

SKE protokol

Aleksej Gaj*

15/08/2023

Obsah

1	Zadání	3
2	Dataset	4
3	Parametrické a neparametrické modely	6
3.1	Parametrické modely	6
3.1.1	Pacienti léčení lékem	6
3.1.2	Pacienti léčení placebem	9
3.2	Neparametrické modely	10
3.2.1	Pacienti léčení lékem	10
3.2.2	Pacienti léčení placebem	11
4	Porovnání podskupin lék vs. placebo	11
5	Coxův regresní model	12
5.1	Ověření předpokladů	12
5.2	Model	12
6	Závěr	13

*email: aleksejalex@gmail.com

Postulate 1 (Jen pro mně: censored vs died). In the context of survival analysis, censored data refers to observations that are incomplete or not fully observed. Censoring occurs when the exact event time (such as death) for a subject is unknown or has not yet occurred at the end of the study or the time of analysis.

In your dataset with patients, if the event of interest is whether they have died or not, the "censored" status indicates that the patient's event status is unknown because they were still alive or their follow-up time ended before the event occurred. It means that the patient was observed for a certain period but the event (death) did not occur within that period. Censoring can happen due to various reasons, such as the end of the study, loss to follow-up, or the patient being still alive at the time of analysis.

In survival analysis, censored data is an important consideration. Statistical methods, like the Cox proportional hazards model or Kaplan-Meier estimator, handle censored data appropriately and use the available information to estimate survival probabilities and hazard rates.

When analyzing your dataset, you would need to take into account both the "died" events and the censored observations to obtain valid estimates of survival probabilities or hazard ratios. The censored observations contribute information about the time the patient was followed until censoring, which is valuable in estimating the survival function beyond the observed events.

It's crucial to appropriately handle and interpret censored data in survival analysis to obtain reliable conclusions about the survival experience of patients in your study.

1 Zadání

A) Pomocí parametrických a neparametrických metod pro cenzorovaná data odhadněte vhodný spolehlivostní model pro časy dožití (survif T_j) obou vybraných podskupin pacientů. Pro kontrolu fitu parametrické rodiny užijte Kaplan-Meierův plot nebo Nelson-Aalenův ‘hazard plot’ (nejlépe v jednom obrázku spolu s parametrickým průběhem), resp. QQ/PP při RC.

B) SROVNĚJTE tyto vybrané podskupiny vzhledem k jejich

- průběhu spolehlivosti (survival function) $R(t)$, resp.
- intenzitě poruch (survivals) $\lambda(t)$ (IFR/DFR/CFR), resp.
- kumulativní intenzitě poruch (survivals) $\Lambda(t)$, resp.
- střední době života MTTF, resp.
- mediánové době života t_{med} , resp.
- ... (jiné vlastní, pokud vás něco osloví)

C) Graficky srovnejte log-logR ploty pro obě podskupiny a na jejich základě zdůvodněte vhodnost/nevhodnost užití Coxova PH (proportional hazard) modelu.

Skupina II.: treat=1(standard) versus treat=2(placebo) pro cell=2(small)

2 Dataset

Poskytnutý dataset představuje záznam testování vlivu jistého léčiva na dobu přežití pacienta. Data se skládají ze 137 pozorování 8 proměnných, viz Tabulka 1. Cílem je modelovat dobu dožití (*survival time*), tedy **survt** je vysvětlovaná proměnná. Ta je censorována podle proměnné **cens**. Další proměnné, které jsou k dispozici, představují věk pacienta, typ buněk, Karnofsky score (představující závažnost nemoci¹), trvání nemoci (proměnná **didur**) a zda pacient už absolvoval léčbu v minulosti.

V této práci se zaměříme na skupinu pacientů s typem buněk **cell=2**.

Název prom.	Komentář
treat	treatment (1 = standard/lék, 2 = test/placebo)
cell	cell type (1 = squamous, 2 = small, 3 = adeno, 4 = large)
survt	survival time (days)
cens	status (0 = censored, 1 = died)
KAR	performance status – Karnofsky score (0 = worst,..., 100 = best)
didur	disease duration from diagnosis to treatment (months)
age	age (years)
prith	prior therapy (0 = none, 10= some)

Tabulka 1: Popis proměnných v datasetu

V Tabulce 2 je uveden základní analýza na celém datasetu, v Tabulce 2 pak analýza podskupiny **cell=2**.

