

SKE protokol

Aleksej Gaj*

11/08/2023

Obsah

1	Zadání	3
2	Dataset	4
2.1	Základní analýza	5
2.2	Vizualizace	5
3	Parametrické a neparametrické modely	6
3.1	Parametrické modely	6
3.1.1	Pacienti léčení lékem	6
3.1.2	Pacienti léčení placebem	7
3.2	Neparametrické modely	8
3.2.1	Pacienti léčení lékem	8
3.2.2	Pacienti léčení placebem	9
4	Porovnání podskupin lék vs. placebo	9
5	Coxův regresní model	10
5.1	Ověření předpokladů	10
5.2	Model	10
6	Závěr	11

*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

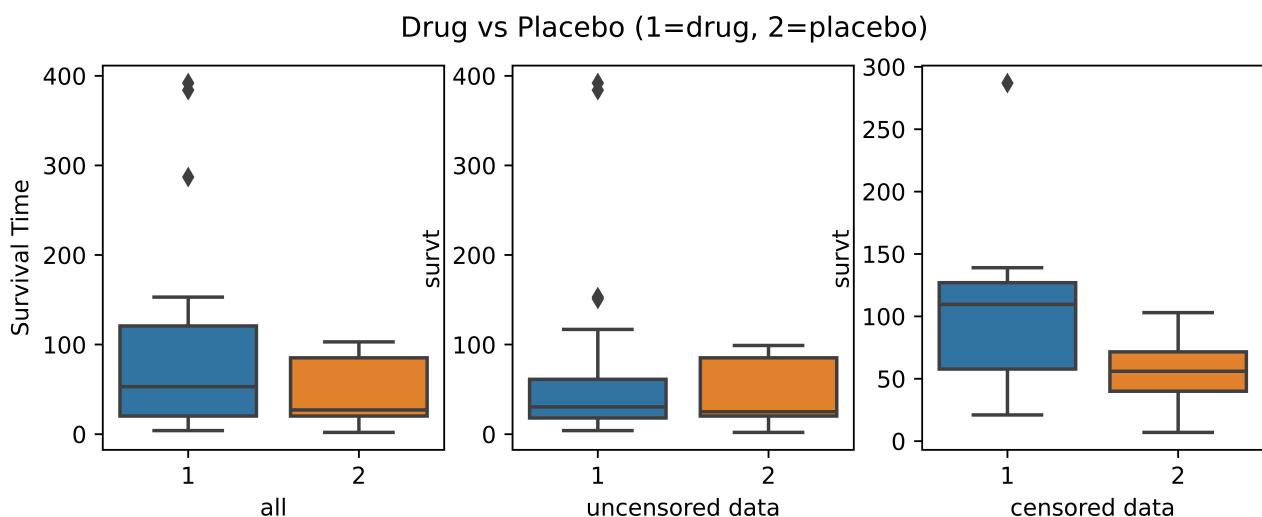
Tabulka 2: Základní analýza na celém datasetu

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

Na Obrázku 1 jsou tři boxploty, ilustrující rozdělení dat v j

2.1 Základní analýza

TODO: popis o co jde, tabulka describe celeho datasetu a mych dat Cílem je zjistit, jak dané léčivo ovlivnilo dobu dožití pacientů v porovnání s placeboem.

