POZVÁNKA NA ROBUST 2024
INVITATION TO ROBUST 2024

Redakce

ROBUST 2024
20+24=44

Vážené kolegyně, vážení kolegové,

dne 8. září 2024 bude v Bardějově slavnostně zahájena dvacátá třetí letní škola **JČMF ROBUST 2024**. Připomeňme, že série letních a zimních škol ROBUST je organizována skupinou pro výpočetní statistiku při **ČMS JČMF** obrok od roku 1980, takže se s řadou z Vás sejdeme po neplánovaných a neuvěřitelných **čtyřiceti čtyřech letech**. Letos mezi spolu-organizátory patří **Česká statistická společnost** a **Slovenská štatistická a demografická spoločnosť**.

ROBUST 2024 bude věnován, tak jako vždy, současným trendům statistiky, pravděpodobnosti, finanční matematiky, optimalizace a analýzy dat. Vedle toho bude velká pozornost věnována problematice výuky našich oborů v době Internetu a existence volného přístupu k nástrojům typu Wikipedie, Google Search či Scholar, ChatGPT, systému R, apod., svádějících k mylné představě (nejenom studentů), že si řešení snadno a bezbolestně „vygúglíme“, spočteme či jej někdo „vyřeší za nás“.

Nabídku k přednesení zvaných přednášek přijali (abecedně):

- D. Kraus, MUNI, Brno
- Další pozvaní bodou upřesněni co nejdříve.

Vedle zvaných přednášek se počítá se sděleními účastníků a přednáškově–posterovou sekcí doktorandů a studentů.

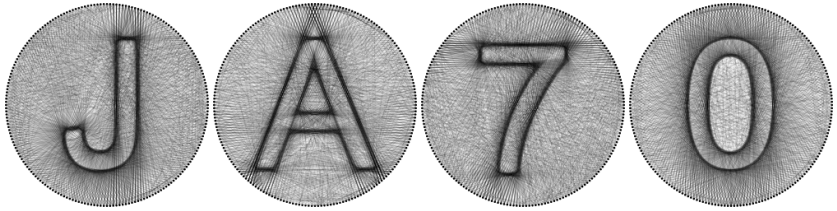
- *Datum a místo konání*: 8.–13. září 2024 (ne-pá), **Bardějov**; GPS $49°17'33.566''N, 21°16'16.726''E$.
- *Ubytování*: Ve dvou a tří lůžkových pokojích. Vzhledem k odlehlosti místa konání pro zájemce lze zajistit za příplatek ubytování již od soboty 7. září 2024. Za příplatek lze též zajistit individuální ubytování.
- *Stravování*: Celodenní (neděle večeře – pátek oběd).
- *Kulturní program*: Na středu plánujeme výlet po **dřevěných chrámech východoslovenských Karpat**, jež jsou zařazeny do světového dědictví UNESCO, a po stopách **karpatsko-dukelské operace**.
- *Technické vybavení*: K dispozici bude notebook, zpětný a datový projektor, tabule a wifi připojení.
- *Přihláška a registrace*: Registrace je otevřena na adrese `http://robust.nipax.cz`. Pro ty, kteří se zúčastnili předchozích Robustů, jsme zachovali již dříve vložené osobní údaje. Zkontrolujte je a v případě změn opravte. Při jakýchkoliv potížích s registrací se obraťte na kolegu Dohnala (`dohnal@nipax.cz`).
- *Abstrakt*: **Na svoji přednášku myslete již dnes**, a abstrakt v TEXu spolu s pdf souborem vložte přes registrační stránku na adrese `robust.nipax.cz` **nejpozději 8. května 2024**.
- *Stipendia*: Česká statistická společnost vypsala několik stipendií; podrobnosti viz `www.statspol.cz`.
- *Publikace*: V případě zájmu účastníků připravíme speciální číslo časopisu Statistika; podrobnosti viz Zprávy.
- *Konferenční poplatek*: 6 250 Kč, resp. 250 Euro. Pro studenty řádného studia a interní PGS studenty bez vedlejšího úvazku 5 500 Kč, resp. 220 Euro. Zahrnuto je ubytování, strava a náklady organizačního výboru.
- *Bankovní spojení*: CZ: Číslo účtu 2001215985/2010 u Fio banky, a.s. IBAN: CZ77 2010 0000 0020 0121 5985 BIC: FIOBCZPPXXX. Příjemce: Česká matematická společnost, sekce Jednoty Českých matematiků a fyziků, Žitná 609/25, CZ – 110 00 Praha 1. Variabilní symbol 20245xxx, kde xxx zvolte sami a vložte do Vaší registrační stránky!
  SK: Číslo účtu 4021244256/7500 u Československé obchodné banky, a.s. IBAN: SK33 7500 0000 0040 2124 4256 SWIFT CEKOSKBX. Příjemce: Slovenská štatistická a demografická spoločnosť, Miletičova 3, SK – 824 67 Bratislava. Variabilní symbol 2024xxx, kde xxx zvolte sami a vložte do Vaší registrační stránky!
- *Poznámky k placení*: Abychom předešli zbytečnému „stěhování“ konferenčních poplatků přes hranice tam a zpět, a snížili výdaje za bankovní směny peněz, doporučujeme účastníkům ze Slovenska zaplatit na účet SŠDS. Jako informaci pro příjemce platby uveďte Vaše jméno (instituci). Daňový doklad obdržíte během konference. Budete-li platit na místě, dohodněte se s námi předem. Pokud potřebujete zálohovou fakturu dříve, obraťte se na paní Naxerovou (platby z CZ) (`naxerova@karlin.mff.cuni.cz`) nebo Stankovičovou (platby z SK) (`iveta.stankovicova@gmail.com`).
- *Další oznámení*: Veškeré informace budou zveřejňovány na `www.karlin.mff.cuni.cz/~antoch`, a dle potřeb distribuovány buď e-mailem nebo klasickou poštou.
- *Adresa pro korespondenci*: ROBUST 2024, MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín, tel. 221 913 287; e-mail: `antoch@karlin.mff.cuni.cz`

V Praze na Svatého Bruna roku 2023; posledni update na Svatého Andreje.

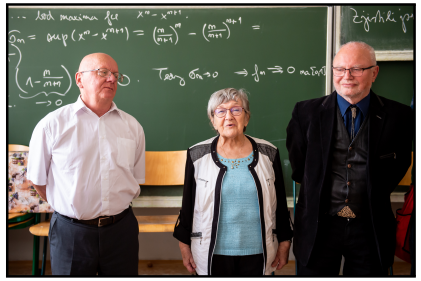
Na setkání se těší: J. Antoch, G. Dohnal, D. Hlubinka, M. Pešta, I. Stankovičová a – především – účastníci.

VŠECHNO NEJLEPŠÍ PROF. ANTOCHOVI!
HAPPY BIRTHDAY, PROF. ANTOCH!

Výbor společnosti



Významné životní jubileum oslavil v květnu bývalý dlouholetý předseda ČStS prof. Antoch. Při té příležitosti se v Praze konalo sympózium. IB otiskuje texty některých zahraničních gratulantů.



Zdroj: String art font, `https://erikdemaine.org/fonts/stringart`.
Foto: Matúš Maciak.

SYMPÓZIUM PRO JAROMÍRA ANTOCHA
SYMPOSIUM FOR JAROMÍR ANTOCH

Michal Pešta

Dňa 12. mája 2023 sa konalo sympózium pre profesora Jaromíra Antocha k oslave jeho životného jubilea. Oslavy 70. narodenín doyena českej štatistiky sa uskutočnili na pôde jeho alma mater v Karlíne. Pri tejto jedinečnej príležitosti vystúpili pozvaní zahraniční i českoslovenští kolegovia, spolupracovníci a priatelia pána profesora, ktorí ovplyvnili jeho život – konkrétne (v chronologickom poradí vystúpení, bez titulov a s názvami prednášok):

- Jana Jurečková (UK) – Score functions rediscovered in Jaromír's dissertation.
- Noël Veraverbeke (Hasselt University) – Conditional residual quantiles.
- Daniela Jarušková (ČVUT) – Changepoint analysis of Klementinum temperature series.
- Pascal Sarda (University of Toulouse) – Some incursions into nonparametric statistics and functional data analysis.
- Gejza Wimmer (Slovak Academy of Sciences) Straight-line error-in-variables calibration model.
- Jean-Marc Deshoulliers (University of Bordeaux) – Video pozdrav na dialku.
- Marie Hušková (UK) – Change point – asymptotic and computational issues.
- Francesco Mola (University of Cagliari) – Some personal considerations about Jaromír.
- Michal Černý (VŠE) – Life and work with JA.
- Viktor Witkovský (Slovak Academy of Sciences) – Numerical inversion of characteristic functions for exact probability distribution.
- Michal Pešta (UK) – Last doctorand.

