Regression Models with an Interval-Censored Covariate

Klaus Langohr

PhD Thesis directed by
Dra. Guadalupe Gómez Melis

Universitat Politècnica de Catalunya

Barcelona, April 2004

Für Lesly,
und Mäm,
sowie in Erinnerung an
meine liebe Großmutter†

Acknowledgements

En primer lugar, quiero dar las gracias a Lupe por todo el trabajo que ha supuesto esta tesis también para ti. Ha sido un placer tanto personal como profesionalmente trabajar contigo y espero que sigamos colaborando así.

Thank you very much/ Ganz herzlichen Dank/ Moltes gràcies/ Muchas gracias a:

Axel, für deine allzeit geschaltene S-Plus Hotline. Das nächste Woizn geht auf mich.

Gracias a todo el departamento de EIO de la UPC por hacerme sentir tan a gusto aquí. Especialmente gracias a Celia y a Toni por su paciencia y ayuda en todo momento.

Elisenda, pel teu suport amb les correccions del resum a català.

Dr David M. Gay from AMPL. I appreciate very much all your immediate replies to my questions on AMPL. They were of big help.

Al Grup de Recerca en Anàlisi eStadística de la Supervivència (GRASS) per totes les discussions professionals i pel bon rotllo que hi havia.

A la Generalitat de Catalunya per la beca *2000FI 00408*, la meva.

Guillermo, por darme la oportunidad de participar en el proyecto sobre la vida útil y, en el marco de éste, conocer a tu tierra argentina.

To the external referees whose comments have been helpful to improve this thesis.

En general, a todos los amigos y compañeros que de una manera u otra nos han ayudado a lo largo de estos años, sobre todo en esos tiempos difíciles cuando la tesis estaba en un muy segundo plano. Gracias a vosotros he podido seguir con este trabajo en todo momento.

Daneben sei vor allem meiner Familie in Schwaben gedankt. Ohne eure so saumäßig wertvolle Unterstützung, Mäm, Gerhard, Janice, Gäb und Herrmann, wäre diese Arbeit in der vorliegenden Form keinesfalls möglich gewesen.

Last, but not least, muchísimas gracias a vos, Lesly, por tu gran apoyo en todo momento y en todos los sentidos. Tu fortaleza para superar esa enfermedad ha sido un ejemplo a seguir y me ha ayudado mucho más de lo que puedes imaginar.

Abstract

Survival analysis deals with the evaluation of variables which measure the elapsed time until an event of interest. In the area of clinical and epidemiological studies, this event of interest is often the onset of a disease, the disappearance of a disease's symptoms, or death. One particularity survival analysis has to account for are censored data. These arise whenever the time of interest cannot be measured exactly, but partial information is available. Several types of censoring are distinguished: a right-censored value occurs when the unobserved survival time is known to be superior to that value; left-censoring is present when the survival time is less than an observed time, and in case of interval-censoring, the survival time is observed within a time interval. We speak of doubly-censored data if also the time origin is censored.

Within the area of clinical and epidemiological studies, survival analysis plays a substantial role in studies on HIV and AIDS, given the long incubation periods and high lethality of this disease. Many drugs have been developed to prolong both the AIDS incubation period and the survival time, once AIDS is diagnosed, and improve the quality of life of HIV infected people. A major contribution to yield these objectives has been the development of the Highly Active Antiretroviral Treatments (HAART) in the mid-nineties.

The present PhD thesis has been mainly motivated by two studies on HIV/AIDS. One is on the survival of Tuberculosis patients co-infected with HIV in Barcelona and gives an idea about the extensive magnitude of the problem of HIV and AIDS in Spain. In this study, the time of interest is the elapsed time from start of Tuberculosis treatment until the death of the patients. Not all patients have died before the end of the study, hence, right-censored observations are present. In contrast with that, the observed censoring patterns in the second study are more complex. It is on injecting drug users in Badalona (Spain), most of whom became infected with HIV as a consequence of their drug addiction. The study aim has been to examine the possible association between the elapsed time from first injecting drug use until the moment of HIV infection and the subsequent AIDS incubation period. Whereas the former period is interval-censored since the moment of infection cannot be exactly observed, the latter is doubly-censored. On one hand, its time origin —the infection with HIV— is interval-censored, on the other hand, the AIDS onset has been partly left- and right-censored.

Methods for interval-censored data have received much attention during the last decades. Specific methods are required for these data since procedures for right-censored data, such as

the Kaplan-Meier estimator for the survival function (Kaplan and Meier 1958), are not valid for interval-censored data. Furthermore, replacing the unobserved survival times by the intervals' midpoints furnishes, generally, biased results, especially if the intervals are very broad.

The nonparametric estimation of a random variable's distribution function in the presence of interval-censoring has been addressed by several authors. The most important contribution is the one by Turnbull (1976) who introduces the idea of self-consistency to obtain the maximum likelihood estimator of the distribution function. Regarding the Cox proportional hazard model, Finkelstein (1986) proposes a procedure to estimate the model's parameters if the response variable includes interval-censored data. Nowadays, all main statistical software packages cover these methods.

Regarding doubly-censored data, De Gruttola and Lagakos (1989) propose a nonparametric estimator for the distribution function and apply it to doubly-censored AIDS incubation periods. It is based on a generalization of Turnbull's self-consistency algorithm and estimates simultaneously the distribution functions of both the interval-censored time origin, the infection with HIV, and the AIDS incubation period. Gómez and Lagakos (1994) deal with the same situation, but propose a two-step procedure: first, the time origin's distribution function is estimated, then, the distribution function of the AIDS incubation period. A generalization of this methodology to the case of continuous times is presented by Gómez and Calle (1999).

In contrast with that, a topic which has hardly been addressed in scientific literature, is the case of interval-censored covariates within parametric regression models. An exception is the work of Gómez, Espinal, and Lagakos (2003) who present a linear regression model with such a covariate. However, what if the response variable is a survival time, too? This situation is given in the above mentioned study on injecting drugs users in Badalona. It is the moment of HIV infection which cannot be observed exactly, but can only be diagnosed posterior by means of an HIV test. Therefore, an important part of this PhD thesis will be dedicated to estimation procedures for parametric survival models with an interval-censored covariate.