2.2 Vizualizace

TODO: histogram, boxplot

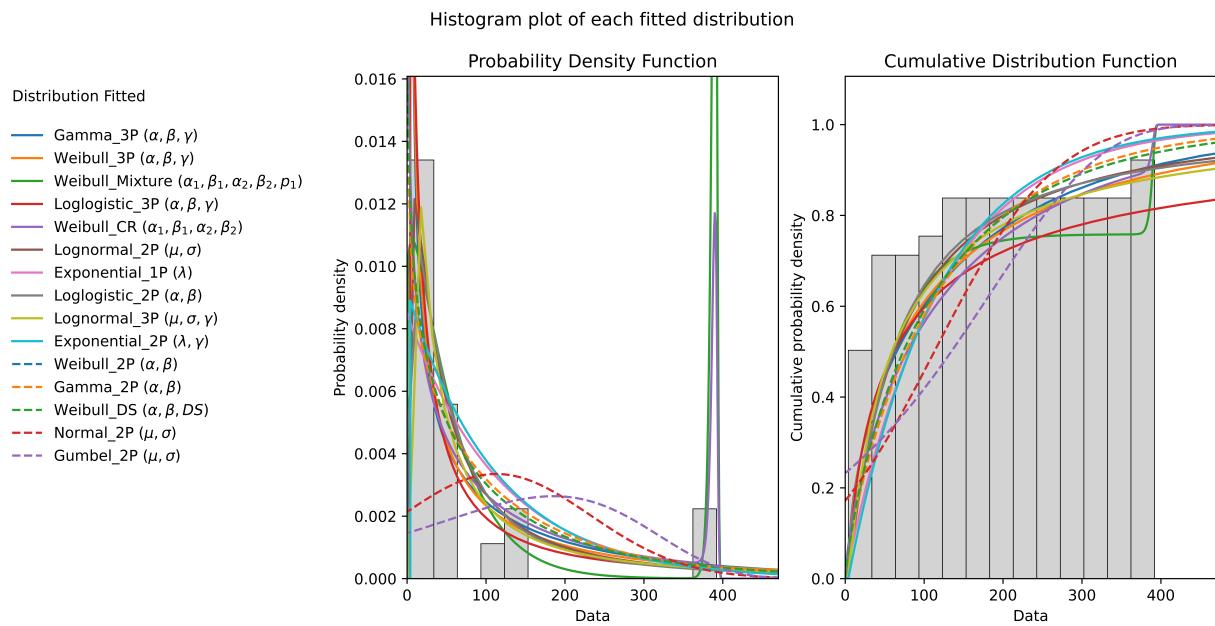
3 Parametrické a neparametrické modely

Nyní se pokusíme najít nejlepší modely pro popis dat, tedy vytvoříme pro každou skupinu pacientů vlastní model. Použijeme k tomu existující implementaci v knihovně **reliability**.

