are today recognized as human carcinogens (see (S. C. Maltoni C. 1979) and (C. B. Maltoni C. 1983) for the example of benzene, or (Belpoggi F 1995) for Methyl-tert-Butyl Ether). Nevertheless, in some occasions, the results obtained from its experiments failed to be recognized by the regulatory organs in charge because they were not sufficiently strong and universally accepted from a methodological point of view, rather than because of a weak scientific evidence.

Deepening the knowledge and understanding of the methods used for the statistical analysis of their results is one of the strategies to enhance the quality of the research, and it is the ground on which a strong scientific base can be built. This work is an attempt in this direction: it aims to go beyond the techniques that are performed routinely, to explore the characteristics of the data and to try to understand the mechanisms that determined them, and, in this framework, to propose some methodologies that manage to answer the research questions in a more comprehensive way.

The data were chosen among those that were published long before, did not contain any sensitive or reserved information and resented some features that made them particularly challenging for the researchers to analyse. The choice fell on the Coca- Cola study, a series of 4 experiments that started in the late ‘80s and whose results were not found to be particularly controversial or worrying, and were indeed published only in 2006. They presented the particularity of involving two generations of subjects, the first randomized, the second included in blocks into the experiment, to evaluate the effect of exposing pups in a particular window of susceptibility, the perinatal period. They were also one of the few cases where the exposure determined an important increase in body weight, so that this measure needed to be considered with more attention than usual.

This work focused on two topics that are not the main focus of the interest in carcinogenicity studies, but deserve a thorough analysis for the reasons that were illustrated at the beginning of each of the two chapters dedicated to data analysis.

For the analysis of survival, the time-to-death of each individual was examined, to evaluate the effect of the experimental regime received; two separate analyses were performed, one involving breeders, the other offspring only, because we assume the former observations to be independent, while the latter are taken on siblings, so independence can’t be assumed. The effect of the exposure was evaluated, while it was necessary to control for the influence of other likely risk factors: this was possible thanks to the modelling approach. Proportional hazard models were fitted, and the Accelerated failure-time model, parametrized using a generalized gamma distribution, was used when the assumption of PH didn’t hold. The use of frailties was evaluated: a univariate term was introduced for the breeders, at the individual level, to control for possible unobserved heterogeneity; a shared frailty was used to account for the lack of independence among subjects from the second generation, that contained many siblings. The results say that the treatment appears relevant in accelerating death among the individuals exposed from adult life; the frailty could not be introduced in the best-fitting model because of computational problems, so its influence remains to be verified. The models estimated for the rats of second generation initially confirmed the role of the treatment in increasing the risk of death, but when a term for a common, unobserved source of heterogeneity was introduced to consider the effect of familiarity among members of the same litters, it became the only relevant variable.

These findings underline some very important principles: the first is the importance of properly choosing the methods and specifying the models, where properly means in a data-and-experience-driven way. A thorough knowledge of the data and of the dynamics that contribute to determine them is always a good starting point to build plausible, really representative and meaningful models. Another crucial point is the fact that model checking and verification of the respect of the assumptions that lie at the foundations of any method should become a routine embedded in every analysis, while it remains not yet so common (or, at least, not always explicitly reported) in the literature regarding carcinogenicity bioassays.

The methods for the analysis of time-to-event are really flexible: here they were use for the most “literal” application, to evaluate the general survival of the population under study. They would be suitable for several other applications, if the collection of additional information was possible: it would be interesting for example to build a classification of tumours based on their lethality or incidentality; being able to attribute a cause of death to each animal, or to register the time of onset for those (rare) type of lesions that are detectable while the animals are still alive. The collection of such additional information would greatly enhance the possibilities for analysis, making possible to consider cause-specific mortality or methods to handle competing risks. Another interesting possibility that can immediately be evaluated and implemented with the information available so far is the joint modelling of longitudinal measures (such as body weights) with survival data.

The second part of this work was dedicated to the analysis of the longitudinal measurements of body weights from the group of rats of second generation. The main purpose was to find the best modelling approach to adequately describe the process of growth of these animals, from their youth to the reaching of the adult size; once this is assessed, the influence of the exposure to Coca-Cola and the effect of other covariates has been also evaluated.

The data consisted of weekly recordings of body weights of young rats, followed from the age of 8 weeks until spontaneous death: the observations are very numerous and were collected regularly; they present a high variability among individuals that is at least partially expected, because males and females can have very different sizes, and sometimes a very high variability within individuals as well, that is due to quite extreme weight increases or decreases due to health issues or mortality.