In Chapter 1 of the present work, we first give a survey on statistical methods for interval-censored data, including both parametric and nonparametric approaches. Therein, among others, the procedures of Turnbull, Finkelstein, and De Gruttola and Lagakos mentioned above shall be presented. Most of these procedures are based on the assumption that the censored data generating process is noninformative. Noninformativity means that the observed intervals do not carry any further information on the unobserved survival time; rather, this time lies within those intervals. Without this assumption, the construction of the likelihood function would have to account for the censoring process, and the distribution function of the survival time could not be identified. In Section 1.2, we address this important issue with more detail. Given the importance of optimization procedures in the further chapters of the thesis, the final section of Chapter 1 is about optimization theory. This includes some optimization algorithms, as well as the presentation of optimization tools, which have played an important role in the elaboration

of this work. We have used the mathematical programming language AMPL (Fourer, Gay, and Kernighan 2003) to solve the maximization problems which have arisen. One of its main features is that optimization problems written in the AMPL code can be sent to the internet facility 'NEOS: Server for Optimization' (Ferris, Mesnier, and Moré 2000) and be solved by its available solvers.

The data analyzed in this dissertation are presented in Chapter 2. We start presenting the data on the survival of Tuberculosis patients co-infected with HIV in Barcelona (Falqués, Langohr, Gómez, Olalla, Jansà, and Caylà 1999). The data set consists of 1135 individuals who have been registered by the Barcelona Tuberculosis Prevention and Control Program from 1988 through 1993. The second study on HIV/AIDS is the one on injecting drug users. The study cohort includes 361 injecting drug users from Badalona and surroundings who have been admitted to the detoxication unit of the *Hospital Trias i Pujol* between 1987 and 2000 (Langohr, Gómez, and Muga 2004). A completely different area to epidemiological studies are studies on the shelf life of food products. The presence of interval-censored data in a study on the shelf life of yogurt has motivated the work of Hough, Langohr, Gómez, and Curia (2003), who have introduced the methodology for interval-censored data to this area. We shall present this study in detail at the end of Chapter 2.

Chapter 3 deals with the theoretical background of an accelerated failure time model with an interval-censored covariate. An accelerated failure time model with such a covariate is presented and the corresponding likelihood function is constructed. The maximization of the resulting log likelihood function, not considered by statistical software, is complicated by the fact that the parameters of the covariate's unknown distribution function are present. For computational reasons, the covariate shall be assumed discrete. We propose an estimation procedure which includes the simultaneous maximization of the log likelihood function with respect to all parameters by means of optimization techniques. In particular, the optimization solver SNOPT (Gill, Murray, and Saunders 1999) is used to tackle the maximization problem. This solver uses a sequential quadratic programming method and is especially suitable for large nonlinearly constrained optimization problems. In order to include general cases, we consider two different situations: the response variable may be left- and right-censored or doubly-censored. The latter case is somewhat more complicated since it requires modeling the distribution of the response variable's time origin, too.

In Chapter 4, we present further regression models with interval-censored covariates and address the estimation of the corresponding parameters by means of simultaneous maximization of the likelihood function. One of these models is a linear regression model as presented by Gómez, Espinal, and Lagakos, others are logistic regression and Cox's proportional hazard model. The corresponding programmes written with the AMPL code are attached in Section C.1 in the appendix.

Besides simultaneous maximization, other estimation procedures can be applied to estimate

the parameters of the presented model. In Chapter 5, we present approaches based on imputation techniques for interval-censored data, similar to the procedure used by Pan (2000a) in the context of the proportional hazards model with an interval-censored response variable. In our case, these methods can be utilized within a two-step procedure: first, the missing value of the covariate is imputed, then, the model parameters can be estimated by standard statistical procedures available in any statistical software package. As imputed values, we can use, for example, the midpoints of the observational intervals or the expected values based on the nonparametric estimation of the covariate's distribution function.

The simultaneous maximization method has been implemented by the author in the mathematical programming language AMPL. Its application to the given data set from Badalona, the results, and their interpretation are addressed in Chapter 6. The respective programmes can be found in the appendix, in Section C. According to the obtained results, the longer an injecting drug user remains HIV-free, the longer is his or her AIDS incubation period once HIV infection occurs. These results still hold when introducing further covariates in the model. Possible reasons for this finding might be a genetically based strong immune system resisting both HIV infection and AIDS onset, or hygienic precautions taken by the individuals. However, further studies controlling for medical markers such as the CD4 count would be necessary to confirm the findings. Furthermore, longer AIDS incubation periods are observed for women and younger injecting drug users.

A simulation study has been carried out and its results are given in Chapter 7. The objective of these simulations has been to compare the results of simultaneous maximization with two procedures based on imputation for the accelerated failure time model. These are chosen since they have the advantage that the whole estimation procedure can be programmed in statistical software such as S-Plus. According to the obtained results, functions of the model's parameters, such as the median time of the response variable given different values of the covariate, are estimated with least bias by simultaneous maximization. For some of the simulation scenarios, single parameters are estimated with less bias by the imputation methods.

Finally, in the closing Chapter 8, results are summarized and several aspects which remain unsolved or might be approximated in another way are addressed. Among the unresolved aspects, one refers to methods to judge the goodness-of-fit of the adjusted model. We present a first approach to this issue, but further investigation is required to tackle it satisfactorily. Summarizing the present thesis, one of its main important aspects is the use of optimization techniques to solve optimization problems in statistics. Knowledge about these techniques can make life easier for statisticians.

Resum

L'anàlisi de supervivència tracta de l'avaluació estadística de variables que mesuren el temps fins a un esdeveniment d'interès. En l'àrea d'estudis clínics i epidemiològics, aquest esdeveniment és moltes vegades l'inici d'una malaltia, la desaparició de símptomes d'una malaltia o la mort. L'anàlisi de supervivència ha de considerar una particularitat: les dades censurades. Aquestes apareixen quan el temps d'interès no pot ser observat exactament i la informació sobre això només és parcial. Es distingeixen diferents tipus de censura: censura per la dreta és aquella quan que el temps de supervivència desconegut supera un temps observat; la censura per l'esquerra es dóna si la supervivència desconeguda és menor que un temps observat. En el cas de censura en un interval, el temps està a un interval del temps observat, i el cas de doble censura apareix quan, també, l'origen del temps de supervivència està censurat.

Dins l'àrea d'estudis clínics i epidemiològics, l'anàlisi de supervivència juga un rol substancial en els estudis sobre el VIH i la SIDA a causa dels llargs temps d'incubació i l'alta letalitat d'aquesta malaltia. S'han desenvolupat molts medicaments, d'una banda, per perllongar el temps d'incubació i la supervivència fins a la mort, i, de l'altra, per millorar la qualitat de vida de les persones infectades pel VIH. Una contribució molt valuosa per aconseguir aquests objectius ha estat el desenvolupament de les teràpies antiretrovirals altament actives (HAART) a mitjan dels anys noranta.