Prof. RNDr. Jaromír Antoch, CSc. (8. 5. 1953), sa ako medzinárodne uznávaný vedec a obľúbený vysokoškolský pedagóg obrovskou mierou zasadil o rozvoj matematickej a výpočetnej štatistiky v Českej republike.

Bol okrem iného prezidentom International Association for Statistical Computing IASC (2007–2009), predsedom European Regional Section of IASC (1992–1996), predsedom České statistické společnosti (2001–2007), člen rady International Statistical Institute (2007–2011).

Pro akademické a vědecké pracovníky (a možná i pro další zájemce) bude jistě zajímavá sedmá kapitola nazvaná „Věda a temná data: povaha objevování“. Zajímavá je v ní zejména diskuse týkající se replikační krize (str. 158–183). Hand obhajuje názor, že „k narušení vědeckého procesu *nedochází*“ a že „věda prokazatelně funguje“. Zajímavé je, že po tomto prohlášení předkládá dlouhý výčet neduhů, kterými věda trpí, a které by bylo možno považovat za protiargumenty výše uvedených tvrzení. Píše např. o publikačním zkreslení (s. 160), p-hackingu resp. problému mnohonásobného testování (s. 165), HARKingu (Hypothesis After the Result is Known, s. 168). Problému (ne)replikovatelnosti výsledků vysvětluje principem regrese k průměru. Hand (s pro mě překvapivým klidem) konstatuje: „Neměli bychom být překvapeni, pokud anomální výsledek zmizí, a měli bychom očekávat že 'deklarované výsledky výzkumu jsou často falešné'. . . “. Hand tedy vědeckému procesu věří navzdory jeho četným neduhům, jichž si je vědom. Pozornost věnuje také podvodům ve vědě. Za povšimnutí v této souvislosti stojí citace téměř 200 let starého díla Charlese Babbage [1], který v roce 1830 psal o tom, že „Vědecká zkoumání jsou více než jakákoliv jiná vystavena nájezdům podfukářů.“ Popsal přitom čtyři (hlavní) druhy podvádění: mystifikace, padělání, ořezávání a vaření. Hand tyto pojmy (resp. „techniky“) detailně vysvětluje (s. 171–180).

Zatímco v prvních sedmi kapitolách knihy se Hand věnuje „původu a důsledkům“ temných dat, ve druhé části knihy tvořené třemi kapitolami vysvětluje nejprve, jak s temnými daty nakládat (kapitola 8 je úvodem k tématu práce s chybějícími hodnotami) a „jak mít z temných dat užitek“ (v kapitole 9 pojednává o technikách jako např. simulace, boosting, bootstrap, ale také o bayesovské inferenci). V závěrečné desáté kapitole pak nabízí „kategorizaci“ temných dat (je zmíněno patnáct různých případů temných dat).

Kniha je psána populárně, bez matematických vzorců, je plná zajímavých příkladů. Když jsem si však položil otázku, zda je publikace určena široké veřejnosti, dospěl jsem k přesvědčení, že spíše ji ocení lidé, kteří s daty sami pracují. Rozhodně bych ji doporučil studentům oborů matematika, informatika, statistika, data science, a pochopitelně také nejrůznějších přírodních věd, jako motivaci k dalšímu studiu, resp. k zapsání přednášek týkajících se pokročilých statistických metod.

Reference

[1] Babbage, Ch. (1830): Reflections on the Decline of Science in England, and on Some of Its Causes. Londýn, B. Fellowes, 1830. `https://www.gutenberg.org/files/1216/1216-h/1216-h.htm` *cit. 29*

[2] Kulich, M. (2021): O datech. `http://www.statspol.cz/wp-content/uploads/2021/02/o_datech_publ.pdf` *cit. 28*

k dispozici. Pokud nechápou, jak data vznikla a co v nich může chybět, vystavují se vážnému riziku, že budou hledat pouze tam, kam dohlédnou, a ne tam, kde by mohly ležet odpovědi.“

Dovolím si nyní krátce připomenout text Michala Kulicha *O datech* z února 2021 [2], tedy z doby nej(h)různějších analýz „covidových dat“, jejichž autoři výše uvedené Handovo varování nejspíš neznali, nebo si ho nebrali k srdci. Sami posuďte, jak Kulichův text s Handovým souzní: „Potřebujeme aspoň nějaká data, která [...] jsou sbírána na pečlivě vybraných vzorcích podle předem připraveného plánu a naprosto konzistentními metodami, které minimalizují možnost chyb a omylů. Takhle se mají dělat pořádné studie. Přitom každá taková studie musí mít jasně stanovený účel, který determinuje, jaká data je potřeba sbírat, kdy a na kom.“ A Handův text zase souzní s Kulichovým: „... data, která potřebujete shromáždit, analýza, kterou budete provádět, i odpověď, kterou dostanete, to vše závisí na tom, co chcete zjistit.“ (Hand, str. 81)

Myšlenku potřeby znalosti procesu vzniku dat Hand detailně rozvedl ve druhé kapitole své knihy, v níž shrnul tři základní strategie tvorby datových souborů: první dva způsoby se týkají observačních dat, přičemž v prvním případě jsou zaznamenávány údaje o všech statistických jednotkách, zatímco ve druhém případě jen o některých. Třetím typem dat jsou pak data experimentální, tedy vzniklá při „experimentu“ spočívajícím v cíleném ovlivňování podmínek, za nichž jsou data sbírána. Hand přitom u všech tří způsobů sběru dat věnuje velkou pozornost možnosti výskytu temných dat.

Není mým cílem převyprávět obsah celé knihy. Přesto bych se rád zastavil u některých pasáží, ke kterým se budu v budoucnu zřejmě opakovaně vracet.

Jedním z těchto míst je pátá kapitola, kde je mj. zmíněna směrnice EU z roku 2004 o rovnosti žen a mužů, která „má bojovat proti diskriminaci na základě pohlaví“. Hand (na str. 120) zmiňuje dopad této směrnice na výši pojistného u pojištění automobilů. To bylo dříve nižší pro ženy, neboť data ukazovala, že je u nich menší pravděpodobnost, že budou mít nehodu. V momentě, kdy tato daty podložená skutečnost (rozdílnost v rizikovosti žen oproti mužům) musí být dle práva ignorována (Hand zmiňuje rozhodnutí soudu z roku 2013), je praktickým důsledkem této ignorace zvýšení pojistného u žen (méně rizikových) a snížení u mužů (více rizikových). Hand se pak zamýšlí nad společenským přínosem tohoto rozhodnutí a diskutuje pochopení pojmu „spravedlnost“ – klade si otázku, zda je spravedlivé, aby mužům a ženám, kteří se ve všech ostatních vlastnostech ve statistickém modelu shodují, bylo účtováno rozdílné pojistné, když data ukazují, že mají rozdílná rizika? Poznamenejme ještě, že Hand interpretuje nemožnost zohlednění určitého faktoru (např. pohlaví pojištěného) tak, že se z tohoto faktoru stávají „temná data“.

ÚVODNÉ SLOVÁ K PREDNÁŠKE PROF. WIMMERA NA SYMPÓZIU PRI PRÍLEŽITOSTI 70. NARODENÍN PROF. ANTOCHA

INTRODUCTION TO THE LECTURE DELIVERED BY PROF. WIMMER FOR THE 70TH BIRTHDAY OF PROFESSOR ANTOCH

Gejza Wimmer

Dear Professor Antoch,

I offer my hearty congratulations, much satisfaction, professional and personal success on the occasion of your significant life anniversary on behalf of the Mathematical Institute of the Slovak Academy of Sciences, on behalf of the Slovak mathematical community, especially on behalf of Slovak statisticians, scientists and workers involved in solving applied statistical problems in practice and working in the teaching of statistics.

You are one of the most important Czech scientific personalities, who contributed to the very fruitful cooperation between the Czech and Slovak statistical community, to the professional growth of our students, doctoral and post-doctoral students in statistics, as well as scientific workers and workers involved in solving applied statistical problems in practice and working in the teaching of statistics. You have achieved all this above by being the soul of regular ROBUST conferences for a long time. As part of these conferences, there is an excellent possibility of constant contact between Slovak statisticians and their Czech colleagues, the possibility of establishing professional cooperation and carrying out consultations.

Dear professor, thank you very much for that.