	treat	cell	survt	cens	KAR	didur	age	prith
count	137.00	137.00	137.00	137.00	137.00	137.00	137.00	137.00
mean	1.50	2.34	121.63	0.74	58.57	8.77	58.31	2.92
std	0.50	1.07	157.82	0.44	20.04	10.61	10.54	4.56
min	1.00	1.00	1.00	0.00	10.00	1.00	34.00	0.00
25%	1.00	1.00	25.00	0.00	40.00	3.00	51.00	0.00
50%	1.00	2.00	80.00	1.00	60.00	5.00	62.00	0.00
75%	2.00	3.00	144.00	1.00	75.00	11.00	66.00	10.00
max	2.00	4.00	999.00	1.00	99.00	87.00	81.00	10.00

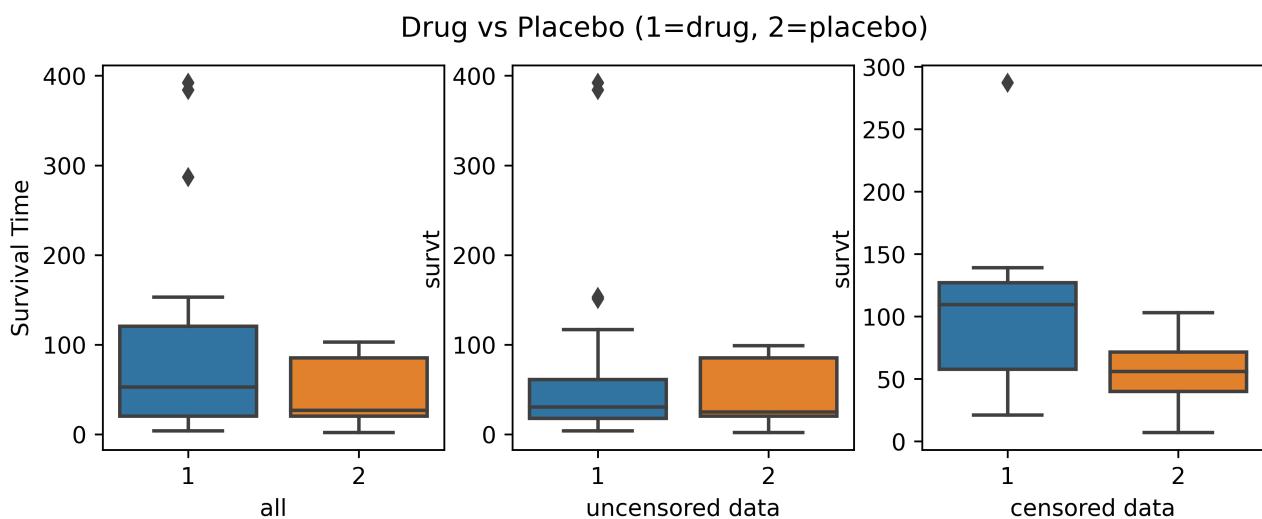
Tabulka 2: Základní analýza na celém datasetu

¹Pozn: Hodnoty Karnofského skóre znamenají: $KAR \leq 30$ – úplná hospitalizace, $30 < KAR \leq 60$ – částečná hospitalizace, $KAR > 60$ – vlastní péče bez hospitalizace.

	treat	cell	survt	cens	KAR	didur	age	prith
count	48.00	48.00	48.00	48.00	48.00	48.00	48.00	48.00
mean	1.38	2.00	71.67	0.75	53.54	9.25	59.88	2.29
std	0.49	0.00	85.77	0.44	19.10	13.91	9.92	4.25
min	1.00	2.00	2.00	0.00	20.00	1.00	35.00	0.00
25%	1.00	2.00	20.00	0.75	40.00	2.00	54.75	0.00
50%	1.00	2.00	51.00	1.00	60.00	4.00	62.50	0.00
75%	2.00	2.00	97.50	1.00	70.00	11.00	67.00	0.00
max	2.00	2.00	392.00	1.00	85.00	87.00	72.00	10.00

Tabulka 3: Základní analýza na podvybraném datasetu (**cell = 2**)

Podvybraný dataset obsahuje pouze 48 pozorování, jedná se o pacienty průměrně o něco starší a s výrazně kratší průměrnou dobou dožití. Dataset je rozdělen nerovnoměrně: 30 pacientů z 48 bylo léčeno skutečným lékem, a jen 18 z 48 placeboem.



Obrázek 1: Boxplot doby dožití (nalevo - celá analyzovaná skupina, uprostřed - pouze pro necensorovaná data, napravo - pouze pro censorovaná data)

Na Obrázku 1 jsou tři boxploty, ilustrující rozdělení pacientů léčených lékem a placebem mezi censorovanými² a necensorovanými případmi. Na všech třech boxplotech je vidět, že střední hodnota doby přežití je u pacientů léčených lékem vyšší, než u pacientů léčených placebem. Dokonce mezi skutečně léčenými pacienty je několik outlierů s výrazně nadprůměrnou dobou dožití.