All analyses were made in the framework of mixed- effects models, because they are the best approach to treat longitudinal data and allow to include additional level of grouping, that in this case consists of litters. It is therefore possible to account for the structural effect of covariates, that we expect to act the same way on all individuals, and to add random effects that introduce a correlation among the subjects within the same group (here: within the pups of the same litter.

Mixed effects can be included in a variety of models: here several were explored, and their suitability has been evaluated. The first was a linear formulation, which is reasonable only if variables undergo a transformation to linearize the trend of body weights, that increase very rapidly in the first period and gradually slow down. The log-transformation of the time variable was used. The second was to build a linear function with polynomial terms for time, to allow the trajectory to assume a curve trend; the cubic function of time, that allows for two inflections points, was found to be the best option. Finally, growth models were fitted: they are non-linear models that were specifically created to study human development. We selected those that could easily be adapted to the animal context, since they were less sophisticated to catch characteristics specific of human and infant growth. The Berkey and Reed model proved the best for this purpose.

Even if it's not possible to directly compare the results of these models using the measures of information criteria, we can still draw some conclusions. Growth models represent an interesting way to analyse the body weights of young rats, since not only they allow to evaluate the differences among experimental groups, but also give information about the growth process. Choosing the appropriate formulation for the data allows to represent quite precisely the shape of the trajectories; they may become useful in those studies where the development and sexual maturation of the animals are investigated, because they can be used to obtain an estimate of the start and development of the adolescent period.

They were not explicitly considered here, but these analyses of course are suitable for adult rats as well; for this purpose, a linear model or a quadratic polynomial of time should be the best options, because rarely growth models are built to model the weight trajectories during the whole lifespan of individuals, so a simpler and more efficient alternative is to be preferred.

Some issues remain open, like the problem of how to consider and treat the extreme trends that some individual experience in the last part of their life; it is an interesting feature, that is usually associated with the worsening of the health conditions, but it can also affect heavily the estimates and the overall likelihood of the models. It could be useful to this regard to further study and develop the application of general additive mixed models to these data, that were only briefly explored during this analysis but seemed quite promising. Allowing to include a nonparametric, smoothed function of time in the classic linear regression function, this approach can handle variables with an irregular trend like these body weights, and at the same time maintain a simple and understandable interpretation for the experimental variables.

References

Belpoggi F, Soffritti M, Maltoni C. “Methyl-tert-Butyl Ether (MTBE) – a gasoline additive – causes testicular and lymphohaematopoietic cancers in rats.” *Toxicol Ind Health* 11 (1995): 119-149.

Bretagnolle J., Huber-Carol C. “Effects of omitting covariates in Cox’s model for survival data.” *Scandinavian Journal of Statistic* 15 (1988): 125–138.

Buckley I. V., James I. “Linear regression with censored data.” *Biometrika*, 1979: 429–436.

Calle E.E., Rodriguez C., Walker-Thurmond K., Thun M.J. “Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. Adults.” *New English Journal of Medicine*, 2003: 1625-38.

Cancer, International Agency for Research on. “Preamble.” In *IARC monographs on the Evaluation of Carcinogenic Risks to Humans*. Lyon: WHO- IARC, 2006, last update September 2015.

Clayton, D.G. “A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence.” *Biometrika* 65 (1978): 141 - 151.

Clayton, D.G. “A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence.” *Biometrika* 65 (1978): 141-151.

Collaboration for Global Burden of Disease Cancer. “The global burden of cancer 2013.” *JAMA Oncology* 1, no. 4 (2015): 505-527.

Collett, D. *Modelling survival data in medical research.* 2nd edition. Boca Raton: Chapman &Hall/CRC, 2003.

Cook R.D., Johnson M.E. “A family of distributions for modeling non-elliptically symmetric multivariate data.” *The Journal of the Royal Statistical Society* B, no. 43 (1981): 210 - 218.

Cox, D. R. “Regression models and life tables (with discussion).” *Journal of the Royal Statistical Society* B, no. 74 (1972): 187-220.

Fitzmaurice G, Molenberghs G. “Advances in longitudinal data analysis: an historical perspective.” In *Longitudinal data analysis: a Handbook of modern statistical methods*, by Davidian M, Verbeke G, Molenberghs G. Fitzmaurice G, 3-30. London: Chapman & Hall/CRC Press, 2008.

Hothorn, L. “Statistical evaluation of toxicological bioassays- a review.” *Toxicol Res* 4 (2014): 418.

Hougaard, P. *Analysis of multivariate survival data.* New York: Springer, 2000.