GRATULACE Z JAPONSKA
CONGRATULATIONS FROM JAPAN

Ryozo Miura

Adresa: Professor Emeritus Hitotsubashi University, Japan

I would like to sincerely congratulate Professor Jaromir Antoch on his 70th birthday and would also like to thank him for his friendship, kindness and hospitality provided to me for years.

In the middle of 1980s, I came to know the name of Jaromir Antoch as a young statistician working on the asymptotic theory of statistics which I was also working on, in a conversation with a Czech probabilist Dalibor Volny who was visiting Osaka. Then I wrote a letter to Jaromir to start exchanging our research papers. Based on these communications, I brought up a plan of visiting Prague in 1988, but this was not realized regrettably for some reason: my physical conditions.

In 1989, I moved to Tokyo from Osaka, from Osaka City University to Hitotsubashi University and then I became very busy on working in a new field of now so-called Financial Engineering/Mathematics. Thus, I was quite absent from statistical meetings and journals for a long time. However, 20 years later in 2008, I found the name of J. Antoch in the list of main participants at the International Conference for Computational Statistics held in Yokohama, Japan. After making sure that this is the statistician I had known during 1980s, we met each other in Yokohama. There, Jaromir asked me if I still want to visit Prague. I said "Oh, Yes!", and my visit to Charles University was realized the following year.

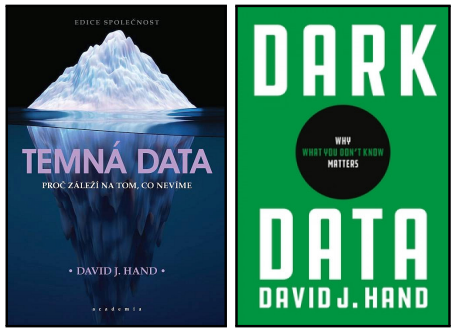
My interests in visiting Prague and Czechia were two-fold; academics and history/culture. Academic interest is in Rank statistics, Asymptotic Theory of Statistics. I used to read the book Theory of Rank Test by Hajek and Sidak, and I knew Prague is the center for Rank statistic. Cultural Interests is, as well as Music performances, History/Culture of Czechia especially the Introduction of Christianity (by Saints Cyril and Methodius) during the 9th century and the movement by Jan Hus, Petr Chelcicky and their followers, with the related matters.

On my first visit to Prague in November 2009, Jaromir arranged a meeting with Professor Jana Jureckova, it was my first time to meet her. During 1980s she recognized my work on adaptive estimates of location which I published in a very local journal of my department of commerce (in English),

RECENZE KNIHY TEMNÁ DATA
BOOK REVIEW: DARK DATA

Ondřej Vencálek

E-mail: ondrej.vencalek@upol.cz



Nakladatelství Academia letos (2023) vydalo překlad knihy *Temná data* předního britského statistika Davida J. Handa. Kniha má 272 stran. Anglický originál je z roku 2020. A jde o zajímavé čtení, které bych doporučil všem, kdo pracují s daty.

Slovy *temná data* označuje Hand data, která z nejrůznějších důvodů (ať už vědomě či nevědomě, záměrně či nechtěně) nejsou k dispozici. Není proto divu, že takto široce chápaná (nepozorovaná) data nachází prakticky „na každém rohu", tedy spíše ve všech možných situacích, a to včetně těch, kdy zdánlivě máme k dispozici pozorování všech statistických jednotek. Autor, který celý život pracoval jako statistik, těží ze svých bohatých zkušeností – na řadě zajímavých a často netriviálních příkladů vysvětluje, *proč záleží na tom, co nevíme* (tak zní podtitul knihy). Právě zdůraznění důležitosti přemýšlení o tom, co v datech není, činí knihu výjimečnou. Hand nabádá nejen všechny analytiky k ostražitosti a vede je k pochopení rizik, která s sebou temná data nesou.

Poselství knihy velmi dobře vystihuje vtipné přirovnání v samotném závěru knihy: Vypráví starou anekdotu o opilci, který hledá klíče pod sloupem veřejného osvětlení. Proč zrovna tam? Kupodivu nikoliv proto, že by mu spadly právě tam, ale proto, že jedině tam je dost světla, aby je uviděl. Hand pak uzavírá: „Výzkumníci, analytici a vlastně všichni, kdo se snaží z dat získat význam, jsou jako onen opilec, omezí-li se jen na data, která mají

References

[1] Bagdonavičius, V., Nikulin, M. (2002): *Accelerated Life Models, Modeling and Statistical Analysis.* Chapman & Hall/CRC, Boca Raton. *cit. 19, 20*

[2] Buckley, J., James, I. (1979): Linear regression with censored data. *Biometrika* **66**, 429–436. *cit. 14, 16*

[3] Cox, D. R. (1972): Regression models and life-tables (with Discussion). *J. R. Stat. Soc.* **B 34**, 187–220. *cit. 14*

[4] James, I. R., Smith, P. J. (1984): Consistency results for linear regression with censored data. *Ann. Statist.* **12**, 590–600. *cit. 17*

[5] Jin, Z., Lin, D. Y., Ying, Z. (2006): On least-squares regression with censored data. *Biometrika* **93**, 147–161. *cit. 17*

[6] Kalbfleisch, J. D., Prentice, R. L. (2002): *The Statistical Analysis of Failure Time Data.* Wiley, New York. *cit. 19*

[7] Koul, H., Susarla, V., Van Ryzin, J. (1981): Regression analysis with randomly right-censored data. *Ann. Statist.* **9**, 1276–88. *cit. 14, 17*

[8] Lin, D. Y., Wei, L. J., Ying, Z. (1998): Accelerated failure time models for counting processes. *Biometrika* **85**, 605–618. *cit. 20*

[9] Miller, R. G. (1976): Least squares regression with censored data. *Biometrika* **63**, 449–464. *cit. 14, 15*

[10] Novák, P. (2013): Goodness-of-fit test for the accelerated failure time model based on martingale residuals. *Kybernetika* **49**, 40–59. *cit. 20*

[11] Volf, P., Timková, J. (2014): On selection of optimal stochastic model for accelerated life testing. *Reliability Engineering & System Safety* **131**, 291–297. *cit. 20, 25*

and she referred it in a chapter of Handbook for Nonparametric Statistics. (I wanted to meet her to express my gratitude for this matter.) After this visit, I was fortunate to have opportunities to give a seminar talk and then in 2011 a series of lectures on my work in the fields of Mathematical Statistics, Mathematical Finance and Financial Engineering, and Financial Data Analysis, at the department of Statistics in Charles University.

Jaromir knew my cultural interest in the country of Czechia as well as my academic ones. He thus arranged my visit to Liberec Technical University in 2011 where I gave a few general lectures, and Professor Linka kindly took me for an excursion to a mountain area called Paradise. As for my historical Interest in Chelcicky, there was a very kind help and hospitality of Professor Jana Klicnarova at Faculty of Economics of South Bohemia University. I was introduced to her by Professor Dalibor Volny on my first visit with him (an academic visit) to Ceske Budejovice in November, 2012. This was a beginning of my recurrent visit to the university. Later on, I was given an opportunity to give lectures to undergraduate students in English (maybe under Erasmus program). She understood my historical interests and kindly guided me to the Chelcicky museum and to several historical places (including Tabor 2) related Chelcicky. Jaromir also arranged my visit to Masaryk University in Brno in November 2012 where I gave a lecture. Professor Ivanka Horova arranged for me some visit to the interesting places in Brno, e.g. Mendel Museum. Right after Brno, Professor Marek Omelka on his holidays at his hometown, helped me very much for my stay and visit to Velehrad and the Church there where I could learn about Saints Cyril and Methodius. Several years later, Jaromir introduced me to Professor Ondrej Vencalek to make my visit to Palacky University in Olomouc in April 2018. Then I could fortunately make a visit again in the next year 2019 to have some academic exchange with him and his faculty. Ondrej kindly took me for a drive to Velehrad and other nearby towns/cities in Moravia.

One day I discussed with Jaromir about an academic exchange with Japanese statisticians. Later this was realized in a form of the academic exchange agreement between Department of Statistics Charles University Prague and Institute of Statistical Mathematics Tokyo Japan. Based on this agreement, a team of three statistician: Professors Marek Omelka, Daniela Jaruskova, and Jaromir Antoch together visited the Institute of statistical mathematics in March, 2012. Then, Professor Marie Huskova visited the Institute in September 2013. Jaromir visited Japan on another occasion as well. It was in November 2017. He was invited by Professor Yasumasa Matsuda to visit Department of Economics of Tohoku University in Sendai for a workshop. On this occasion I took Jaromir to Sakunami Onsen near Sendai over

the weekend to experience a Japanese open door hot spring (Iwa-buro): a rock bath with a pretty river-side view. Also, we discussed on the development of rank statistics for linear regression models and I made many questions about the theoretical assumptions set for deriving properties of the tests and estimators, especially about the assumptions set for the independent variables (explaining variables). He explained me the most recent form of the theoretical assumptions. A summary of this will be coming up as a paper for publication.