Mezi skutečně léčenými pacienty je několik outlierů s výrazně nadprůměrnou dobou dožití. Může se jednat o případy, pro které testovaný lék je obzvláště účinný nebo pouze o přirozenou výjimku. Skutečnost, že taková odlehlá pozorování jsou pouze v případě léku a pouze na jednu stranu od průměru, napovídá k první variantě.

²V kontextu spolehlivostní analýzy censorované pozorování je neúplné nebo ne zcela pozorované měření. V tomto případě censorovaná pozorování jsou například pacienti, kteří přežili po celou dobu experimentu (a tedy okamžik úmrtí nebyl zaznamenán) nebo ukončili léčbu předčasně. Tedy necensorovaná pozorování představují pacienty, jejichž úmrtí bylo pozorováno před koncem experimentu.

Na Obrázku 2 je základní vizualizace datasetu (barevně jsou odděleny podskupiny, tj. lék a placebo).

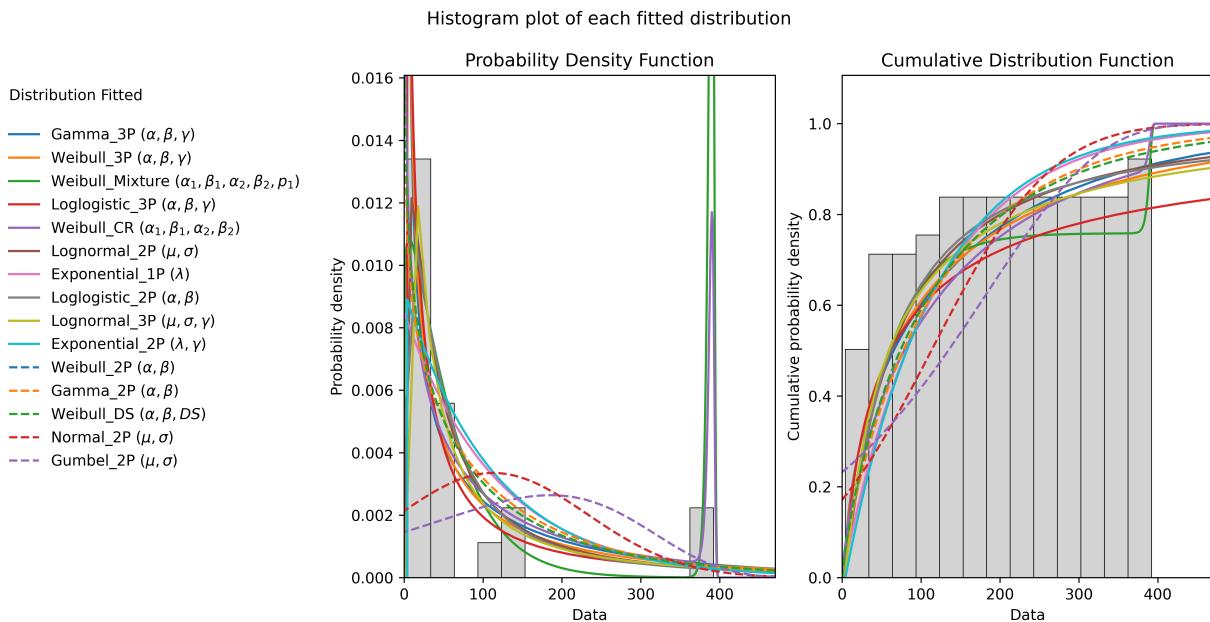
3 Parametrické a neparametrické modely

Nyní se pokusíme najít nejlepší modely pro popis dat, tedy vytvoříme pro každou skupinu pacientů vlastní model. Nejprve vyzkoušíme pro každou podskupinu parametrický přístup, poté neparametrický a následně výsledné modely porovnáme.