Laird N M, Ware J H. “Random effects models for longitudinal data.” *Biometrics* 38 (1982): 963–974.

Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K. “Body fatness and cancer – viewpoint of the IARC Working Group.” *The New England Journal of Medicine,* 2016: published online, 25 August 2016.

Lindstrom M., Bates D. “Nonlinear mxed effects models for repeated measures data.” *Biometrics* 46, no. 3 (September 1990): 673-668.

Maltoni C., Conti B., Cotti G. “Benzene: A multipotential carcinogen. Results of long-term bioassays performed at the Bologna Institute of oncology. .. ,), pp. .,.” *American Journal of Industrial Medicine* 4, no. 5 (1983): 589-630.

Maltoni C., Scarnato C. “First experimental demonstration of the carcinogenic effects of benzene. Long-term bioassays on Sprague-Dawley rats by oral administration.” *Medicina del Lavoro* 70 , no. 5 (1979): 352-357.

OECD. “Guidance document 116 on the conduct and design of chronic toxicity and carcinogenicity studies, supporting Test Guidelines 451, 452 and 453, 2nd edition.” Paris: OECD, 13 April 2012.

OECD. *Guidance Notes for Analysis and Evaluation of Chronic Toxicity and Carcinogenicity Studies.* Organisation for Economic Co-operation and Development, 2002.

OECD. *Test No. 451: Carcinogenicity Studies.* Paris: OECD Publishing, 2009.

RAPP, K. et al. “Obesity and incidence of cancer: a large cohort study of over 145000 adults in Austria.” *British Journal of Cancer* 93 (2005): 1062–1106.

Sengupta, P. “The laboratory rat: Relating its age with human’s.” *International Journal of Preventive Medicine* 4 (2013): 624-30.

Singer J. D., Willet J. B. *Applied longitudinal data analysis. Modeling change and event occurrence.* Oxford: Oxford University Press, 2003.

Therneau T.M., Grambsch P.M. *Modeling Survival Data.* New York: Springer, 2000.

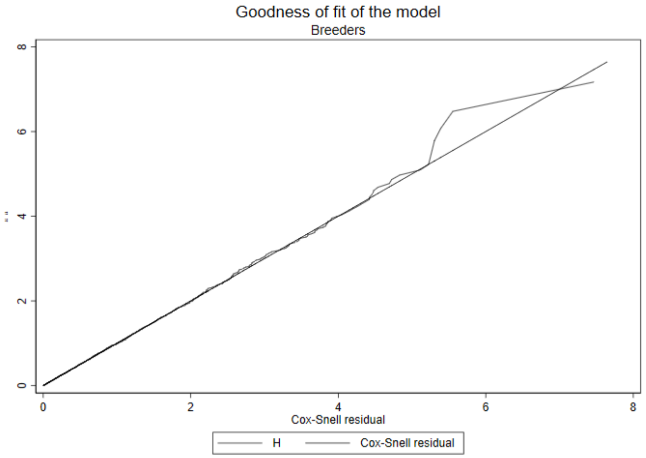
Vaupel J.W., Manton K.G., Stallard E. “The impact of heterogeneity in individual frailty on the dynamics of mortality.” *Demography* 16 (1979): 439–454.

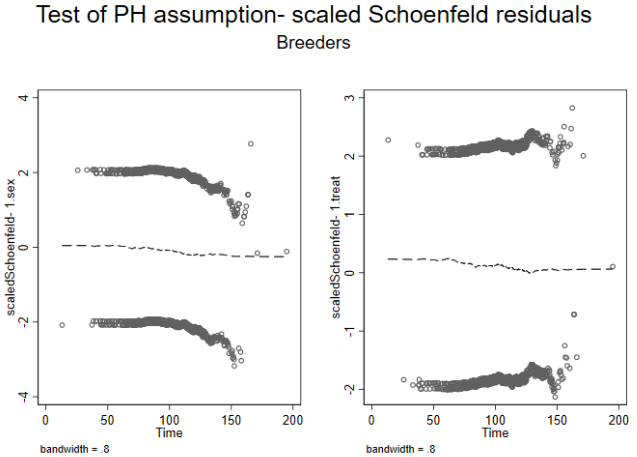
Wei, L.J. “The accelerated failure time model: A useful alternative to the cox regression model in survival analysis.” *Statistics in Medicine* 11, no. 14-15 (1992): 1871–1879.