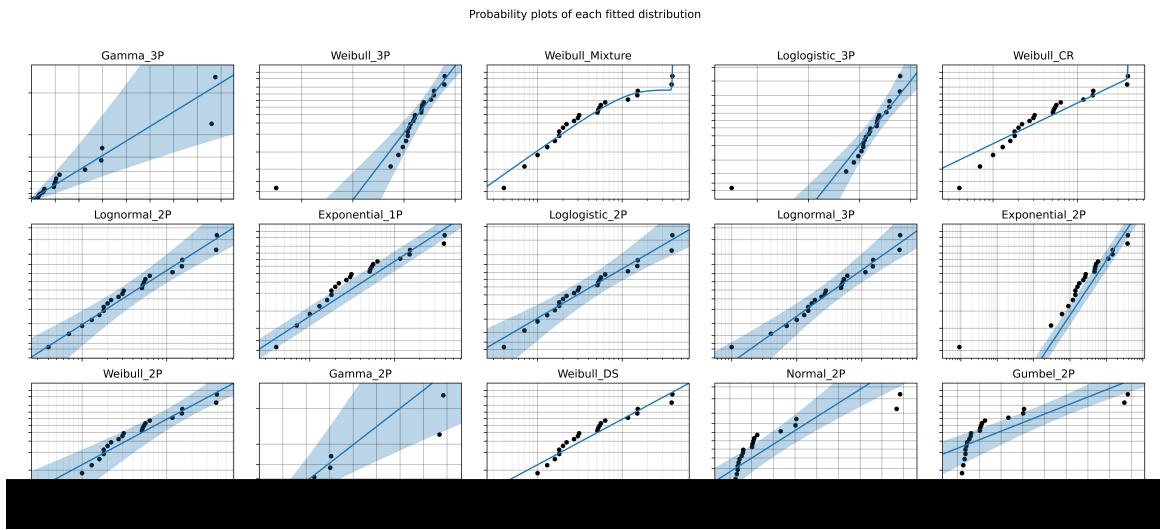
3.1 Parametrické modely

V Podsekčích 3.1.1, 3.1.2, 3.2.1, 3.2.2

3.1.1 Pacienti léčení lékem



Obrázek 2: fit all lék

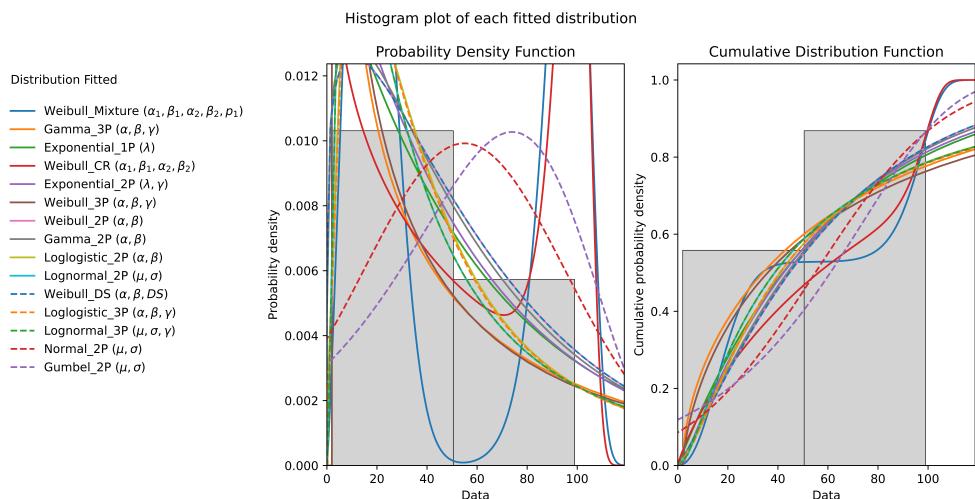


Obrázek 3: PP ploty pro jednotlivé fity

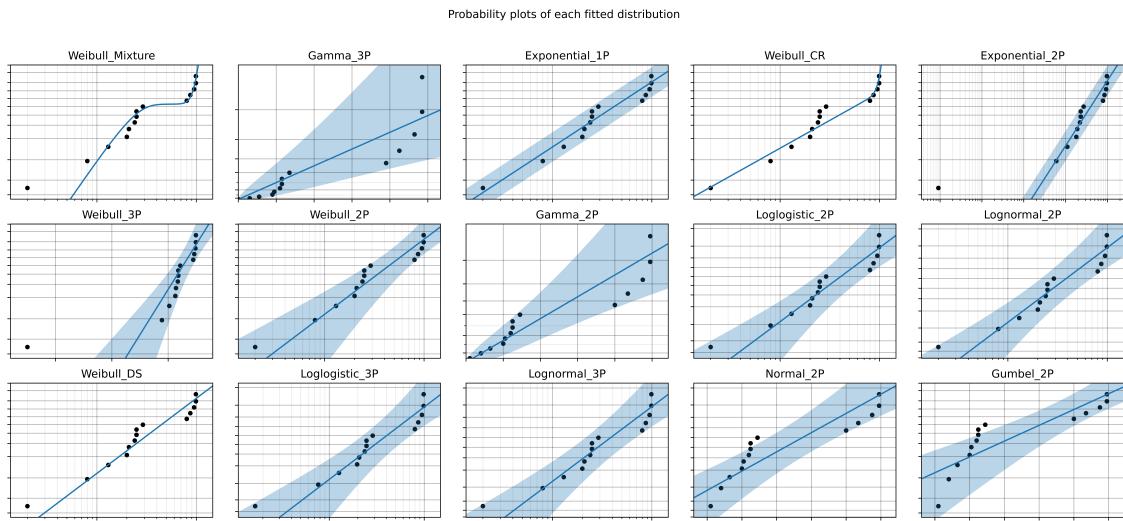
Na Obrázku 6 je porovnání **ceho?** pro

Je vidět, že oba modely (logNormal 2P a Gamma 3P) popisují data podobně dobře, pokud porovnávame tyto parametrické fity s neparametrickými metodami Kaplan-Meiera a Nelson-Aalena, a tedy zvolíme jednodušší model – dvouparametrickou lognormální hustotu.