Aquesta tesi doctoral neix a partir de estudis sobre VIH/SIDA. El primer tracta de la supervivència de pacients de tuberculosi coinfectats amb el VIH a Barcelona i dóna una idea de la magnitud del problema de l'extensió del VIH i de la SIDA a Espanya. El temps de supervivència d'interès d'aquest estudi és el temps transcorregut des de l'inici del tractament de la tuberculosi fins a la mort dels pacients. Com que no tots els pacients van morir abans de la fi de l'estudi, hi hagut temps de supervivència censurats per la dreta. En canvi, els patrons de censura del segon estudi són més complexos. Aquest és sobre addictes de drogues intravenoses a Badalona, molts dels quals van resultar infectats pel VIH com a conseqüència de la seva drogaaddicció. L'objectiu de l'estudi ha estat examinar la possible relació entre el temps transcorregut des del primer consum de drogues intravenoses i la infecció amb VIH i després el temps de la incubació de la SIDA. Mentre el primer temps està censurat en un interval, ja que no es pot observar exactament el moment de la infecció amb VIH, el segon temps està doblement censurat: d'una banda, el seu origen, la infecció amb VIH, està censurat en un interval; de l'altra, el començament de la SIDA

ha estat parcialment censurat per la dreta i per l'esquerra, respectivament.

Els mètodes per a dades censurades en un interval han rebut molta atenció durant les darreres dècades. Aquestes dades requereixen mètodes específics, perquè procediments per a dades censurades per la dreta, com ara l'estimador de Kaplan-Meier (Kaplan and Meier 1958), no són aplicables a dades censurades en un interval. A més a més, substituir el temps de supervivència no observada pel punt mig de l'interval només proporciona resultats biaixats, especialment, si els intervals són molt amples.

L'estimació no paramètrica d'una variable aleatòria censurada en un interval ha estat tractada per molts autors. La contribució més important és la de Turnbull (1976), que introdueix la idea de la autoconsistència per a l'estimació de màxima versemblança de la funció de distribució. Respecte al model de riscos proporcionals de Cox, Finkelstein (1986) proposa un procediment per estimar els paràmetres del model si la variable de resposta inclou dades censurades en un interval. Avui en dia, tots els paquets de programari estadístic inclouen aquests mètodes.

Pel que fa a les dades doblement censurades, De Gruttola i Lagakos (1989) proposen un estimador no paramètric per a la funció de distribució i l'apliquen al temps d'incubació de la SIDA doblement censurat. Aquest es basa en una generalització de l'algorisme de autoconsistència de Turnbull i estima simultàniament les funcions de distribució tant de l'origen, la infecció amb VIH censurada en un interval, com de la incubació de la SIDA. Gómez i Lagakos (1994) tracten la mateixa situació, però proposen un procediment de dos passos: primer s'estima la funció de distribució de l'origen, després la de la incubació de la SIDA. Gómez i Calle (1999) presenten una generalització d'aquest mètode al cas de temps continus.

En canvi, un tema que ha rebut menys atenció en la literatura científica és el cas de les covariants censurades en un interval en un model de regressió paramètric. Una excepció és el treball de Gómez, Espinal i Lagakos (2003) que presenten un model de regressió lineal amb una covariant d'aquest tipus. No obstant això, com es pot procedir si la variable de resposta és també un temps de supervivència? Aquesta situació es dóna a l'estudi sobre usuaris de drogues intravenoses a Badalona mencionat anteriorment. En aquest cas, el moment de la infecció per VIH no es pot observar exactament. El que si es pot observar posteriorment, mitjançant una prova d'VIH, és el seu diagnòstic. Per aquesta raó, una part important d'aquesta tesi doctoral tracta de procediments d'estimació en models de supervivència paramètrics amb una covariant censurada en un interval.

La primera part del Capítol 1 d'aquesta tesi conté un resum de la metodologia estadística per a dades censurades en un interval, incloent-hi tant mètodes paramètrics com no paramètrics. Entre d'altres, es presentaran els mencionats procediments de Turnbull, Finkelstein, i De Gruttola i Lagakos. La majoria d'aquests mètodes suposen que la censura de les dades és no informativa. La no-informativitat significa que els intervals observats no porten més informació sobre el temps de supervivència desconegut que la mera informació que aquest temps està contingut en els intervals. Sense aquesta suposició, la funció de versemblança hauria de tenir en compte el procés

de generació de les dades censurades i la funció de distribució del temps de supervivència no podria ser identificada. Exposem més detalladament aquest aspecte a la Secció 1.2. Atesa la importància dels mètodes de l'àrea d'optimització en els capítols restants, la Secció 1.3 tracta de la teoria d'optimització. Això inclou uns quants algorismes d'optimització i la presentació d'eines d'optimització que han estat cabdals durant l'elaboració d'aquesta tesi. S'ha utilitzat el llenguatge de programació matemàtica AMPL (Fourer, Gay, and Kernighan 2003) per resoldre els problemes de maximització que han sorgit. Una dels característiques més rellevants d'AMPL és la possibilitat d'enviar problemes d'optimització, programats amb el seu codi al servidor 'NEOS: Server for Optimization' (Ferris, Mesnier, and Moré 2000) a Internet perquè siguin solucionats per aquest servidor.

En el Capítol 2, es presenten els conjunts de dades que han estat analitzats per a aquesta tesi. El primer estudi és sobre la supervivència de pacients de tuberculosi coinfectats pel VIH a Barcelona. Aquest conjunt de dades està format per 1135 individus que van ser registrats pel programa de prevenció i control de la tuberculosi a Barcelona entre els anys 1988 i 1993 (Falqués, Langohr, Gómez, Olalla, Jansà, and Caylà 1999). L'estudi següent, de l'àrea d'VIH/SIDA, és el mencionat abans sobre usuaris de drogues intravenoses. La mostra de l'estudi inclou 361 individus de Badalona i rodalia que varen ser admesos a la unitat de desintoxicació de l'Hospital Trias i Pujol entre els anys 1987 i 2000 (Langohr, Gómez, and Muga 2004). Un àrea completament diferent als estudis epidemiològics són els estudis sobre la vida útil d'aliments. La presència de censura en un interval en un estudi sobre la vida útil del iogurt ha motivat el treball de Hough, Langohr, Gómez i Curia (2003), que han introduït la metodologia per a dades censurades en un interval en aquest àrea. Aquest estudi es presenta amb detall al final del Capítol 2.