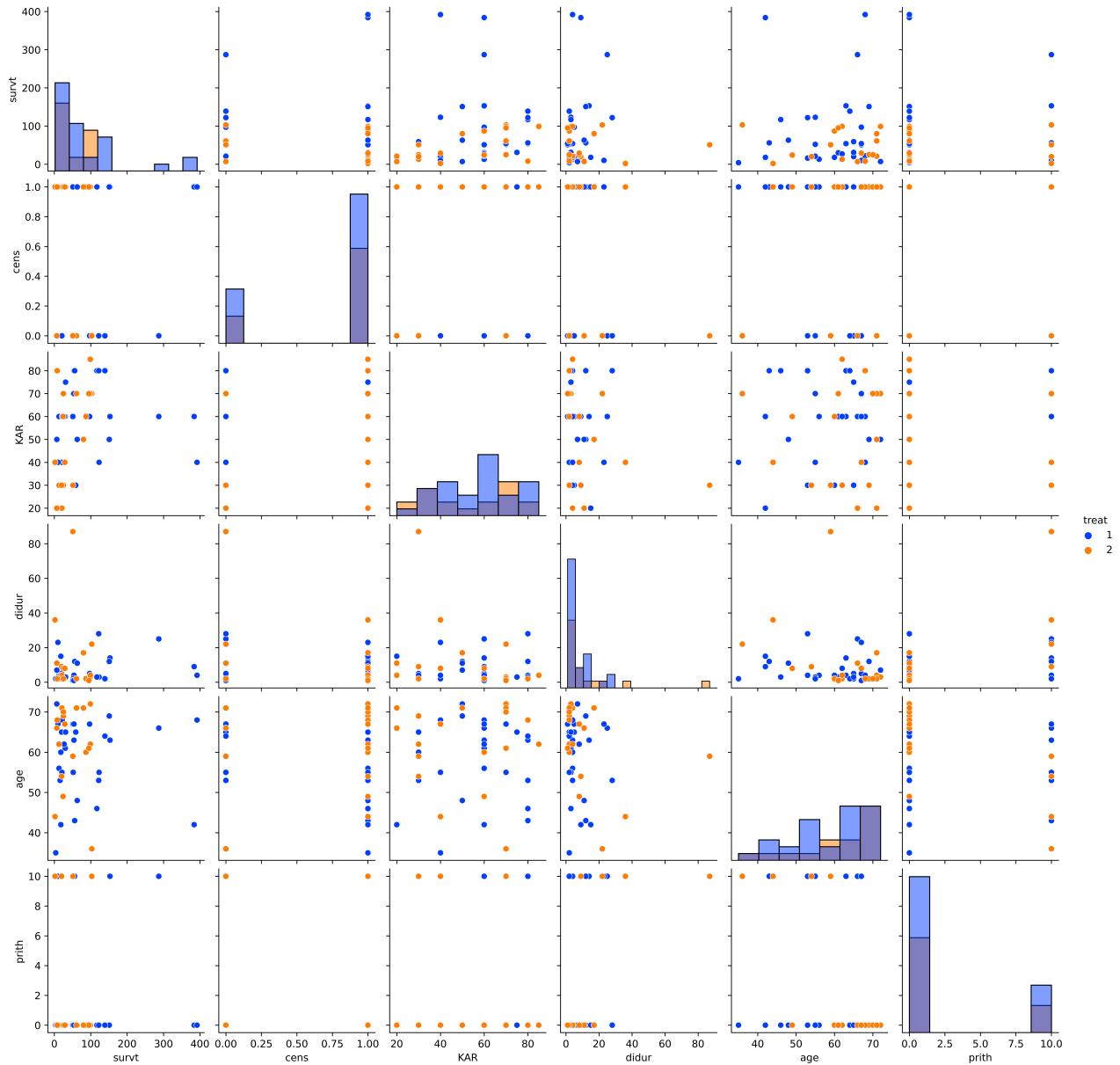
3.1 Parametrické modely

Ke hledání vhodného parametrického modelu použijeme existující implementaci v knihovně `reliability`. V Podsekčích 3.1.1 a 3.1.2 jsou pokaždé uvedeny grafické výstupy funkce `Fit_Everything` a stručná diskuse, vysvětlující volbu konkrétní hustoty. Funkce `Fit_Everything` proloží poskytnutá data pomocí 15 předdefinovaných hustot a pro každou odhadne hodnoty parametrů. Následně vykreslí PP ploty a probability ploty pro každý fit.