3.1.2 Pacienti léčení placeboem



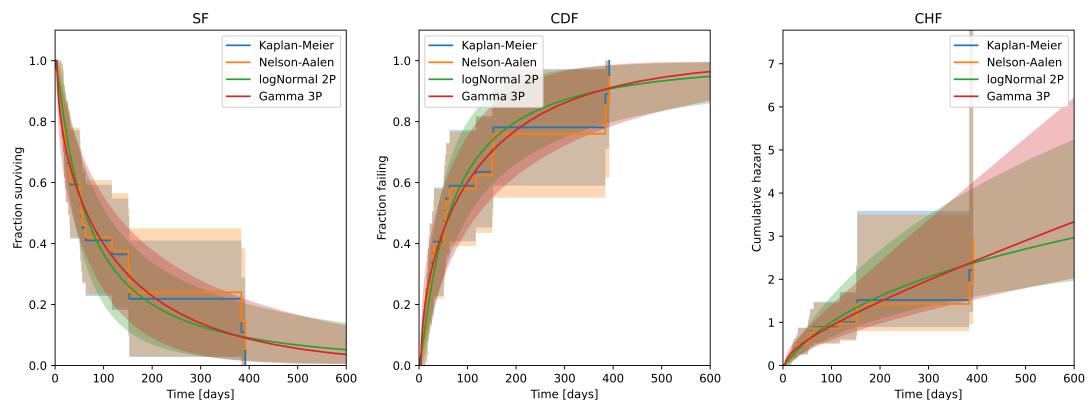
Obrázek 4: Fit all Placebo



Obrázek 5: PP ploty pro jednotlivé fity

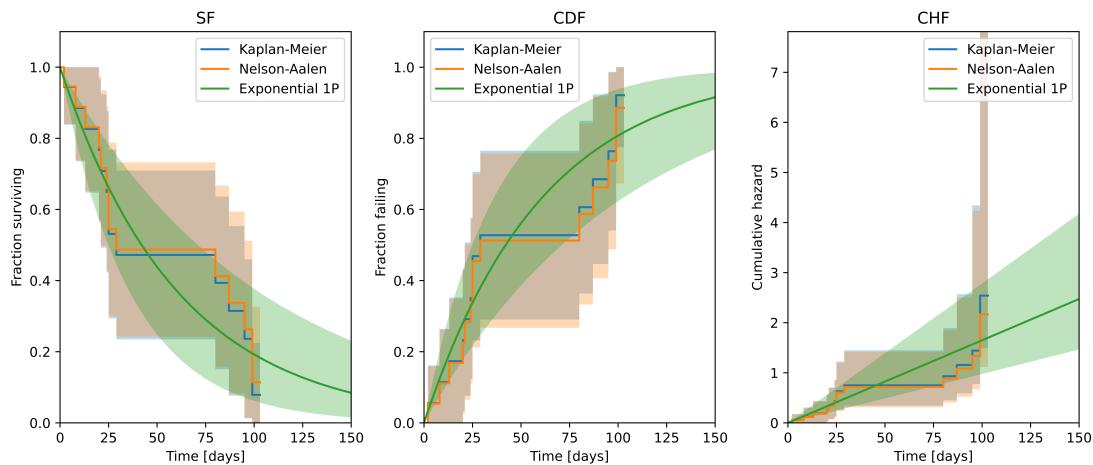
3.2 Neparametrické modely

3.2.1 Pacienti léčení lékem

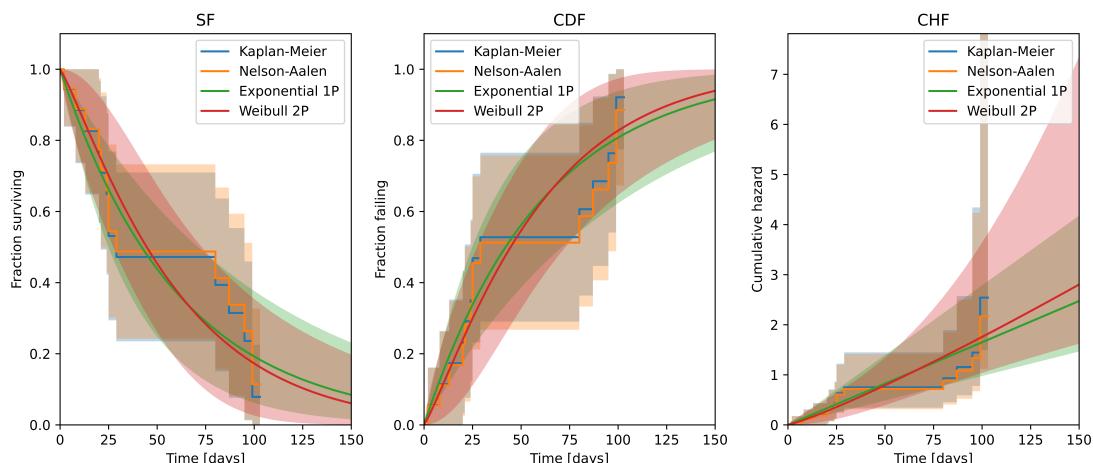


Obrázek 6: Four men drugs

3.2.2 Pacienti léčení placeboem



Obrázek 7: Four men placebo



Obrázek 8: Four men placebo compared to parametrics

4 Porovnání podskupin lék vs. placebo