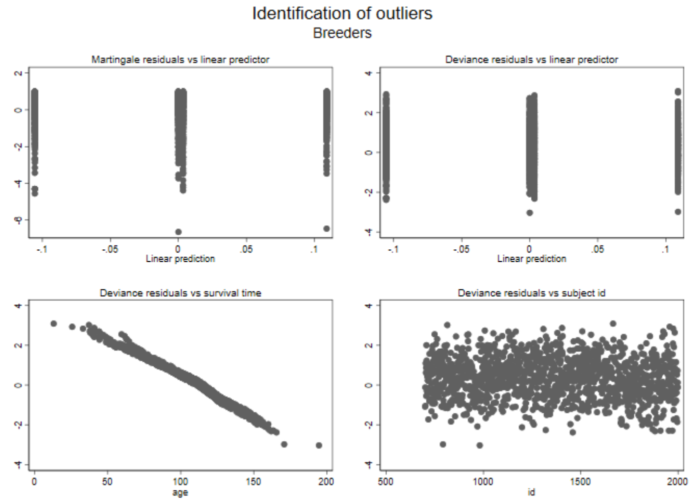
WHO. *IPCS Environmental Health Criteria 70: Principles for the safety assessment of food additives and contaminants in food.* Geneva: World Health Organization, 1987.

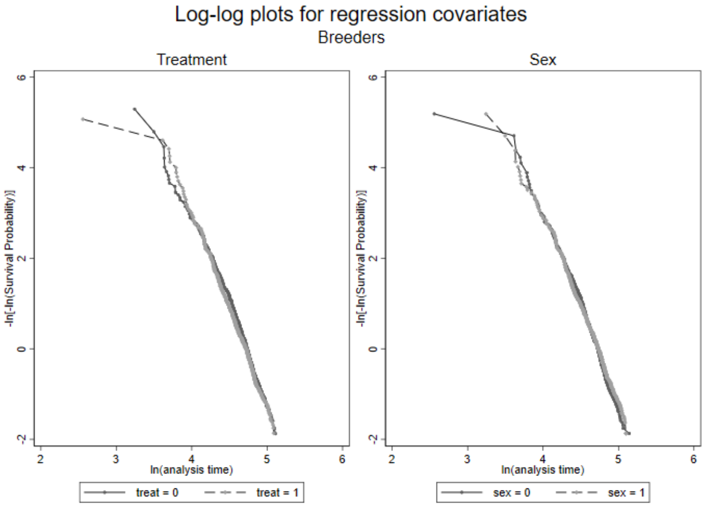
Appendix 1: Analysis of residuals for survival analysis

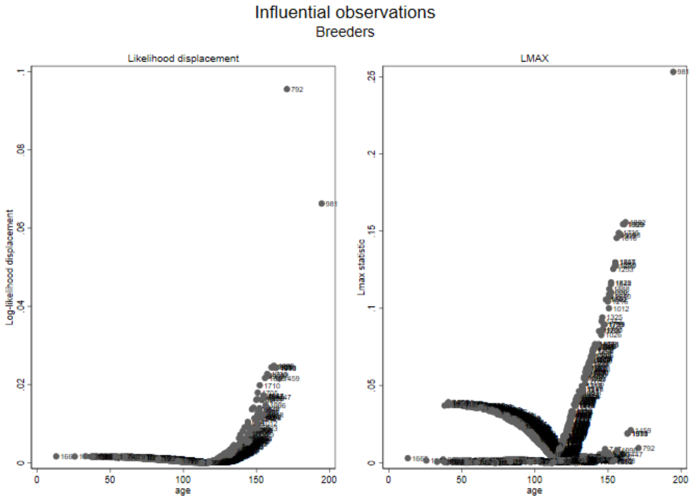
1.1 Cox Proportional Hazard model, breeders