Since 2009, I visited Czech every year (except in 2013) until 2019. In 2020, a meeting of "Robust Statistics" was planned to hold in Bardejov, Slovakia where I wanted to join. But because of the Covid-19 Pandemic this meeting was cancelled for the year. For me, every visit to Czechia starts from arriving at the airport in Prague: Vaclav Havel Airport, then stay for a few days at least to visit the department of statistics to meet Jaromir. Every time I visited his office, it was my opportunity for gaining my knowledge to recover my absence from the field of theory of statistics for 20 years. I questioned him on recent developments on theory of Rank Test and Estimation, then he kindly guided me to the related papers and the related chapters of books. Alongside to academics, Jaromir guided me to many places to attract my cultural/historical interests such as Rudolfinum, Vysehrad, Strahov Monastery/Library, Japan Garden, etc. And he also took me even for a long drive to visit monumental places outside of Prague area. Most recent and memorable one is Mitsuko exhibition at Horsovsky Tyn.

Yes, I owe him a lot and had a good time with him. I would like to thank him again and wish him good health and his continued success in research and teaching.

Ryozo Miura, September 2023

to improve its precision with growing data size. The method of Koul et al., however, has again demonstrated that the large variability of "synthetic" data can cause an instability of trend estimation.

5. Concluding remarks

We have recalled three "historical" methods analyzing the LRM with censored data, compared them mutually, with a direct (numerical) solution via the MLE under assumption of normally distributed errors, and also with a solution in the framework of the semi-parametric AFT regression model. The results seem to support a conjecture that the BJ estimator is the most reliable from the three traditional approaches. Very good results were achieved by the last two methods. It could be said that they represent contemporary approaches to the problem of linear regression with censored data. A variant methods are nowadays based on the Bayes approach and the use of the Markov chain Monte Carlo (MCMC) procedures (as for example in Volf and Timková, 2014).

Let us here also mention a connection of the LR problem with the Cox model, at least with its special case. Let us assume that the baseline distribution is the Weibull one, with the distribution function $F_0(t) = 1 - \exp\{-(t/a)^b\}$ defined for $t > 0$ and with positive parameters, scale a and shape b. Further, let the dependence on a covariate x follows the Cox (proportional hazard) regression model, let us denote the Cox regression parameter by γ. Then the cumulative hazard rate equals:

$$H(t|X = x) = H_0(t) \cdot \exp(\gamma\, x) = (t/a)^b \cdot \exp(\gamma\, x) = \left(\frac{t}{a(x)}\right)^b, \qquad (12)$$

where $a(x) = a \cdot \exp\{-\gamma\, x/b\}$. Thus, the scale parameter $a(x)$ depends on covariate, while the shape parameter remains b. Simultaneously, $H(t|X = x)$ can be rewritten as

$$H(t|X = x) = \left(\frac{t \cdot \exp(\gamma/b \cdot x)}{a}\right)^b = H_0(t \cdot \exp(\beta\, x)), \qquad (13)$$

which corresponds to the AFT model with regression parameter $\beta = \gamma/b$. Hence, if in the case of Weibull baseline distribution the estimation uses the Cox model framework, we then obtain the estimate of $\gamma = \beta \cdot b$ instead of β, while the estimates of $H_0(t)$ in both Cox and AFT models should be close one to each other (because both are consistent, both should tend to the Weibull(a, b) cumulative hazard rate).

4. *Direct MLE* under the assumption of normally distributed residuals yielded $a = 2.0208$, $b = -1.0103$, $s_r = 0.9853$.
5. *Semi-parametric estimator* in the framework of the AFT model: We are able to estimate just the slope parameter β, while the rest of model is hidden in the non-parametric estimate of the cumulative baseline hazard rate. In our case, this estimate should be close to the CHR of the log-normal distribution with parameters $\mu = \alpha = 2$, $\sigma = 0.1$. The proximity of both curves is shown in Figure 4. Estimate of β was $b = -1.0369$.

4.2. Example 2

The purpose was to explore the reliability of all methods in more detail. In particular, we were interested in the precision of estimates of the regression parameter β. Therefore, each simulation (in the framework of the same model configuration as in Example 1) was repeated 200 times, for each method. In such a way, from each method, a set of 200 estimates of β was obtained. They were sorted, the interval from the 6th to 195th value taken as representing a 95% "precision interval" characterizing each method. The results are in the following Table 1, its left part. It contains also the mean of 200 estimates.

Method	$N = 100$: 95% "PI"	Mean	$N = 500$: 95% "PI"	Mean
Miller	$(-0.9740, -0.8410)$	-0.9164	$(-0.9990, -0.9550)$	-0.9774
BJ	$(-1.0188, -0.9189)$	-0.9735	$(-1.0220, -0.9739)$	-0.9987
KSVR	$(-2.2786, 0.0590)$	-0.6009	$(-1.2042, -0.5121)$	-0.7991
MLE	$(-1.0400, -0.9650)$	-1.0002	$(-1.0100, -0.9800)$	-0.9981
AFT	$(-1.0434, -0.9510)$	-0.9989	$(-1.0206, -0.9819)$	-1.0016

Table 1: Results of estimation of regression parameter β, 200 repetitions, from data of size $N = 100$ and $N = 500$.

The right part of Table 1 shows the results of estimation from larger data-sets, with $N = 500$, each experiment was again repeated 200 times. It is seen that the parametric MLE method and the estimation in the framework of semi-parametric AFT model have provided the most reliable estimates of the regression parameter. The performance of the Buckley and James method was just slightly weaker, while the method of Miller seemed to be biased systematically. Simultaneously, all methods have shown a tendency

PŘÁNÍ K 70. NAROZENINÁM PROF. ANTOCHA
HAPPY 70TH BIRTHDAY FOR PROFESSOR JAROMIR ANTOCH

Noël Veraverbeke

On May 12, 2023 a small symposium was organized at the department of Probability and Statistics of Charles University to celebrate the 70th birthday of our colleague Jaro Antoch. I felt honoured to be invited and this brought back many nice memories of this dear colleague and friend.

I met Jaro at many occasions in the past when I visited his department. My first contacts with the group in Prague were through Jana Jureckova, Marie Huskova and of course Jaro. I met Jana already in 1977 in Vilnius and in 1984 she came for a research stay of two months to my university in Hasselt, Belgium.

I visited Prague for the first time in 1985 and that is the first time that I met Jaro. I was allowed to stay in a beautiful house if the university in Celetna street and Jaro took me around in town. Prague looked very different, compared to nowadays. It was winter and there were almost no tourists, which is difficult to imagine today. Jaro was very proud of his town and he showed me the many nice spots in the city. We tasted the Czech beers and wines. I remember very well that one night Jaro also took me to the Rudolfinum to listen to the Czech Philharmonic Orchestra. Really unforgettable.

Jaro was invited to our university in Belgium in the academic year 1987–1988 for a four months research stay. During that period, he really became a good friend and colleague of many of us. Jaro gave us a seminar, every Friday, for the whole group of our statisticians. He did this in his own style and with his own humour. The topic was CART (Classification and Regression Trees) and he followed the book of Leo Breiman that was published in 1984. But Jaro did much more than this. He also spent a lot of time in exploring our software and the (modest) computer infrastructure on our campus in Diepenbeek.

A nice anecdote is that Jaro organized towards the end of his stay a "First Czech evening in Diepenbeek". To prepare this event, Jaro and I went by car to Brussels where we stopped at a rather mysterious house which turned out to be the Czechoslovak Embassy. He went inside and I waited in the car. I remember that I got worried because it took very long. But finally, he come out, fully loaded with plenty of material for his plan.

It became a great evening, with promotion films, beautiful tourist books, Czech food, beer and of course also Becherovka. On my bookshelf at home, there is still a nice present from this evening: a beautiful book "Walks through Prague" with 500 black and white photos of all the famous spots in the city.

I returned many times to Prague for conferences, research or teaching and each time it is always a great pleasure to meet each other again.

Dear Jaro: Ad multos annos!