aaaaaa

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

- a
- b

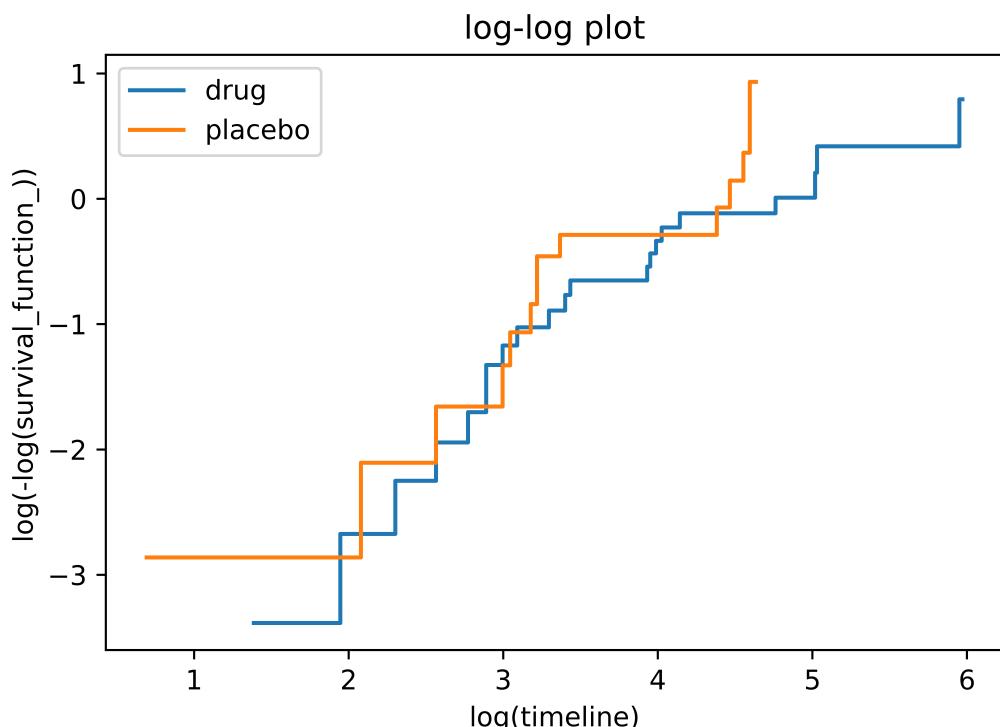
5 Coxův regresní model

Coxův regresní model je založen na Coxově proporcionalním hazardovém předpokladu, který říká, že poměrné riziko dvou jedinců je konstantní v čase. Tento předpoklad umožňuje 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

- text na logrank test
- pic



Obrázek 9: log-log plot lék vs placebo

5.2 Model

6 Závěr

tba tba