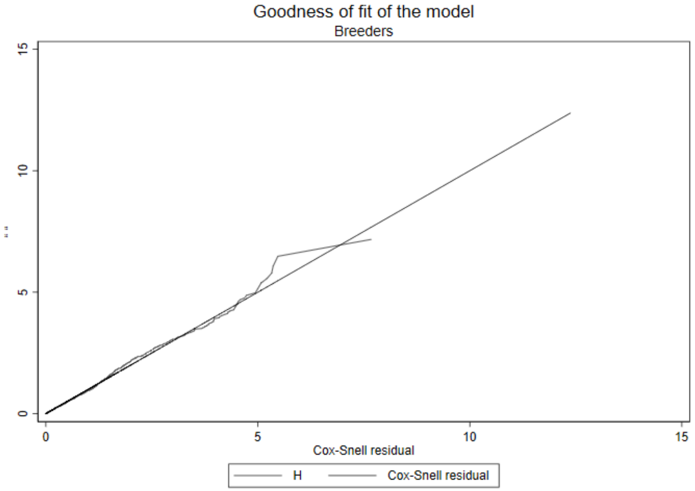


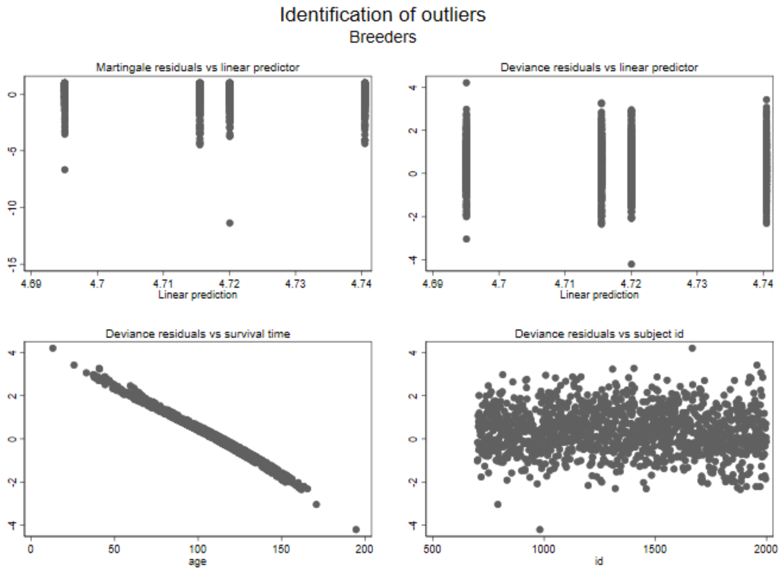


```
Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      0b.sex      |           .             .        1             .
      1.sex       |   -0.05423           3.84        1         0.0499
      0b.treat    |           .             .        1             .
      1.treat     |   -0.02560           0.85        1         0.3570
      ------------+---------------------------------------------------
      global test |                      4.78        2         0.0915
      ----------------------------------------------------------------


      Time:  Log(t)
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      0b.sex      |           .             .        1             .
      1.sex       |   -0.05282           3.65        1         0.0562
      0b.treat    |           .             .        1             .
      1.treat     |   -0.02284           0.68        1         0.4113
      ------------+---------------------------------------------------
      global test |                      4.40        2         0.1107
      ----------------------------------------------------------------
```

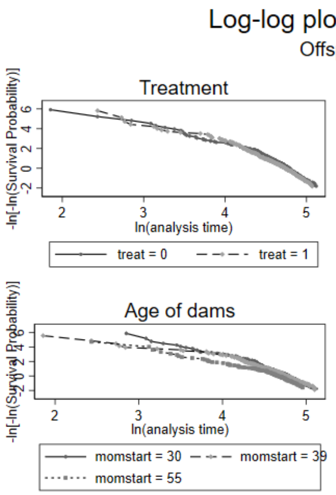
```
Time:  Kaplan-Meier
----------------------------------------------------------------
            |       rho            chi2       df       Prob>chi2
------------+---------------------------------------------------
0b.sex      |            .            .        1              .
1.sex       |     -0.05994         4.70        1         0.0302
0b.treat    |            .            .        1              .
1.treat     |     -0.03101         1.24        1         0.2645
------------+---------------------------------------------------
global test |                      6.06        2         0.0483
----------------------------------------------------------------
```

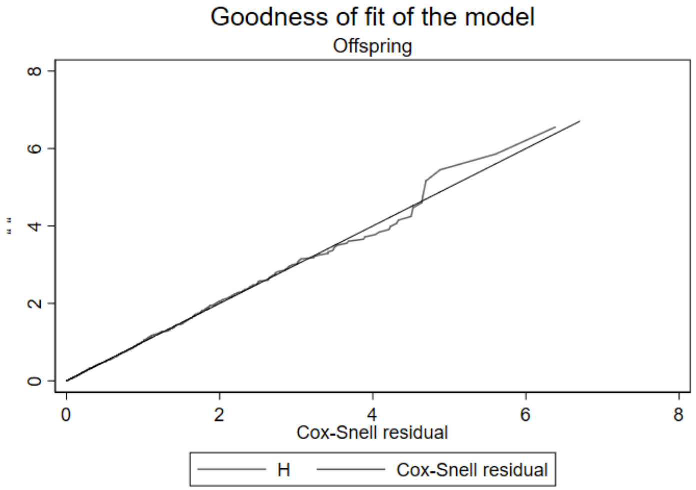
1.2 Accelerated failure-time model, generalized Gamma distribution, breeder