GRATULACE ZE SLOVENSKA
CONGRATULATIONS FROM SLOVAKIA

Karol Nemoga

MATEMATICKÝ ÚSTAV
SLOVENSKÁ AKADÉMIA VIED, v. v. i.
Štefánikova 49, 814 73 B r a t i s l a v a

TEL: (0) 57510414, 5249 7316

E-mail: mathinst@mat.savba.sk
http://www.mat.savba.sk

Vážený pán
Prof. RNDr. Jaromír Antoch, CSc.
Katedra pravděpodobnosti a matematické statistiky
MFF UK
Sokolovská 49/53
Praha

Bratislava, 8. V. 2023

Vážený pán profesor,

dovoľte mi, aby som Vám za seba, za Matematický ústav SAV, v.v.i., za celú slovenskú matematickú obec, špeciálne za bratislavských aj mimo bratislavských štatistikov, zaprial veľa spokojnosti, profesionálnych a osobných úspechov pri príležitosti Vášho významného životného jubilea.
Ste jednou z najvýznamnejších českých vedeckých osobností, ktorá sa zaslúžila o veľmi plodnú spoluprácu medzi českou a slovenskou štatistickou obcou, o odborný rast našich študentov, aj doktorandov a post-doktorandov štatistiky, ale aj vedeckých pracovníkov a pracovníkov pôsobiacich pri riešení aplikačných štatistických problémov v praxi a pri vyučovaní štatistiky. Toto všetko ste dokázali predovšetkým tým, že ste dlhodobo dušou pravidelných konferencií ROBUST. V rámci týchto konferencií existuje výborná možnosť neustáleho kontaktu slovenských štatistikov s ich českými kolegami, možnosť nadviazať odbornú spoluprácu, realizovať konzultácie. Patrí Vám za to, vážený pán profesor, veľká vďaka.

Želáme Vám, vážený pán profesor, veľa pevného zdravia a ešte veľa tvorčích úspechov a spokojnosti v rodinnom kruhu.

Ad multos annos !

doc. RNDr. Karol Nemoga, CSc.
riaditeľ MÚ SAV, v.v.i.

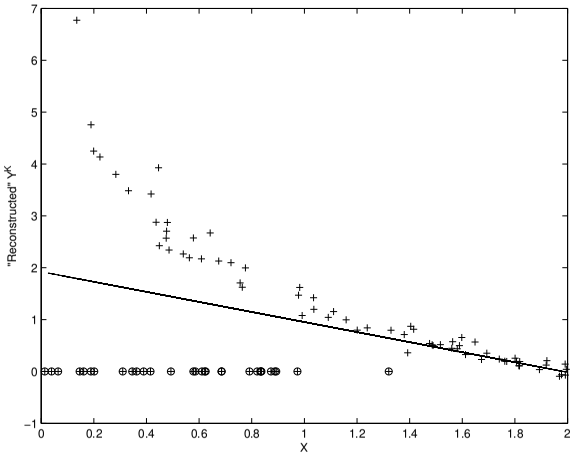


Figure 3: Values $Y^K(X)$ reconstructed via the Koul et al. method (covariate x is on horizontal axis). All censored values, denoted by 'o', were set to zero.

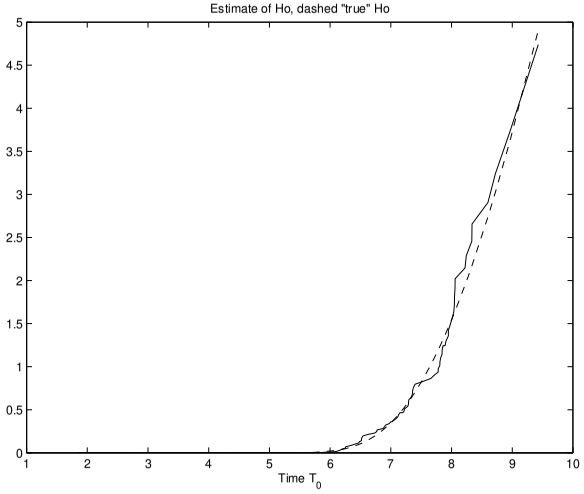


Figure 4: AFT model: Nonparametric estimate of the baseline CHR (full curve) and its comparison with 'true' log-normal CHR (dashed curve).

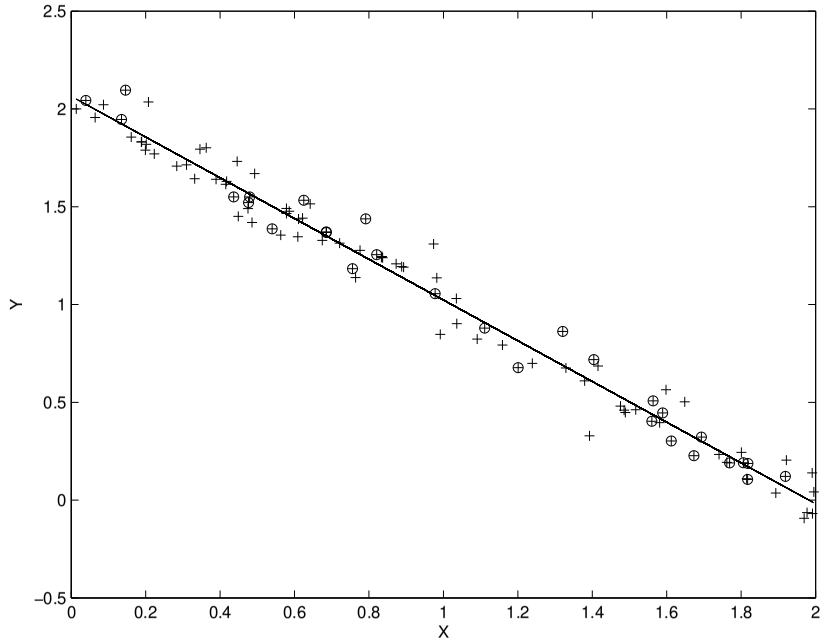


Figure 2: Results of the BJ procedure after 2nd iteration. Reconstructed censored values are denoted by 'o'.

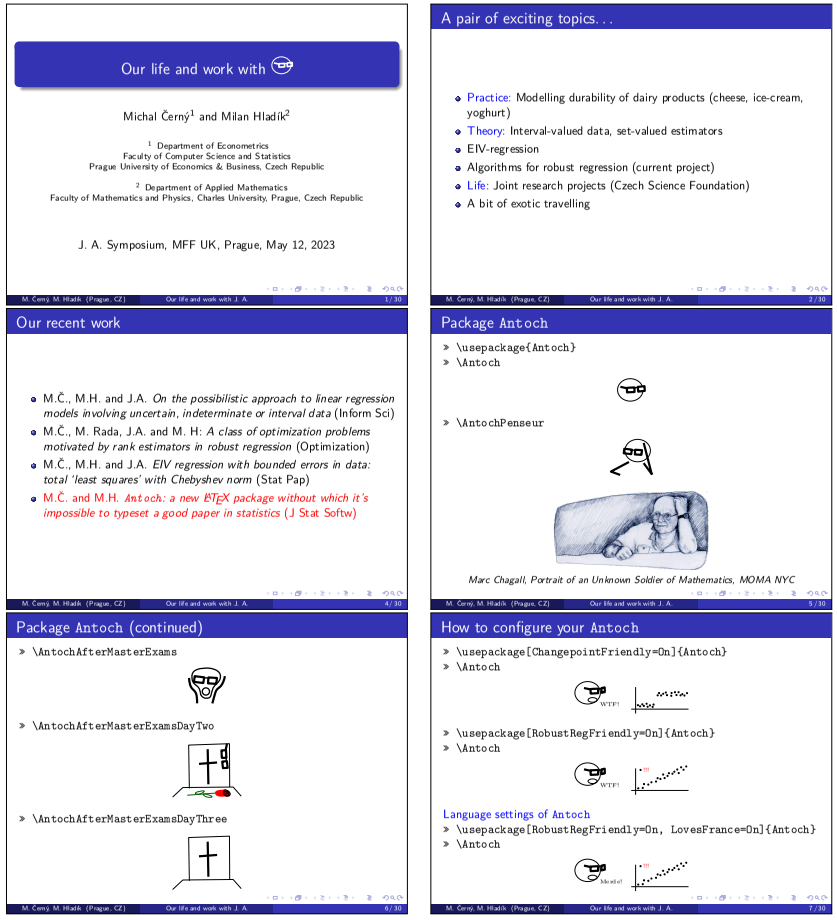
Then, the methods of estimation described above were tested on the same data, with the following results:

1. *Miller's estimator* yielded $a = 1.904$, $b = -0.936$, $s_r = 0.165$.
2. *Buckley and James method:* Figure 2 presents the result of 2nd iteration of their procedure. It was revealed that there the smallest standard deviation of residuals was achieved, $s_r = 0.0912$, corresponding estimates of parameters were $a = 2.0645$, $b = -1,0415$. After that, the results of further steps started to oscillate.
3. *Koul et al. estimator:* Figure 3 shows the data after proposed way of "reconstruction". It illustrates rather clearly the drawback of this method. Nevertheless, in this instance, the estimates were quite reasonable, $a = 1.9228$, $b = -0.9696$. However, as expected, estimated $s_r = 0.3731$ was much higher.