El Capítol 3 tracta el fons teòric d'un model paramètric de supervivència amb una covariant censurada en un interval. S'hi presenta un model de vida accelerada amb una covariant d'aquest tipus i s'hi desenvolupen les funcions de versemblança corresponents. La maximització de la funció de versemblança resultant, no considerada per paquets de programari estadístic, es complica pel fet que aquesta inclou els paràmetres de la funció de distribució desconeguda de la covariant. Per raons computacionals, la covariant es considera una variable aleatòria discreta. Proposem un procediment d'estimació que inclou la maximització simultània del logaritme de la funció de versemblança amb respecte a tots els paràmetres mitjançant tècniques d'optimització. En particular, s'utilitza el *solver* SNOPT (Gill, Murray, and Saunders 1999) per solucionar el problema de maximització present. Aquest *solver* utilitza un mètode de programació quadràtica seqüencial i és especialment adequat per a problemes d'optimització no lineals amb restriccions amb moltes variables. Per incloure-hi casos generals, es consideren dues situacions diferents: la variable de resposta pot ser censurada per l'esquerra i la dreta o pot estar doblement censurada. Aquest últim cas és més complicat perquè requereix també la modelització de la distribució de l'origen de la variable de resposta.

Al Capítol 4 es presenten altres models de regressió amb covariants censurades en un interval i s'aborda l'estimació dels paràmetres corresponents amb maximització simultània de la funció

de versemblança. Un d'aquests models és el de la regressió lineal presentat per Gómez, Espinal i Lagakos, els altres són la regressió logística i el model de riscos proporcionals de Cox. Els programes corresponents escrits amb el codi d'AMPL es troben a la Secció C.1 de l'apèndix.

A part de la maximització simultània, es poden aplicar altres procediments d'estimació a l'estimació dels paràmetres dels models. Al Capítol 5 presentem aproximacions basades en tècniques d'imputació per a dades censurades en un interval, semblants a les utilitzades per Pan (2000a) en el context del model de riscos proporcionals amb la variable de resposta censurada en un interval. En el nostre cas, aquests mètodes poden ser utilitzats en un procediment de dos passos: primer, s'imputa el valor desconegut de la covariant, després, es poden estimar els paràmetres del model amb procediments estadístics estàndards disponibles a qualsevol paquet de programari estadístic. Per a la imputació es poden utilitzar, per exemple, el punt mig dels intervals o el valor desitjat que es basa en l'estimació no paramètrica de la funció de distribució de la covariant.

El mètode de maximització simultània ha estat implementat per l'autor amb el codi del llenguatge de programació matemàtica AMPL. La seva aplicació al conjunt de dades de Badalona, els resultats i la seva interpretació són abordats al Capítol 6. Els programes es troben a l'apèndix, a la Secció C. Segons els resultats obtinguts, com més temps un usuari de drogues intravenoses roman seronegatiu, més llarg és el temps d'incubació de la sida una vegada que esta infectat per l'VIH. Aquests resultats continuen sent vàlids quan s'inclouen més covariants al model. Les possibles raons per a aquestes observacions podrien ser un sistema immunologic més fort resistint tant la infecció per l'VIH com el desenvolupament de la SIDA o precaucions higièniques preses per aquests individus. No obstant això, caldran més estudis que controlin indicadors mèdics com ara el nombre de cèl·lules CD4 per confirmar els resultats. A més a més, s'observen temps d'incubació de la SIDA més llargs per a dones i drogoaddictes més joves.

S'ha dut a terme un estudi de simulació els resultats del qual es presenten al Capítol 7. L'objectiu ha estat comparar la maximització simultània amb dos procediments basats en la imputació per al model de vida accelerada. S'han triat aquests mètodes per l'avantatge que tot el procediment pot ser programat només amb un paquet programari estadístic com l'S-Plus. Segons els resultats, funcions dels paràmetres del model, com la mitjana de la resposta atesos els valors diferents de la covariant, són estimades amb biaix mínim per la maximització simultània. En alguns escenaris de simulació, l'estimació dels paràmetres per mètodes d'imputació té biaix mínim.

Per acabar, a l'últim Capítol 8, es resumeixen els resultats i s'exposen diferents aspectes que no han estat resolts o que es podrien ser aproximar de manera diferent. Entre ells, un aspecte molt important és la bondat d'ajust del model ajustat. Presentem una primera aproximació, però farà falta més investigació per resoldre aquest tema de manera satisfactòria. Breument es pot dir que un dels temes cabdals d'aquesta tesi doctoral és l'ús de tècniques i eines d'optimització per resoldre problemes d'optimització sorgits en estadística. El coneixement d'aquests mètodes pot fer la vida més fàcil als estadístics.

Resumen

El análisis de supervivencia trata de la evaluación estadística de variables que miden el tiempo hasta un evento de interés. En el área de estudios clínicos y epidemiológicos, este evento es muchas veces el inicio de una enfermedad, la desaparición de los síntomas de una enfermedad o la muerte. Una particularidad, la cual ha de considerar el análisis de supervivencia, son datos censurados. Éstos aparecen cuando el tiempo de interés no puede ser observado exactamente y la información al respecto es solamente parcial. Se distinguen diferentes tipos de censura: un tiempo censurado por la derecha está presente si el tiempo de supervivencia desconocido es sabido mayor a un tiempo observado; censura por izquierda está dada si la supervivencia desconocida es menor que un tiempo observado. En el caso de censura en un intervalo, el tiempo está en un intervalo de tiempo observado, y el caso de doble censura aparece cuando, también, el origen del tiempo de supervivencia está censurado.

Dentro del área de estudios clínicos y epidemiológicos, el análisis de supervivencia juega un rol sustancial en los estudios sobre el VIH y SIDA debido a los largos periodos de incubación y la alta letalidad de esta enfermedad. Muchos medicamentos han sido desarrollados, por un lado, para prolongar tanto el tiempo de incubación como el tiempo de supervivencia hasta la muerte, y, por otro lado, para mejorar la calidad de vida de las personas infectadas por el VIH. Una contribución muy importante para lograr estos objetivos ha sido el desarrollo de las terapias anti-retrovirales altamente activas (HAART) en la mitad de los años noventa.