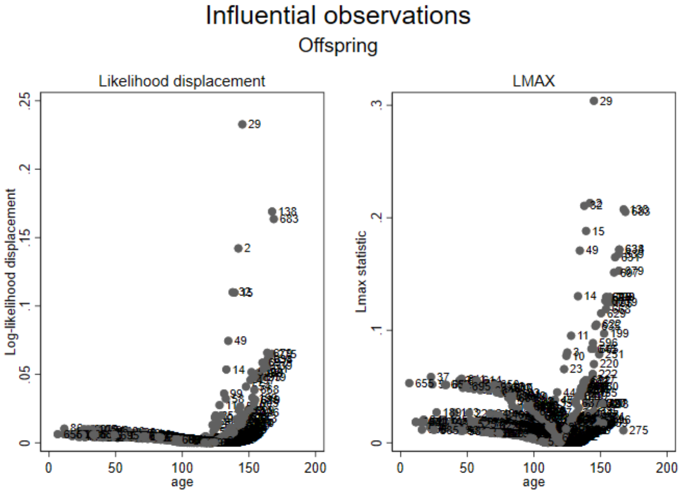


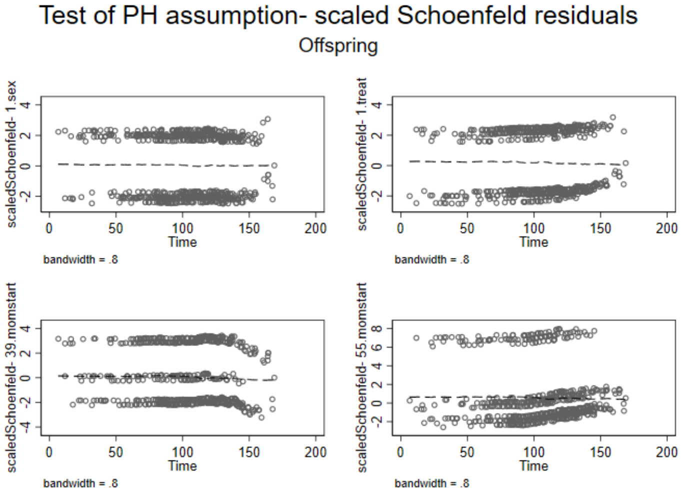


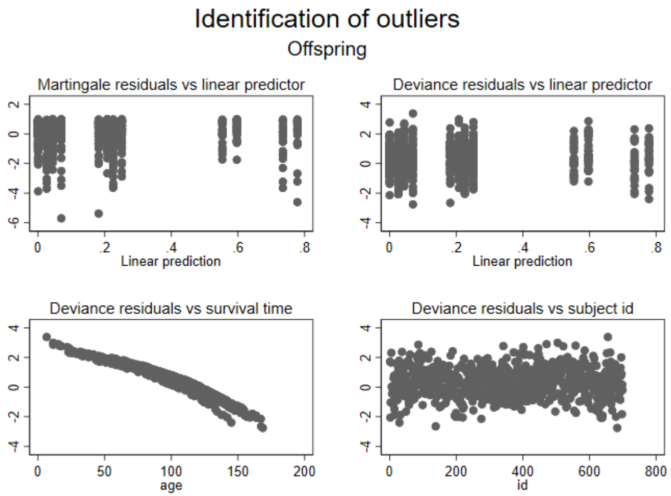
1.3 Cox Proportional Hazard model, offspring