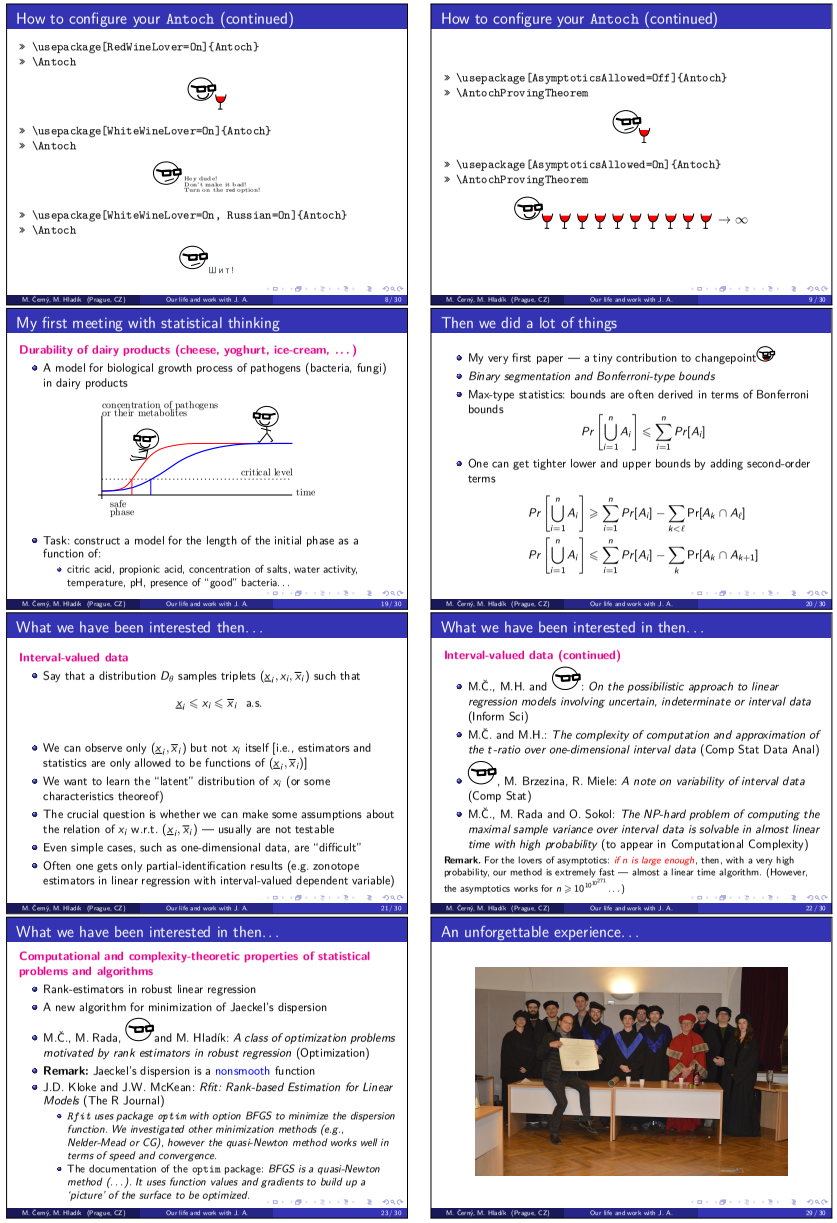
NÁŠ ŽIVOT A SPOLUPRÁCE S PROF. ANTOCHEM
OUR LIFE AND WORK WITH PROF. ANTOCH

Michal Černý, Milan Hladík

E-mail: cernym@vse.cz,hladik@kam.mff.cuni.cz



JA70 Poznámka redakce. Z velkého množství snímků jsme vybrali jen některé.



4.1. Example 1

First, just one set of data was generated, model parameters were estimated from them with the aid of all methods described above. Figure 1 shows the data before and after censoring.

From the non-censored data, estimates were $a = 2.0282\,(1.9911, 2.0653)$, $b = -1.0269\,(-1.0596, -0.9942)$, with 95% confidence intervals in parentheses, standard deviation of residuals was estimated as $s_r = 0.0978$. Optimal line is displayed in Figure 1, left side. For comparison, parameters were estimated also from censored data, not taking censoring into account, in order to show the bias of such a 'naive' approach. The following estimates were obtained: $a = 1.6172$, $b = -0.7868$, $s_r = 0.3568$, corresponding line is shown in the right part of Figure 1. These values were later used as initial estimates in the BJ procedure.

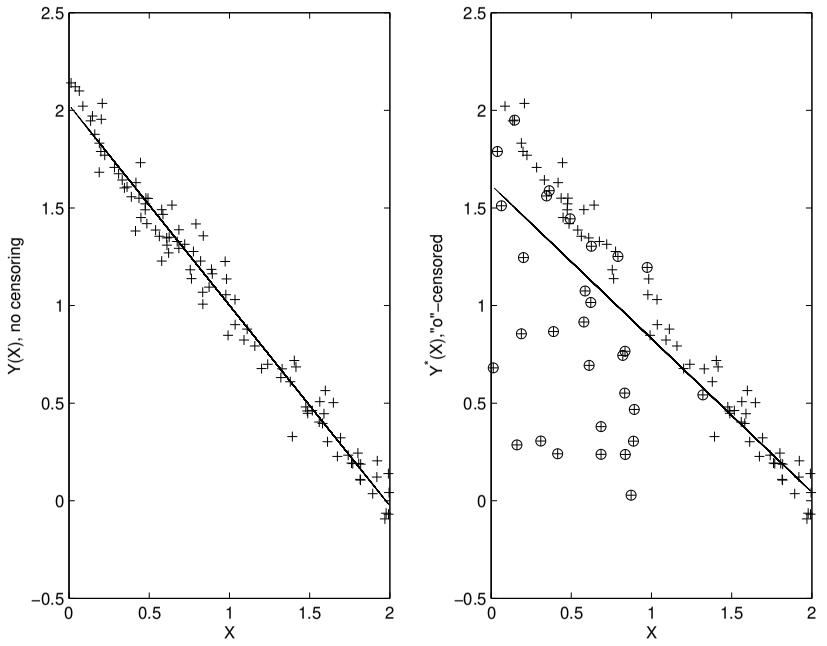


Figure 1: Values of Y before censoring (left) and after, censored values of Y^* are denoted by 'o' (right).

approach based on hazard rates could be preferred. Namely, let $\{T_i^*, x_i, \delta_i, i = 1, .., N\}$ be observed times of failures or censoring of N objects, their covariates, indicators of censoring, respectively. Then the likelihood equals (compare it with (9)):

$$L = \prod_{i=1}^{N} h_i(T_i^*)^{\delta_i} \cdot \exp\left(-\int_0^{T_i^*} h_i(t)\mathrm{d}t\right), \tag{10}$$

where $h_i(t) = h_0(t \cdot \exp(\beta x_i)) \cdot \exp(\beta x_i)$ is the hazard rate of i-th object at time t, h_0 is the baseline hazard rate of T_0. Theory of estimation and asymptotic properties are derived in Lin et al. (1998) and further developed also in Bagdonavičius and Nikulin (2002, Ch. 6). A Bayes approach offers a reasonable alternative, too (see Volf and Timková, 2014).

Lin et al. (1998) have shown that instead of the exact (and rather complicated) score function for β (i.e. obtained by the derivation of the log-likelihood w.r. to β), it is possible to use approximate score functions. Namely, the score function has the form

$$U(\beta) = \sum_{i=1}^{N} \left\langle x_i W_i(s) - \frac{\sum_j x_j W_j(s) I_j^*(T_{0i})}{\sum_k I_k^*(T_{0i})} \right\rangle \delta_i, \tag{11}$$

where $I_k^*(t)$ are indicators of risk in the scale of $T_{0i} = T_i \cdot \exp(\beta x_i)$, hence the unknown parameter β is hidden in them. While exact weights $W_i(s)$ depend on $h_0(s)$ and also on its first derivative, they may be substituted by a set of simpler functions, among them also by $W_i(t) = 1$ for all i and t. Lin et al. (1998) have proved that corresponding estimator of β retains good asymptotic properties. Hence, from such a score function it is possible to estimate β without knowledge of $h_0(t)$, similarly like with the aid of the partial likelihood in the Cox model case. More details can be found also in Novák (2013), who has proposed a method of goodness-of-fit test based on the random generation of residual processes, and has studied the test behavior in various situations.