La presente tesis doctoral ha sido motivada principalmente por dos estudios sobre VIH/SIDA. El primero trata de la supervivencia de pacientes de tuberculosis co-infectados por el VIH en Barcelona y da una idea de la magnitud del problema que conlleva la extensión del VIH y SIDA en España. El tiempo de interés en este estudio es el tiempo trascurrido desde el inicio del tratamiento de la tuberculosis hasta la muerte de los pacientes. Como no todos los pacientes fallecieron antes del fin del estudio, ha habido tiempos de supervivencia censurados por la derecha. En cambio, los patrones de censura del segundo estudio son más complejos. Éste es sobre adictos a drogas intra-venosas en Badalona, muchos de los cuales resultaron infectados por el VIH como consecuencia de su drogadicción. El objetivo del estudio ha sido examinar la posible relación entre el tiempo trascurrido desde el primer uso de drogas intra-venosas y la infección por VIH y el siguiente tiempo de incubación del SIDA. Mientras el primer tiempo está censurado en un intervalo, ya que el momento exacto de la infección por VIH no se puede observar exactamente,

el segundo tiempo está doblemente censurado: por un lado, su origen, la infección por VIH, está censurado en un intervalo, por otro lado, el inicio del SIDA ha sido parcialmente censurado por la derecha y por la izquierda, respectivamente.

Métodos para datos censurados en un intervalo han recibido mucha atención durante las últimas décadas. Estos datos requieren métodos específicos, ya que procedimientos para datos censurados por la derecha, como por ejemplo el estimador de Kaplan-Meier (Kaplan and Meier 1958), no son aplicables a datos censurados en un intervalo. Además, sustituir el tiempo de supervivencia no-observado por el punto medio del intervalo suele proporcionar resultados sesgados, especialmente, si los intervalos son muy anchos.

La estimación no-paramétrica de una variable aleatoria censurada en un intervalo ha sido tratada por muchos autores. La contribución más importante es la por parte de Turnbull (1976), quien introduce la idea de la auto-consistencia para la estimación de máxima verosimilitud de la función de distribución. Con respecto al modelo de riesgos proporcionales de Cox, Finkelstein (1986) propone un procedimiento para estimar los parámetros del modelo si la variable de respuesta incluye datos censurados en un intervalo. Hoy en día, todos los paquetes de software estadístico incluyen estos métodos.

Respecto a datos doblemente censurados, De Gruttola y Lagakos (1989) proponen un estimador no-paramétrico para la función de distribución y lo aplican a tiempos de incubación de SIDA doblemente censurados. Éste se basa en una generalización del algoritmo de auto-consistencia de Turnbull y estima simultáneamente las funciones de distribución tanto del origen, la infección por VIH censurada en un intervalo, como de la incubación del SIDA. Gómez y Lagakos (1994) tratan la misma situación, pero proponen un procedimiento de dos pasos: primero se estima la función de distribución del origen, después la de la incubación del SIDA. Una generalización de este método al caso de tiempos continuos es presentado por Gómez y Calle (1999).

En contraste, un tema que ha recibido menos atención en la literatura científica, es el caso de covariantes censuradas en un intervalo en un modelo de regresión paramétrico. Una excepción es el trabajo de Gómez, Espinal y Lagakos (2003), quienes presentan un modelo de regresión lineal con una covariante de este índole. Sin embargo, ¿cómo proceder si la variable de respuesta es también un tiempo de supervivencia? Esta situación se da en el estudio sobre adictos a drogas intra-venosas en Badalona mencionado anteriormente. En tal caso, el momento de la infección por VIH no se puede observar exactamente. Lo que sí se puede observar posteriormente, mediante una prueba de VIH, es su diagnóstico. Por esta razón, una parte importante de la presente tesis doctoral trata de procedimientos de estimación en modelos de supervivencia paramétricos con una covariante censurada en un intervalo.

La primera parte del Capítulo 1 de ésta tesis contiene un resumen de la metodología estadística para datos censurados en un intervalo, incluyendo tanto métodos paramétricos como no-paramétricos. Entre otros, se presentarán los mencionados procedimientos de Turnbull, Finkelstein, y De Gruttola y Lagakos. La mayoría de estos métodos suponen que la censura en los datos

es no-informativa. La no-informatividad significa que los intervalos observados no llevan más información sobre el tiempo de supervivencia desconocido que la mera información que dicho tiempo está contenido en los intervalos. Sin esta suposición, la función de verosimilitud necesitaría tener en cuenta el proceso de generación de los datos censurados y la función de distribución del tiempo de supervivencia no podría ser identificada. Abordamos este aspecto importante con más detalle en la Sección 1.2. Dada la importancia de los métodos del área de optimización en los demás capítulos, la Sección 1.3 trata de la teoría de optimización. Esto incluye varios algoritmos de optimización y la presentación de herramientas de optimización que han sido de mucha importancia en la elaboración de esta tesis. Se ha utilizado el lenguaje de programación matemática AMPL (Fourer, Gay, and Kernighan 2003) para resolver los problemas de maximización que han surgido. Una de las características más importantes de AMPL es la posibilidad de enviar problemas de optimización, programados en su código al servidor 'NEOS: Server for Optimization' (Ferris, Mesnier, and Moré 2000) en internet para que sean solucionados por ese servidor.

En el Capítulo 2, se presentan los conjuntos de datos que han sido analizados para esta tesis. El primer estudio es sobre la supervivencia de pacientes de tuberculosis co-infectados por el VIH en Barcelona. Este conjunto de datos consiste de 1135 individuos que fueron registrados por el programa de prevención y control de la tuberculosis en Barcelona durante los años 1988 hasta 1993 (Falqués, Langohr, Gómez, Olalla, Jansà, and Caylà 1999). El siguiente estudio, del área de VIH/SIDA, es el mencionado sobre adictos a drogas intra-venosas. El cohorte del estudio incluye 361 usuarios de drogas intra-venosas de Badalona y alrededores que fueron admitidos a la unidad de desintoxicación del *Hospital Trias i Pujol* entre los años 1987 y 2000 (Langohr, Gómez, and Muga 2004). Un área completamente diferente a los estudios epidemiológicos son los estudios sobre la vida útil de alimentos. La presencia de censura en un intervalo en un estudio sobre la vida útil de yogur ha motivado el trabajo de Hough, Langohr, Gómez y Curia (2003), quienes han introducido la metodología para datos censurados en un intervalo en este área. Este estudio se presenta con detalle al final del Capítulo 2.