3.1.1 Pacienti léčení lékem



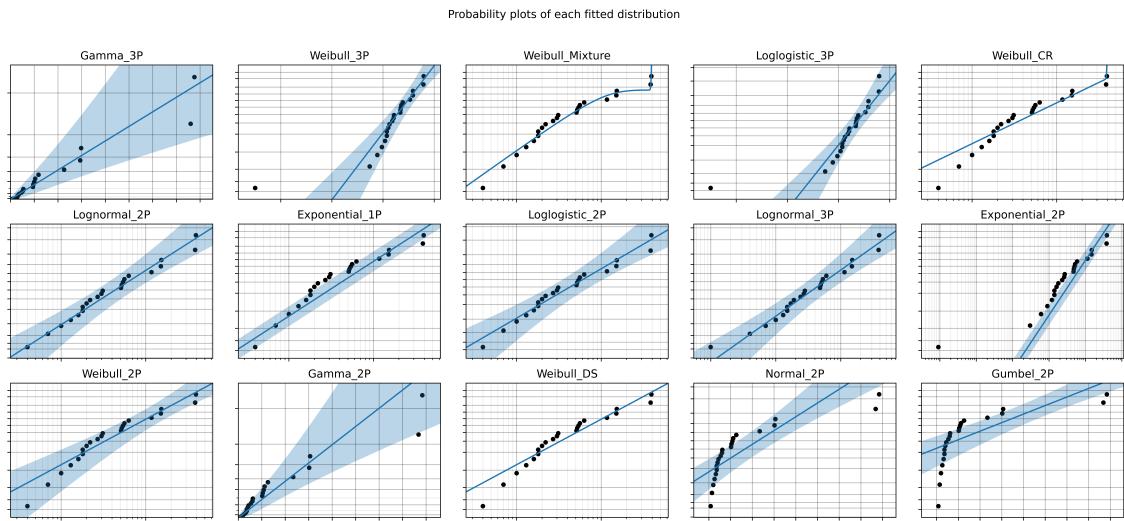
Obrázek 3: fit all lék



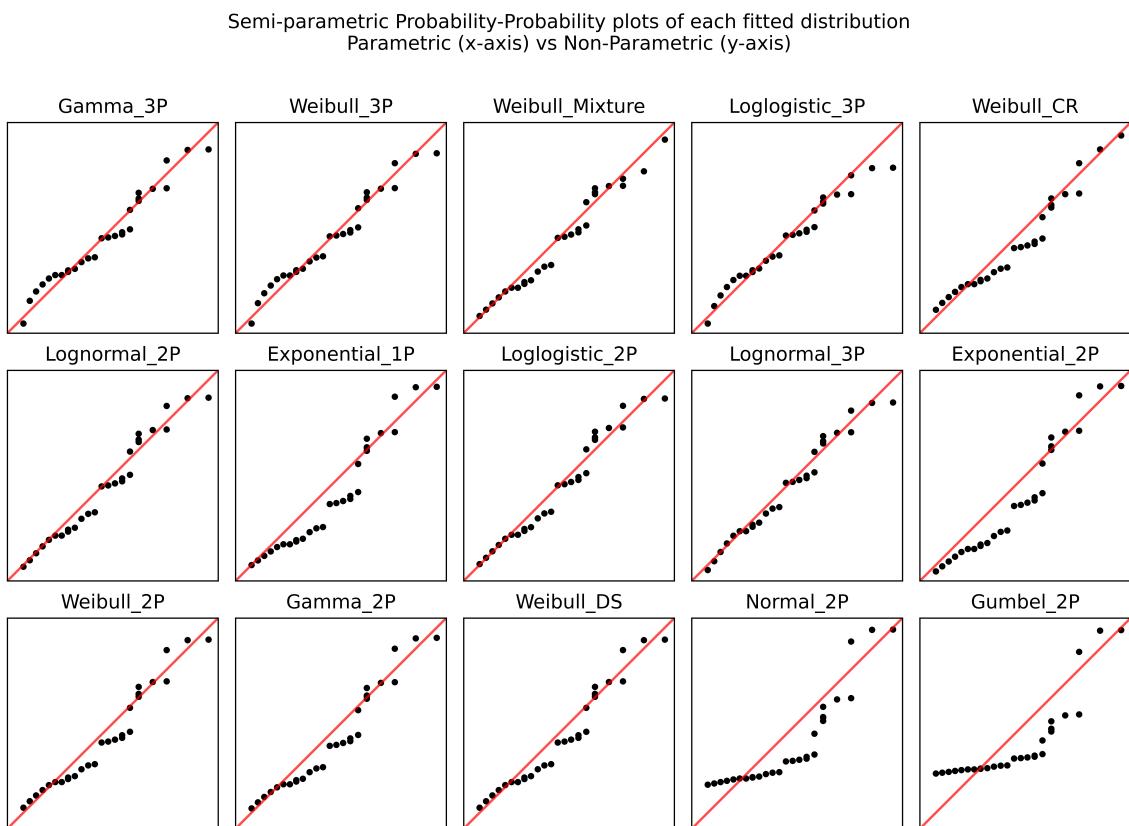
Obrázek 2: Pairplot

3 PARAMETRICKÉ A NEPARAMETRICKÉ MODELY

8



Obrázek 4: P ploty pro jednotlivé fity



Obrázek 5: PP ploty pro jednotlivé fity

V Tabulce 4 jsou shrnuty hodnoty odhadnutých parametrů a také hodnoty log-likelihood, AIC a BIC.