4. Examples

We shall use here the following configuration of input variables and parameters: In the LRM (1), let us set the number of data $N = 100$, the covariates X_i distributed randomly uniformly in (0,2), regression parameter $\beta = -1$, intercept $\alpha = 2$, residuals ε_i following the normal distribution with $\mu = 0$, $\sigma = 0.1$. Further, let the censoring values be selected uniformly in the interval $(0, 3)$, the rate of censoring is then about 30%.

PŘEHLED VYBRANÝCH METOD ŘEŠENÍ ÚLOHY LINEÁRNÍ REGRESE S CENZOROVANÝMI DATY

AN OVERVIEW OF METHODS OF LINEAR REGRESSION ANALYSIS WITH CENSORED DATA

Petr Volf

Adresa: Faculty of Sciences, Humanities and Education, Technical University of Liberec, Czech Republic

E-mail: petr.volf@tul.cz

Abstrakt: Po úspěšném rozšíření a používání Coxova regresního modelu v 70. letech (a někdy až jeho nadužívání bez ohledu na splnění předpokladu proporcionality rizik) se pochopitelně zvýšila snaha najít vhodné metody také pro analýzu lineárního regresního modelu s cenzorovanými daty. Tento příspěvek přibližuje tři základní "historické" metody (autorů Miller, 1976; Buckley and James, 1979; Koul et al., 1981) a vzájemně je porovnává (na jednoduchém příkladu s náhodně generovanými daty). Výsledky jsou porovnány i s řešením pomocí maxima věrohodnosti při předpokladu normálně rozdělených odchylek, i s obecnějším semi-parametrickým modelem se zrychleným časem (AFT model).

Klíčová slova: Lineární regresní model, metoda nejmenších čtverců, cenzorovaná data, AFT model, Coxův model.

Abstract: As a reaction to results of D. R. Cox and others in '70s proposing regression models able to work with censored data, a number of statisticians (e.g. Miller, 1976; Buckley and James, 1979; Koul et al., 1981) have tried to extend the method of the least-squares in the same sense, i.e. to adapt it to censoring. The present contribution contains an overview of these approaches, compares them with direct numerical solution of corresponding maximal likelihood task and recalls also the AFT (Accelerated Failure Time) model as a semi-parametric generalization of the linear regression model. Performance of methods will be shown on simple numerical examples.

Keywords: Linear regression model, method of least squares, censoring, AFT model, Cox model.

1. Introduction

The standard linear regression model for N pairs of variables (X_i, Y_i), $i = 1, ..., N$ has the form

$$Y_i = \alpha + \beta \cdot X_i + \varepsilon_i, \quad i = 1, ..., N, \tag{1}$$

with ε_i assumed to be the i.i.d. variables with zero mean. The primary type of censorship considered here will be the random censoring from the right side. It means that there exist random variables C_i, say, independent mutually and of Y_i, and instead Y_i the values of $Y_i^* = \min(Y_i, C_i)$ are observed. As a rule, it is known whether an observation is censored or not, which is denoted with an indicator $\delta_i = 1$ if $Y_i \leq C_i$, $\delta_i = 0$ if $Y_i > C_i$.

After the introduction of the Cox model (Cox, 1972), the effort of statisticians has concentrated also to the solution of linear regression problem with censoring. In the present overview we shall recall several most important and more or less successful approaches provided in Miller (1976), Buckley and James (1979), and Koul et al. (1981). It is also interesting to follow these methods development from the point of view of their assumptions and of their real performance. Let us summarize first certain common assumptions:

1. Residuals ε_i are assumed to be i.i.d. with zero mean and finite variance.
2. Censoring variables C_i are independent mutually and of ε_i.
3. In general, C_i can depend on covariates X_i, however some approaches do not allow it, namely that of Koul et al. There the i.i.d. C_i-s are required.
4. When the estimation is based on the maximum likelihood principle, it has to be assumed that variables C_i do not contain any information on model parameters (this is called "non-informative" censoring, see also the assumptions for estimation in the Cox regression model); only then we can omit the likelihood part containing distribution of C_i.

This contribution recapitulates the above mentioned approaches first. Then the approach based on the MLE is presented (based on assumed type of the distribution of residuals). Further, the relationship of the linear model with the Accelerated Failure Time (AFT) regression model is shown. The performance of all methods is illustrated and compared with the aid of an example. Finally, also the Cox regression model is recalled, in order to show a connection of its parametrized version (with Weibull baseline distribution) with the linear regression problem

2. Overview of methods

All three traditional methods described bellow have one common feature, they all use the Kaplan-Meier "Product Limit Estimate" (PLE), though in slightly different ways. As the PLE is a nonparametric estimator of distribution function (from censored data), its use may lead to a lower precision

Gumbel (doubly-exponential, obtained by the logarithmization of the Weibull distribution), normal, logistic, exp-gamma, and other similar distributions.

Thus, depending on assumed distribution of errors ε_i, a direct maximization of the log-likelihood function corresponding to this distribution, with the aid of a convenient optimization procedure, could be a good choice. Let us recall that the "relevant" log-likelihood in the right-censored data case has the form (provided the censoring is "noninformative", only then the part concerning the distribution of variables C may be omitted):

$$L = \sum_{i=1}^{N} [\delta_i \cdot \ln(f_i(Y_i^*, \theta)) + (1 - \delta_i) \cdot \ln(1 - F_i(Y_i^*, \theta))], \tag{9}$$

where f_i F_i now denote the density and distribution function of variables Y_i, hence they depend on covariates, θ here denotes the model parameters. The MLE, via solving equation $\mathrm{d}L/\mathrm{d}\theta = 0$, can sometimes be found quite easily (in several steps of the Newton-Raphson algorithm); this is, among others, the instance of the Gumbel (log-Weibull) distribution, while the case with normal distribution has to be solved with the aid of a more clever iterative search procedure including repeated evaluation of normal distribution function.

3.1. Semi-parametric AFT regression model

Let us recall here also a general version of this model. For more information see e.g. Bagdonavičius and Nikulin (2002). , The accelerated failure time model is often considered as an alternative to the popular proportional hazard model of Cox, when the proportionality of hazards does not hold. The model assumes that the individual speed of time is changed by a factor depending on covariates. Standardly this factor has the form $\exp(\beta x)$, where x is a covariate, mostly constant in time. It follows that the distribution of time to failure T_i of an object with covariate value x has the distribution function $F(t) = F_0(t \cdot \exp(\beta x))$, where F_0 characterizes a baseline distribution (of a random variable T_0 with covariate $x = 0$). It also means that $T_{0i} = T_i \cdot \exp(\beta x_i)$ is an i.i.d. representation of T_0. The logarithmic transformation of the model then leads to the linear regression model as shown in (8). Notice also changed sign of parameter β in both models.

Hence, the statistical inference in the AFT model setting has to deal with unknown baseline distribution of T_0, the analysis is often complicated by the presence of censored data. As we have seen, the estimation in a linear model with censoring is not quite easy task (especially when the distribution of errors is not known). Therefore, similarly like in the case of Cox model, the

where $\overline{G}(.)$ denotes the survival function of the censoring variable C, which, if unknown, is substituted by its Kaplan-Meier PL estimator. A consequence of such a transformation is that the regression of Y^K on X is strongly heteroscedastic; notice that all censored values are now set to zero, while uncensored values are increased, sometimes rather considerably. Nevertheless, it still holds that $E(Y^K|X = x) = E(Y|X = x)$. It could again be proved easily:

In the model (1) denote $F()$ the distribution function of errors ε. Then

$$E(\delta \cdot Y^*) = \int_{-\infty}^{\infty} t\overline{G}(t)\,\mathrm{d}F(t-\alpha-\beta x), \text{ hence}$$

$$E\left(\frac{\delta \cdot Y^*}{\overline{G}(Y^*)}\right) = \int_{-\infty}^{\infty} t\,\mathrm{d}F(t-\alpha-\beta x) = \alpha+\beta x.$$

However, as the estimation method is based on the least squares (using reconstructed data), in fact its weighted variant should be used. Even the authors have pointed out that the variability of synthetic data around the response curve is neither constant nor symmetric and may expand, therefore the estimation of response curve may not be reliable. Figure 3 of Examples shows a result of such a "data reconstruction".

3. A general formulation of likelihood with censored data

All approaches described above utilize the method of least squares. This method corresponds in fact to the maximum likelihood estimation under the assumption of normally distributed errors. Therefore, the solution based on normal distribution of errors should be equivalent, moreover it avoids the use of the non-parametric PLE.