El Capítulo 3 trata del marco teórico de un modelo paramétrico de supervivencia con una covariante censurada en un intervalo. Un modelo de vida acelerada con una covariante de este índole es presentado y las funciones de verosimilitud correspondientes son desarrolladas. La maximización de la función de verosimilitud resultante, no considerada por paquetes de software estadístico, se complica por el hecho que ésta incluye los parámetros de la función de distribución desconocida de la covariante. Por razones computacionales, la covariante se considera una variable aleatoria discreta. Proponemos un procedimiento de estimación que incluye la maximización simultánea del logaritmo de la función de verosimilitud con respecto a todos los parámetros por medio de técnicas de optimización. En particular, se utiliza el *solver* SNOPT (Gill, Murray, and Saunders 1999) para solucionar el problema de maximización presente. Este *solver* usa un método de programación cuadrática secuencial y es especialmente adecuado para problemas grandes de optimización no-lineales con restricciones. Para incluir casos generales, se consideran dos situaciones diferentes: la variable de respuesta puede ser censurada por la izquierda y la

derecha o censurada doblemente. Este último caso es algo más complicado debido a que requiere también la modelización de la distribución del origen de la variable respuesta.

En el Capítulo 4 son presentados otros modelos de regresión con covariantes censuradas en un intervalo y se aborda la estimación de los parámetros correspondientes con maximización simultanea de la función de verosimilitud. Uno de estos modelos es el de la regresión lineal presentado por Gómez, Espinal y Lagakos, otros son la regresión logística y el modelo de riesgos proporcionales de Cox. Los programas correspondientes escritos en código AMPL se encuentran en la Sección C.1 en el apéndice.

Aparte de la maximización simultánea, otros procedimientos de estimación pueden ser aplicados para la estimación de parámetros de los modelos. En el Capítulo 5 presentamos aproximaciones basadas en técnicas de imputación para datos censurados en un intervalo, parecidas a las utilizadas por Pan (2000a) en el contexto del modelo de riesgos proporcionales con la variable de respuesta censurada en un intervalo. En nuestro caso, estos métodos pueden ser usados en un procedimiento de dos pasos: primero, se imputa el valor desconocido de la covariante, después, se pueden estimar los parámetros del modelo con procedimientos estadísticos estándares disponibles en cualquier paquete de software estadístico. Para la imputación se pueden usar, por ejemplo, el punto medio de los intervalos o el valor esperado basado en la estimación no-paramétrica de la función de distribución de la covariante.

El método de maximización simultanea ha sido implementado por el autor con el código del lenguaje de programación matemática AMPL. Su aplicación al conjunto de datos de Badalona, los resultados y su interpretación son abordados en el Capítulo 6. Los programas se presentan en el apéndice, en la Sección C. Según los resultados obtenidos, cuanto más tiempo permanece sero-negativo un usuario de drogas intra-venosas, más largo es el tiempo de incubación del sida una vez que ha sido infectado por el VIH. Estos resultados continúan siendo válidos al incluir más covariantes en el modelo. Posibles razones para estas observaciones podrían ser un sistema inmunológico más fuerte, resistiendo tanto la infección por el VIH como el desarrollo del sida o precauciones higiénicas tomadas por esos individuos. Sin embargo, serán necesarios más estudios controlando indicadores médicos como, por ejemplo, el número de células CD4 para confirmar los resultados. Además, se observan tiempos de incubación del sida más largos para mujeres y drogadictos más jóvenes.

Se ha llevado a cabo un estudio de simulación cuyos resultados se dan en el Capítulo 7. Ha sido el objetivo comparar la maximización simultanea con dos procedimientos basados en la imputación para el modelo de vida acelerada. Se han escogido estos dos métodos debido a la ventaja de que todo el procedimiento puede ser programado con un paquete de software estadístico como S-Plus. Según los resultados, funciones de los parámetros del modelo, como la mediana de la respuesta dados valores diferentes de la covariante, son estimadas con sesgo mínimo por la maximización simultanea. En algunos escenarios de simulación la estimación de los parámetros por métodos de imputación tiene sesgo mínimo.

Finalmente, en el último Capítulo 8, se resumen los resultados y se abordan diferentes aspectos que aún permanecen sin ser resueltos o podrían ser aproximados de manera diferente. Entre ellos, un aspecto muy importante es la bondad de ajuste del modelo usado. Nosotros presentamos una primera aproximación, pero hará falta más investigación para resolver este tema de forma satisfactoria. Uno de los aspectos más importantes que trata la presente tesis doctoral es el uso de técnicas de optimización para resolver problemas de optimización surgidos en la estadística. El conocimiento de estos métodos puede hacer la vida más fácil a los estadísticos.

Zusammenfassung

Die Überlebenszeitanalyse beschäftigt sich mit der statistischen Auswertung von Variablen, die die Zeit bis zu einem interessierenden Ereignis messen. Im Bereich klinischer und epidemiologischer Studien ist solch ein Ereignis oft der Krankheitsausbruch, das Verschwinden bestimmter Krankheitssyndrome oder der Tod. Eine Besonderheit, welche die Überlebenszeitanalyse berücksichtigt, sind so genannte zensierte Daten. Diese tauchen dann auf, wenn die interessierende Überlebenszeit nicht genau beobachtet kann, jedoch teilweise Information hierüber vorliegt. Es werden verschieden Arten zensierter Daten unterschieden: Rechtszensierung liegt vor, falls die Überlebenszeit einen beobachteten Wert überschreitet, selbst aber nicht beobachtet werden kann; im Falle von Linkszensierung ist die Überlebenszeit geringer als die beobachtete Zeit. Intervallzensierung beschreibt den Fall, dass die unbekannte Überlebenszeit innerhalb eines beobachteten Zeitintervalls liegt, und doppelte Zensierung oder Doppelzensierung liegt vor, wenn auch der Ursprung der Überlebenszeit zensiert ist.

Auf dem Gebiet klinischer und epidemiologischer Studien spielt die Überlebenszeitanalyse eine sehr wichtige Rolle bei Studien über HIV und AIDS. Dies liegt vornehmlich an den langen Inkubationszeiten und der hohen Mortalität dieser Krankheit. Viele Medikamente sind entwickelt worden, um einerseits die AIDS Inkubationszeiten und die Überlebenszeit nach Ausbruch der Krankheit zu verlängern, und um anderseits die Lebensqualität HIV-infizierter Personen zu verbessern. Ein sehr wichtiger Beitrag hierzu war die Entwicklung der antiretroviralen Kombinationstherapien (HAART) Mitte der 90er Jahre.