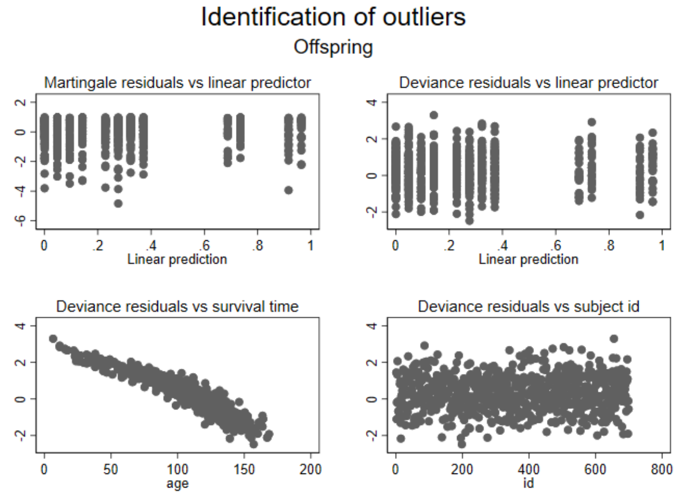


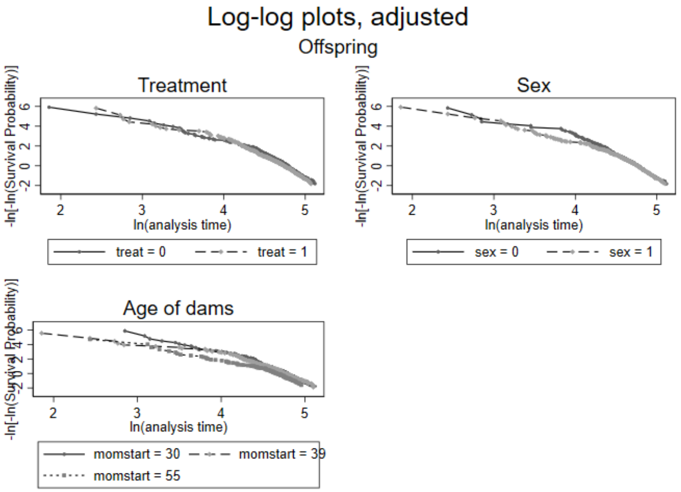


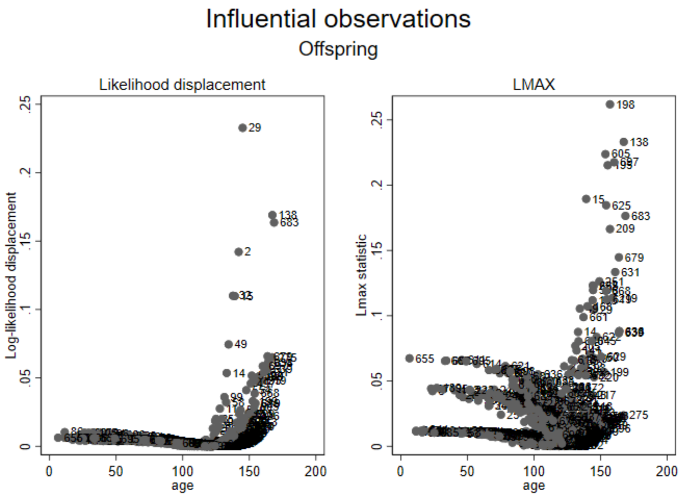
```
Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      0b.sex      |            .            .        1             .
      1.sex       |     -0.05304         2.01        1         0.1568
      0b.treat    |            .            .        1             .
      1.treat     |     -0.00458         0.01        1         0.9029
      30b.momstart|            .            .        1             .
      39.momstart |     -0.07824         4.30        1         0.0382
      55.momstart |     -0.07424         3.86        1         0.0494
      ------------+---------------------------------------------------
      global test |                      7.81        4         0.0986
      ----------------------------------------------------------------


      Time:  Kaplan-Meier
      ----------------------------------------------------------------
                  |       rho            chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      0b.sex      |            .            .        1             .
      1.sex       |     -0.03200         0.73        1         0.3928
      0b.treat    |            .            .        1             .
      1.treat     |     -0.02653         0.50        1         0.4796
      30b.momstart|            .            .        1             .
      39.momstart |     -0.07956         4.44        1         0.0350
      55.momstart |     -0.05326         1.99        1         0.1585
      ------------+---------------------------------------------------
      global test |                      6.25        4         0.1815
      ----------------------------------------------------------------
```

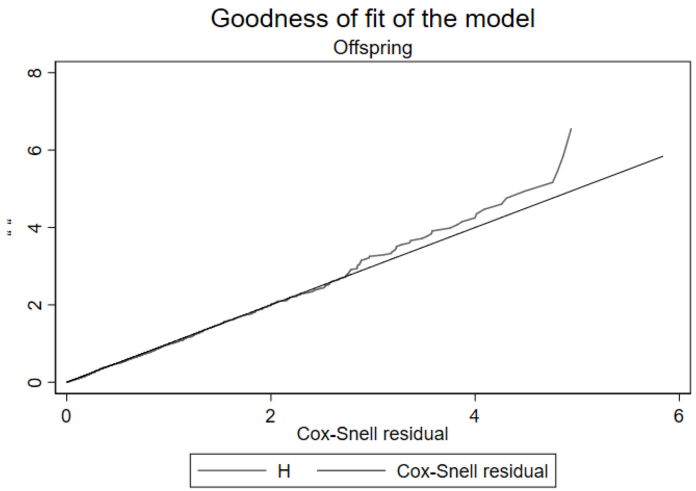
```
Time:  Rank(t)
----------------------------------------------------------------
             |       rho            chi2       df       Prob>chi2
-------------+--------------------------------------------------
0b.sex       |            .            .        1             .
1.sex        |     -0.03189         0.72        1        0.3945
0b.treat     |            .            .        1             .
1.treat      |     -0.02669         0.51        1        0.4769
30b.momstart|            .            .        1             .
39.momstart  |     -0.07962         4.45        1        0.0349
55.momstart  |     -0.05328         1.99        1        0.1584
-------------+--------------------------------------------------
global test  |                      6.25        4        0.1809
----------------------------------------------------------------
```