Distribution	α	β	γ	α_1	β_1	α_2	β_2	Prop. 1	DS	μ	σ	λ	Log-like.	AIC	BIC	
Gamma3P	270.72	0.50	4.00										-120.39	247.71	250.99	
Weibull3P	106.13	0.61	4.00										-120.93	248.79	252.07	
WeibullMixt.				56.57	1.10	389.96	116.37	0.76					-118.04	248.57	253.08	
Loglogistic3P	57.80	0.78	4.00										-122.66	252.25	255.53	
WeibullCR				128.49	0.74	390.17	120.53						-121.13	251.86	255.86	
Lognormal2P									4.13	1.39			-124.59	253.62	255.98	
Exponential1P											0.01		-126.91	255.97	257.23	
Loglogistic2P	60.37	1.19											-125.30	255.04	257.39	
Lognormal3P			3.00						4.02	1.63			-123.98	254.87	258.15	
Exponential2P			4.00								0.01		-125.87	256.18	258.54	
Weibull2P	114.41	0.83											-126.20	256.85	259.21	
Gamma2P	153.69	0.80											-126.49	257.43	259.79	
WeibullIDS	114.41	0.83							1.00				-126.20	259.33	262.61	
Normal2P										112.97	119.09			-141.89	288.23	290.58
Gumbel2P										185.54	139.42			-147.79	300.02	302.38

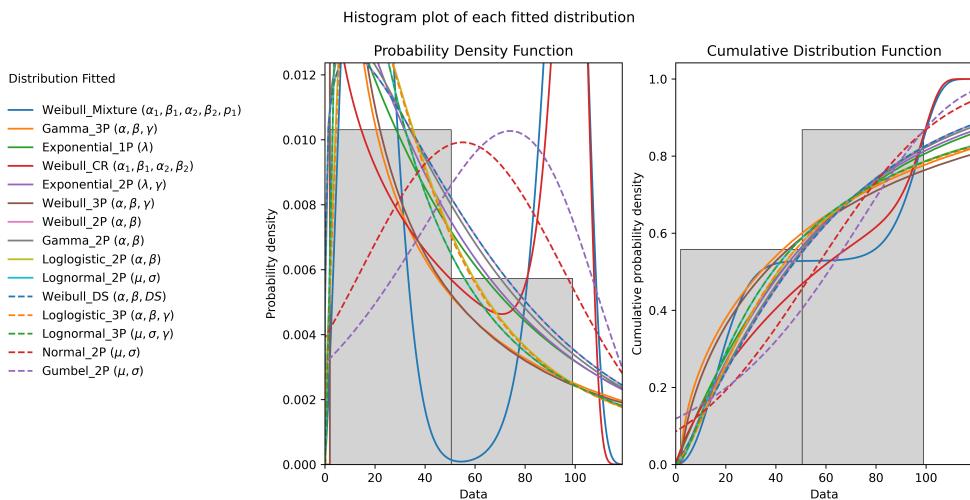
Tabulka 4: Odhadování parametrů pro podskupinu pacientů léčených lékem

Na základě vizuálního vyhodnocení P-plotů a PP-plotů (Obr. 4 a Obr. 5)