Die vorliegende Dissertation ist vornehmlich durch zwei HIV/AIDS-Studien motiviert. Die erste beschäftigt sich mit den Überlebenszeiten HIV-infizierter Tuberkulose Patienten in Barcelona und vermittelt eine Idee des Ausmaßes des HIV/AIDS-Problems in Spanien. Die interessierende Überlebenszeit dieser Studie ist die Zeit zwischen Beginn der Tuberkulose-Behandlung und dem Tod der Patienten. Da bei Studienende nicht alle Patienten verstorben waren, liegen rechtszensierte Daten vor. Im Gegensatz dazu sind die beobachteten Zensierungsmuster der zweiten Studie komplexer, die von intravenös Drogen gebrauchenden Personen in Badalona (Spanien) handelt, von denen sich viele aufgrund ihrer Drogenabhängigkeit mit dem HI-Virus infiziert haben. Das Studienziel bestand darin, den möglichen Zusammenhang der Zeiten zwischen erstem intravenösen Drogengebrauch und der HIV-Infektion einerseits und der nachfolgenden AIDS-Inkubationszeit andererseits zu untersuchen. Während die erste der beiden Zeiten intervall-

zensiert ist, da der Moment der HIV-Infektion nicht genau beobachtet werden kann, ist die zweite doppelzensiert, eben weil die HIV-Infektion intervallzensiert ist und der AIDS-Beginn teilweise rechts- und linkszensiert ist.

Statistische Methoden für intervallzensierte Daten haben in den letzten Jahrzehnten viel Beachtung gefunden. Die Auswertung dieser Daten benötigt spezifische Methoden, da sich Methoden für rechtszensierte Daten, wie zum Beispiel der Kaplan-Meier-Schätzer (Kaplan and Meier 1958), nicht ohne weiteres auf sie anwenden lassen. Genauso wenig ist es ratsam, die nicht beobachtete Überlebenszeit durch den Intervallsmittelpunkt zu ersetzen, da dies in der Regel zu verzerrten Ergebnissen führt, vor allem dann, wenn die Intervalle sehr breit sind.

Mehrere Autoren haben sich mit der Schätzung der Verteilungsfunktion einer intervallzensierten Zufallsvariable beschäftigt. Der wichtigste Beitrag hierzu stammt von Turnbull (1976), der die Idee der Autokonsistenz verwendet, um den Maximum Likelihood Schätzer der Verteilungsfunktion zu bestimmen. Für den Fall einer intervallzensierten Variable innerhalb des Cox'schen proportionalen Hazard-Modells schlägt Finkelstein (1986) eine Schätzprozedur für die Modellparameter vor. Heutzutage sind diese Methoden in allen wichtigen statistischen Software Paketen enthalten.

Für doppelzensierte Daten schlagen De Gruttola und Lagakos (1989) einen nichtparametrischen Schätzer für die Verteilungsfunktion vor und wenden diesen auf doppelzensierte AIDS-Inkubationszeiten an. Er basiert auf einer Verallgemeinerung von Turnbulls Autokonsistenz-Algorithmus und schätzt gleichzeitig die Verteilungsfunktionen des intervallzensierten Zeitpunktes der HIV-Infektion und die der AIDS-Inkubationszeit. Gómez und Lagakos (1994) beschäftigen sich mit derselben Situation, schlagen allerdings eine zweistufige Schätzprozedur vor: in einem ersten Schritt wird die Verteilungsfunktion des Ursprungs der Inkubationszeit geschätzt, anschließend die der AIDS Inkubationszeit. Eine Verallgemeinerung dieser Methode auf den Fall stetig verteilter Zeiten stellen Gómez und Calle (1999) vor.

Im Gegensatz dazu haben Regressionsmodelle mit intervallzensierten Kovariablen bisher nur wenig Beachtung in der wissenschaftlichen Literatur gefunden. Eine der Ausnahmen ist die Arbeit von Gómez, Espinal und Lagakos (2003), die ein lineares Regressionsmodell mit solch einer Regressorvariablen vorstellen. Jedoch, wie hat man vorzugehen, wenn auch die Zielvariable eine Überlebenszeit ist? Dies ist die Situation in der oben genannten Studie über intravenös Drogen gebrauchende Personen in Badalona. Der Moment der HIV-Infektion kann nicht genau beobachtet werden, sondern nur nachträglich mit Hilfe eines HIV-Tests nachgewiesen werden. Aus diesem Grund beschäftigt sich ein Großteil dieser Dissertation mit Schätzprozeduren für parametrische Überlebenszeitmodelle mit einer intervallzensierten Kovariablen.

Der erste Abschnitt von Kapitel 1 der vorliegenden Arbeit beinhaltet einen Überblick über statistische Methoden für intervallzensierte Daten, der sowohl parametrische als auch nichtparametrische Verfahren umfasst. Darin werden unter anderem die Verfahren von Turnbull, Finkelstein sowie De Gruttola und Lagakos vorgestellt. Der Großteil dieser Methoden basiert auf

der Annahme, dass die zensierten Daten nichtinformativ sind, das heißt dass die beobachteten Intervalle keine weitere Information über die unbeobachtete Überlebenszeit beinhalten. Ohne diese Annahme müsste die Likelihoodfunktion die Verteilungsfunktion der Zensierungszeiten mit einbeziehen, womit die Verteilung der Überlebenszeiten nicht identifiziert werden könnte. Dieser wichtige Punkt wird in Abschnitt 1.2 näher beleuchtet. Angesichts der Relevanz von Optimierungsverfahren in den weiteren Kapiteln der Dissertation, beschäftigt sich der abschließende Abschnitt von Kapitel 1 mit diesem Thema. Dies umfasst sowohl Algorithmen von Optimierungsverfahren als auch Software, die bei der Ausarbeitung der vorliegenden Arbeit eine wichtige Rolle gespielt haben. Für die zu lösenden Maximierungsprobleme ist die mathematische Programmiersprache AMPL (Fourer, Gay, and Kernighan 2003) verwendet worden. Ein großer Vorteil dieser Programmiersprache ist die Möglichkeit, dass mit AMPL programmierte Optimierungsprobleme von den *Solvern* des Internetservers 'NEOS: Server for Optimization' (Ferris, Mesnier, and Moré 2000) gelöst werden können.