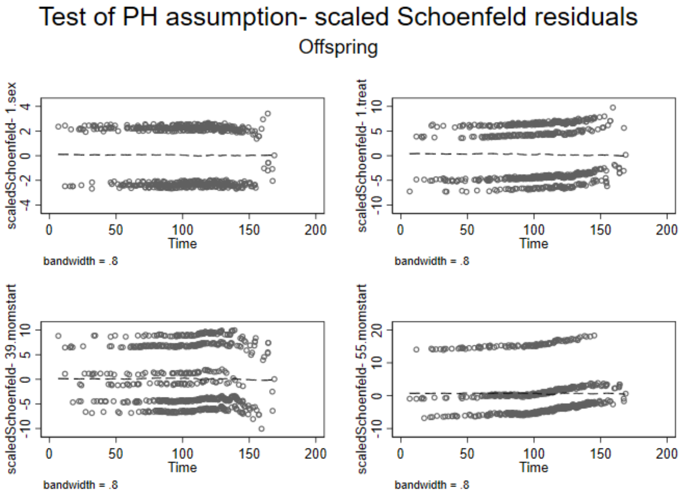
1.4 Cox Proportional Hazard model with shared Gamma frailty, offspring






Influential observations
Offspring
Likelihood displacement
LMAX
Log-likelihood displacement
Lmax statistic
age
0
50
100
150
200
0
.05
.1
.15
.2
.25
29
138
683
2
33
49
14
198
138
605
15
625
683
209
679
631
661
655
275


Goodness of fit of the model
Offspring
0
2
4
6
8
Cox-Snell residual
H
Cox-Snell residual


Test of PH assumption- scaled Schoenfeld residuals
Offspring
scaledSchoenfeld- 1.sex
scaledSchoenfeld- 1.treat
scaledSchoenfeld- 39.momstart
scaledSchoenfeld- 55.momstart
Time
0
50
100
150
200
bandwidth = .8

```
Test of proportional-hazards assumption

   Time:  Time
   ----------------------------------------------------------------
               |       rho            chi2       df       Prob>chi2
   ------------+---------------------------------------------------
   0b.sex      |            .            .        1             .
   1.sex       |     -0.04712         1.78        1         0.1826
   0b.treat    |            .            .        1             .
   1.treat     |      0.00100         0.00        1         0.9659
   30b.momstart|            .            .        1             .
   39.momstart |     -0.03689         2.48        1         0.1151
   55.momstart |     -0.04387         3.17        1         0.0752
   ------------+---------------------------------------------------
   global test |                      5.85        4         0.2108
   ----------------------------------------------------------------


   Time:  Kaplan-Meier
   ----------------------------------------------------------------
               |       rho            chi2       df       Prob>chi2
   ------------+---------------------------------------------------
   0b.sex      |            .            .        1             .
   1.sex       |     -0.02686         0.58        1         0.4474
   0b.treat    |            .            .        1             .
   1.treat     |     -0.01614         0.48        1         0.4887
   30b.momstart|            .            .        1             .
   39.momstart |     -0.03295         1.98        1         0.1593
   55.momstart |     -0.02387         0.94        1         0.3329
   ------------+---------------------------------------------------
   global test |                      3.64        4         0.4572
   ----------------------------------------------------------------


   Time:  Rank(t)
   ----------------------------------------------------------------
               |       rho            chi2       df       Prob>chi2
   ------------+---------------------------------------------------
   0b.sex      |            .            .        1             .
   1.sex       |     -0.02673         0.57        1         0.4495
   0b.treat    |            .            .        1             .
   1.treat     |     -0.01628         0.49        1         0.4850
   30b.momstart|            .            .        1             .
   39.momstart |     -0.03300         1.99        1         0.1587
   55.momstart |     -0.02392         0.94        1         0.3320
   ------------+---------------------------------------------------
   global test |                      3.65        4         0.4551
   ----------------------------------------------------------------
```