On the other hand, in general, we need not to restrict ourselves just to normal model, there are, especially in the context of reliability data studies, other natural possibilities. The linear model can arise from the logarithmization of typical "reliability" models for a variable T (characterizing for instance the time to failure) described for instance by the Weibull, log-normal, log-logistic, gamma, or other similar distributions. Namely, let again a variable Y fulfil the linear regression model, $Y = \beta \cdot X + \varepsilon$. Denote $T = \exp(Y)$ and $T_0 = \exp(\varepsilon)$, their distribution functions by F_T, F_0, resp. Then

$$F_T(t) = P(T \le t) = P(\exp(\beta \cdot X) \cdot T_0 \le t) = F_0(t \cdot \exp(-\beta \cdot X)), \qquad (8)$$

which corresponds to the parametric form of the AFT model. And vice versa, natural choices for distribution of ε in the linear regression model are then the

of estimation (in comparison to the case without censoring). Simultaneously, the analysis of asymptotic properties of estimates is then more difficult, confidence intervals for parameters can be obtained, as a rule, just with the aid of bootstrap.

Let us recall here briefly several basic notions from the area of statistical survival analysis: Consider a positive continuous-type random variable Z, with density f and distribution function F. $\overline{F} = 1 - F$ is then called the survival function, the hazard rate is defined as $h(z) = -\mathrm{d}\ln(\overline{F}(z))/\mathrm{d}z = f(z)/\overline{F}(z)$, the cumulative hazard rate (CHR) then equals $H(z) = \int_0^z h(t)\mathrm{d}t = -\ln(\overline{F}(z))$.

The form of the Kaplan-Meier PLE of distribution function F can be found elsewhere (for instance in the most of references cited here): Let Z_i^* be ordered sample of randomly right-censored realizations of Z, i.e. $Z_1^* \le Z_2^* \le \ldots \le Z_N^*$, δ_i corresponding censoring indicators, then the PLE equals

$$F^{\mathrm{KM}}(z) = 1 - \prod_{Z_i^* \le z} \left(\frac{N-i}{N-i+1}\right)^{\delta_i}.$$

Let us mention here also one practical detail of the PLE construction: the maximal value of data (here Z_N^*) is always taken as non-censored.

2.1. Miller's estimator

The way of construction of the Miller's estimator (Miller, 1976) seems to be rather reasonable, though there also is a common problem with proving estimator properties (its consistency and asymptotic normality). In particular, Miller has just shown his estimator being consistent when the censoring variable fulfils the regression with the same slope, i.e. $C \sim \alpha^c + \beta \cdot X$.

The usual least squares estimates a, b of α, β of model (1) are those values which minimize the sum of squares $\sum_{i=1}^N \varepsilon_i(a,b)^2 = \sum_{i=1}^N (Y_i - a - bX_i)^2$. This is equivalent to minimizing $\sum \varepsilon_i(a,b)^2 \cdot \Delta\widehat{F}(\varepsilon_i(a,b))$, where $\widehat{F}$ is the empirical distribution function (EDF) of residuals $\varepsilon_i(a,b) = Y_i - a - bX_i$ for fixed a, b, $\Delta\widehat{F}$ are the increments of $\widehat{F}$. Residuals $\varepsilon_i(a,b)$ are available just partially, with the same censoring pattern as Y_i^*, i.e. for given a, b we can compute $\varepsilon_i^*(a,b) = Y_i^* - a - bX_i$ and the censoring indicators are the same δ_i. Therefore, a natural step is to use the PLE of distribution function of residuals generalizing its standard EDF to the censoring case. Namely, the task is to minimize, over a, b,

$$\sum \varepsilon_i^*(a,b)^2 \cdot \Delta F^{\mathrm{KM}}(\varepsilon_i^*(a,b)), \qquad (2)$$

where F^{KM} is now the PLE of the distribution of residuals $\varepsilon(a,b)$ for given a,b (the increments ΔF^{KM} are 0 at censored values).

There is no problem to find optimal solution numerically, with the aid of a convenient search method. Further, notice that for fixed b the minimum with respect to a can be computed directly without any iteration, a just shifts the PLE. Similarly like in the standard case, the minimizing a equals the average of $Y_i^* - bX_i$ w.r. to the PLE, namely

$$\widehat{a_b} = \sum \varepsilon_i^*(0,b)^2 \cdot \Delta F^{\mathrm{KM}}(\varepsilon_i^*(0,b)), \tag{3}$$

with $\varepsilon_i^*(0,b) = Y_i^* - bX_i$. In such a way the search can be divided into two iterated steps, each solving just an one-dimensional problem.

Unlike Miller's, the following two methods are based on attempts to reconstruct censored data first, then these reconstructed values are used in the ordinary least squares.

2.2. Buckley and James

This method of analysis in the linear regression model with right-censoring, in fact the most popular one, has been proposed by Buckley and James (1979). It consists in a version of the EM (Expectation-Maximization) algorithm. Namely, the E step reconstructs censored values via conditional expectation, the M step then recomputes the model parameters from these values, using the least squares method.

Let the model have again the form (1). Again, no specific parametric form is assumed for the distribution of the error terms ε_i. They are just assumed to be i.i.d. Buckley and James have proposed to use weighted observations

$$Y_i^w = Y_i \cdot \delta_i + E(Y_i | Y_i > C_i) \cdot (1 - \delta_i). \tag{4}$$

It is seen that uncensored values are preserved, while censored are "corrected", increased. It is easy to show that for each x $E(Y^w|X = x) = E(Y|X = x)$:

For each fixed x (let us omit it for the moment) and for each fixed C we have that

$$E(Y^w|C) = E(Y \cdot \delta|C) + E[E(Y|Y > C) \cdot (1 - \delta)|C] =$$

$$= \int_{-\infty}^{C} \frac{y \cdot \mathrm{d}F_y(y)}{F_y(C)} \cdot P(\delta = 1|C) + \int_{C}^{\infty} \frac{y \cdot \mathrm{d}F_y(y)}{1 - F_y(C)} \cdot P(\delta = 0|C) =$$

$$= \int_{-\infty}^{C} y \cdot \mathrm{d}F_y(y) + \int_{C}^{\infty} y \cdot \mathrm{d}F_y(y) = EY,$$

independently of C, hence $E(Y^w) = E(Y)$ as well. Here F_y denotes the distribution function of Y for given fixed $X = x$.

The procedure of estimation starts from an initial estimate of $a^{(1)}$, $b^{(1)}$, obtained for instance by the least squares method from values (Y_i^*, X_i), and computes corresponding residual values ε_i^*, which again represent censored values of $\varepsilon_i(a^{(1)}, b^{(1)})$, with preserved indicators δ_i. The nonparametric estimator of their distribution, F^{KM}, is again obtained via the Kaplan-Meier PLE. From it we can "reconstruct" censored residuals,

$$\varepsilon_i^{(1)} = E(\varepsilon_i | \varepsilon > \varepsilon_i^*) = \frac{\sum_{\varepsilon_j^* > \varepsilon_i^*} \Delta F^{\mathrm{KM}}(\varepsilon_j^*) \cdot \varepsilon_j^*}{1 - F^{\mathrm{KM}}(\varepsilon_i^*)}, \tag{5}$$

Hence, censored outputs are then reconstructed as well, namely

$$Y_i^{(1)} = Y_i \cdot \delta_i + (a^{(1)} + b^{(1)} \cdot X_i + \varepsilon_i^{(1)})(1 - \delta_i). \tag{6}$$

Then new estimates, $a^{(2)}, b^{(2)}$ are computed via the least squares from innovated outputs $Y_i^{(1)}$. Quite similarly,

$$Y_i^2 = Y_i \cdot \delta_i + (a^{(2)} + b^{(2)} \cdot X_i + \varepsilon_i^{(2)})(1 - \delta_i),$$

where $\varepsilon_i^{(2)} = Y_i - a^{(2)} - b^{(2)} \cdot X_i$. This procedure could be repeated several times till convergence.

Later on, James and Smith (1984) have established the conditions and arguments for the consistency of the BJ estimator. However, the problem lies in the practical computation, in its iterative scheme. The convergence is not guaranteed, it is known that the iterations may end in oscillation between two or more set of values (see for instance Jin et al., 2006).

2.3. Koul, Susarla, and Van Ryzin

While the Buckley–James estimator uses (repeatedly) reconstructed censored values, the estimator of Koul et al. (1981) reconstructs all values, creates "synthetic data". This does not seem to be an ideal concept, as it changes the basic features of the data. Namely, they propose, instead of original censored data, to use

$$Y^K = \frac{\delta \cdot Y^*}{\overline{G}(Y^*)}, \tag{7}$$