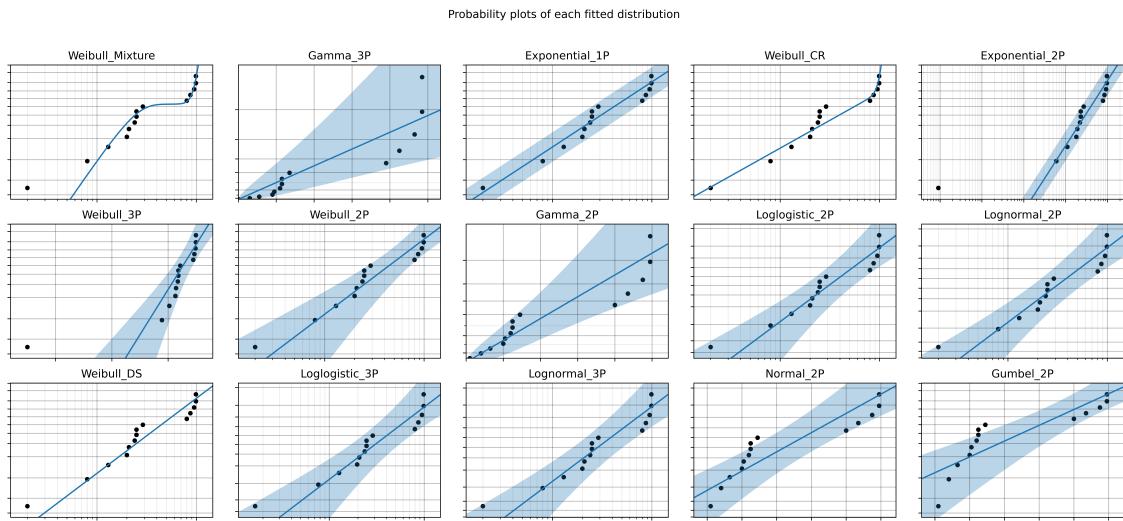
Na Obrázku 8 je porovnání **ceho?** pro

Je vidět, že oba modely (logNormal 2P a Gamma 3P) popisují data podobně dobře, a to i ve srovnání s neparametrickými metodami Kaplan-Meiera a Nelson-Aalena (viz Podsekce ??), a tedy zvolíme jednodušší model – dvouparametrickou lognormální hustotu.

3.1.2 Pacienti léčení placebo



Obrázek 6: Fit all Placebo

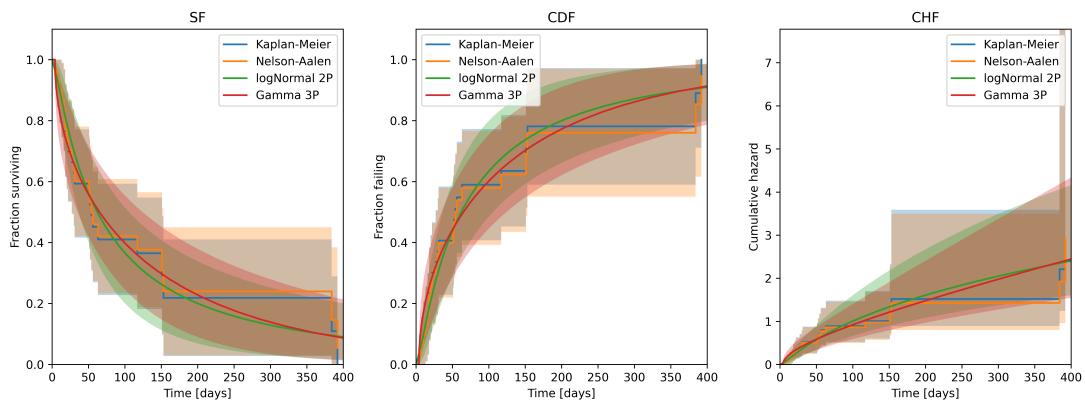


Obrázek 7: PP ploty pro jednotlivé fity

3.2 Neparametrické modely

V Podsekčích 3.2.1 a 3.2.2 jsou pro obě skupiny pacientů (tj. léčení lékem a léčení placebem) uvedeny odhadové Kaplan-Meiera a Nelson-Aalena. Na grafech jsou pro porovnání vynesené také vybrané parametrické odhadové funkce (viz výše).