In Kapitel 2 werden die Datensätze, die für diese Dissertation ausgewertet worden sind, vorgestellt. Dabei handelt es sich zunächst um die Daten der Studie über die Überlebenszeiten von HIV-infizierten Tuberkulose-Patienten in Barcelona (Falqués, Langohr, Gómez, Olalla, Jansà, and Caylà 1999). Der Datensatz umfasst die Daten von 1135 Patienten, die vom Programm zur Vorbeugung und Kontrolle der Tuberkulose in Barcelona zwischen 1988 und 1993 erfasst wurden. Die zweite Studie über HIV/AIDS ist die zuvor erwähnte über intravenös Drogen gebrauchende Personen. Der Studienkohorte besteht aus 361 Drogenabhängigen aus Badalona und Umgebung, die zwischen 1987 und 2000 auf der Entzugsstation des Krankenhauses *Hospital Trias i Pujol* aufgenommen wurden (Langohr, Gómez, and Muga 2004). Ein von epidemiologischen Studien sehr unterschiedliches Gebiet sind Studien über die sensorische Haltbarkeit von Lebensmitteln. Das Auftreten intervallzensierter Daten im Rahmen einer Studie über die Haltbarkeit von Jogurt hat die Arbeit von Hough, Langohr, Gómez und Curia (2003) motiviert, welche die Methodik für intervallzensierte Daten auf diesem Gebiet eingeführt haben. Diese Studie wird im letzten Abschnitt von Kapitel 2 vorgestellt.

Kapitel 3 beschäftigt sich mit dem theoretischen Hintergrund eines loglinearen Überlebenszeit-Modells (*accelerated failure time model*) mit einer intervallzensierten Kovariablen. Zu Beginn wird dieses Modell vorgestellt und die zugehörige Likelihoodfunktion wird hergeleitet. Die Maximierung der resultierenden Loglikelihoodfunktion kann nicht mit herkömmlicher statistischer Software ausgeführt werden und wird dadurch erschwert, dass diese auch bezüglich der Parameter der unbekannten Verteilungsfunktion der Kovariablen durchzuführen ist. Aufgrund rechentechnischer Details nehmen wir an, dass die Regressorvariable eine diskrete Zufallsvariable ist. Für die Maximierung schlagen wir eine Schätzprozedur vor, welche die Loglikelihoodfunktion mit Hilfe von Optimierungsverfahren gleichzeitig bezüglich aller Parameter maximiert. Insbesondere verwenden wir den Optimierungs-Solver SNOPT (Gill, Murray, and Saunders 1999), der eine Methode sequentieller quadratischer Programmierung verwendet und speziell für komplexe nichtlineare Optimierungsprobleme geeignet ist. Zwei unterschiedliche Zensierungsmuster der

Zielvariablen werden betrachtet: diese mag einerseits rechts- und linkszensiert sein oder doppelzensiert. Der letzte Fall ist etwas schwieriger zu modellieren, da dabei die Verteilung des Ursprungs der Zielvariablen ebenfalls modelliert werden muss.

Im darauf folgenden Kapitel 4 werden weitere Regressionsmodelle mit einer intervallzensierten Kovariablen vorgestellt, und die Schätzung der dazu gehörigen Parameter anhand von simultaner Maximierung der Likelihoodfunktion wird behandelt. Eines dieser Modelle ist das von Gómez, Espinal und Lagakos vorgestellte lineare Regressionsmodell, weitere Modelle sind die logistische Regression sowie das proportionale Risikomodell von Cox.

Neben simultaner Maximierung können auch andere Methoden für die Schätzung der Modellparameter angewandt werden. So zum Beispiel die im Kapitel 5 angesprochenen Schätzprozeduren, die Imputationstechniken für intervallzensierten Daten verwenden. Diese Methoden ähneln den von Pan (2000a) im Rahmen des proportionalen Risikomodells mit intervallzensierten Daten verwendeten Verfahren. In unserem Fall bilden sie den ersten Schritt einer zweistufigen Prozedur: zunächst wird der nicht beobachtete Wert der Kovariablen ersetzt, anschließend können herkömmliche Schätzmethoden, die von allen statistischen Software-Paketen angeboten werden, verwendet werden. Für den Imputationsschritt kommen beispielsweise der Intervallsmittelpunkt in Frage, oder der bedingte auf der nichtparametrischen Schätzung der Verteilungsfunktion der Kovariablen basierende Erwartungswert.

Die Methode der simultanen Maximierung ist vom Autor in AMPL programmiert und auf den Datensatz der Studie in Badalona angewandt worden. Die Ergebnisse und zugehörige Schlussfolgerungen finden sich im Kapitel 6, die dazugehörigen Programme im Abschnitt C. Die Ergebnisse deuten an, dass je länger intravenös Drogen gebrauchende Personen von einer HIV-Infektion verschont bleiben, desto länger ist, im Durchschnitt, die AIDS-Inkubationszeit. Dies wurde auch beobachtet als weitere Kovariablen im Modell berücksichtigt wurden. Mögliche Gründe hierfür könnten ein gutes Immunsystem sein, das sowohl die HIV-Infektion als auch den Ausbruch von AIDS hinauszögert, oder auch hygienische Vorsorge von Seiten der Betroffenen. Ebenso wurden längere AIDS-Inkubationszeiten bei Frauen und jüngeren Patienten beobachtet. Dennoch müssen diese Ergebnisse mit Vorsicht betrachtet werden, da medizinische Marker wie die Anzahl an T4-Helferzellen im Blut nicht berücksichtigt worden sind. Dies wäre in weiter führenden Studien nachzuholen.

In Kapitel 7 fassen wir die Ergebnisse einer Simulationsstudie zusammen. Diese wurde mit dem Ziel durchgeführt, die Methoden der simultanen Maximierung der Likelihoodfunktion einerseits und den auf Imputation basierenden Prozeduren andererseits bezüglich der Präzision der Schätzergebnisse zu vergleichen. Das verwendete Modell ist das loglineare Überlebenszeit-Modell aus Kapitel 3. Den Ergebnissen zufolge, werden Funktionen der Modellparameter in der Regel am genauesten mittels simultaner Maximierung geschätzt. Für einige der ausgewählten Simulationsszenarien hingegen wurden einzelne Parameter besser von den Imputationsmethoden geschätzt.

Im abschließenden Kapitel 8 findet sich eine Zusammenfassung der wichtigsten Ergebnisse.

Ebenso werden verschiedene Aspekte der Arbeit angesprochen, zum Beispiel solche, die bislang nur unbefriedigend gelöst worden sind. Ein wichtiges Thema hierbei sind Kriterien, um die Güte der Modellanpassung zu überprüfen. Ein erster Vorschlag hierfür wird in Kapitel 6 vorgestellt, weitere Arbeit ist allerdings notwendig, um die Modellgüte zufriedenstellend beurteilen zu können. Zusammenfassend kann gesagt werden, dass einer der wichtigsten Aspekte dieser Dissertation darin besteht, Techniken aus dem Gebiet der Optimierungverfahren auf Optimierungsprobleme in der Statistik angewandt zu haben. Die Handhabung dieser Methoden kann Statistikern das Leben einfacher machen.

Contents

List of Tables

List of Figures