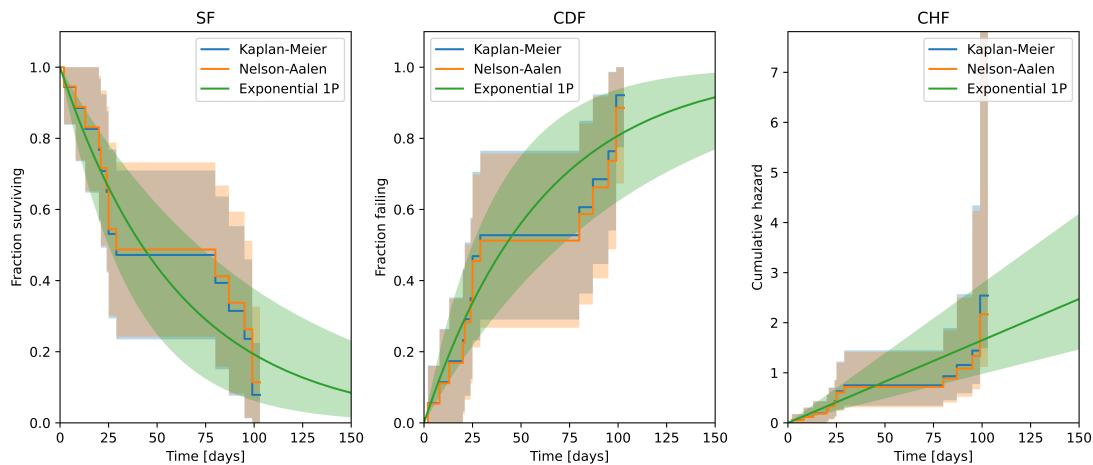
3.2.1 Pacienti léčení lékem



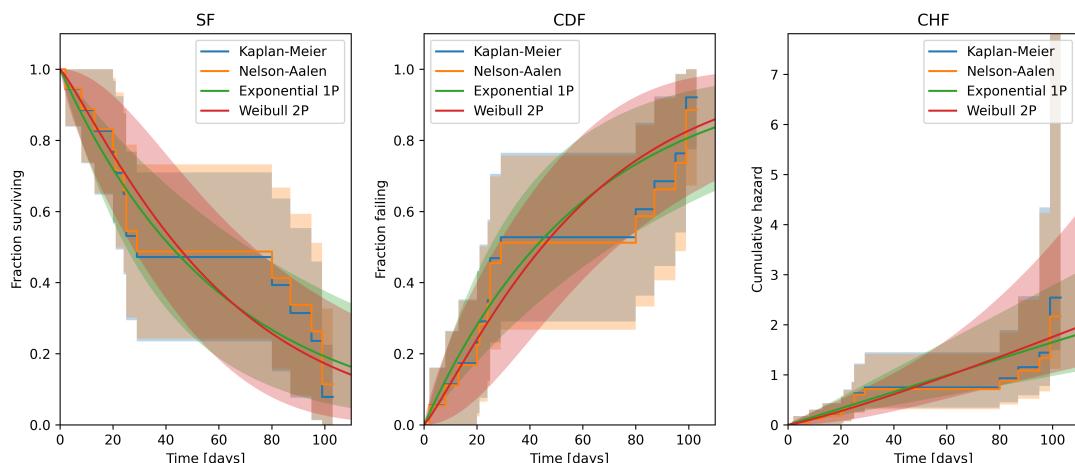
Obrázek 8: Four men drugs

Na ose x je vynesen počet dnů do hodnoty 400, jelikož maximální hodnota `survt` pro pacienty léčené lékem je 392.

3.2.2 Pacienti léčení placeboem



Obrázek 9: Four men placebo



Obrázek 10: Four men placebo compared to parametrics

Zde je počet dnů vynesen pouze do hodnoty 110, jelikož maximální hodnota `survt` pro pacienty léčené placeboem je 103.

4 Porovnání podskupin lék vs. placebo

aaaaaa

V této sekci uvedem porovnání modelů, sestavených v předchozích sekcích. Konkrétně tedy budeme porovnávat

- a
- b

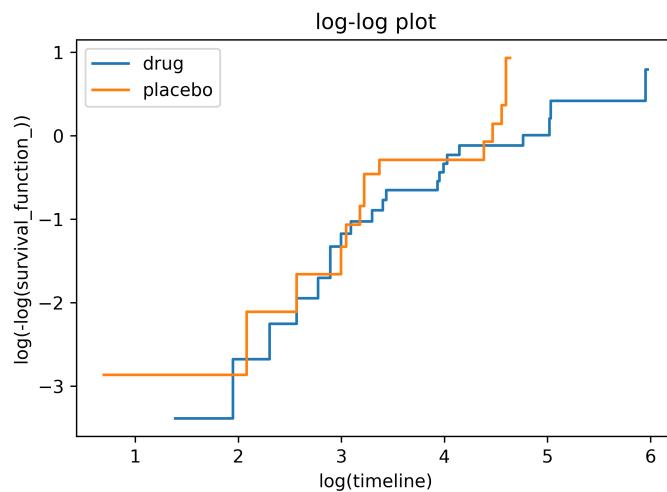
5 Coxův regresní model

Coxův regresní model je založen na Coxově proporcionalním hazardovém předpokladu, který tvrdí, že poměrné riziko dvou skupin je konstantní v čase. Tento předpoklad umožnuje odhadnout vliv různých faktorů na přežití a současně zachovává nezávislost na neovlivňujících proměnných.

Pro použití Coxova regresního modelu je nejprve potřeba mít k dispozici data o čase do události (např. úmrtí) a příslušné prediktory (faktory ovlivňující přežití). Zde byla použita existující implementace Coxova regresního modelu v knihovně **lifelines**.

5.1 Ověření předpokladů

Začneme vizuální kontrolou průběhu log-log \hat{R}_{KM} (Obrázek 11).



Obrázek 11: log-log plot lék vs placebo

Je vidět, že grafy nejsou “rovnoběžné”, na několika místech se kříží. Znamená to, že v tomto případě Coxův model nelze použít.

6 Závěr

Tato práce