

الحمد لله
الرحمن الرحيم



دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان نامه کارشناسی ارشد

مدل رگرسیونی نرخ شفایافتگی آماسیده در صفر و کاربرد آن در تسهیلات بانکی

نگارنده: مونا عابدی مهر

استاد راهنما

دکتر نگار اقبال

استاد مشاور

دکتر حسین باغیثنی

بهمن ۱۳۹۶

شماره:

تاریخ:

باسمه تعالی



مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم مونا عابدی مهر با شماره دانشجویی ۹۴۱۱۶۸۴ رشته آمار گرایش آمار ریاضی تحت عنوان مدل رگرسیونی نرخ شفایافتگی آماسیده در صفر و کاربرد آن در تسهیلات بانکی که در تاریخ ۹۶/۱۱/۱ با حضور هیات محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

☐ مردود	☒ قبول (با درجه: خیلی خوب)
☐ عملی	☐ نظری
نوع تحقیق:	

عضو هیات داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنمای اول	دکتر نگار اقبال	استادیار	
۲- استاد راهنمای دوم	—	—	—
۳- استاد مشاور	دکتر حسین باغیشتی	استادیار	
۴- نماینده تحصیلات تکمیلی	دکتر محمد رضا ربیعی	استادیار	
۵- استاد ممتحن اول	دکتر احمد تراکئی	استادیار	
۶- استاد ممتحن دوم	دکتر داور شاهسوئی	استادیار	



نام و نام خانوادگی رئیس دانشکده: ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند مجدداً در این دانشگاه درخواست دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به:

مقدس‌ترین واژه‌ها در لغت‌نامه دلم
مادر مهربانم که زندگیم را مدیون مهر و
عطوفت آن می‌دانم.

پدر، مهربانی مشفق، بردبار و حامی.
و آرامشم نوید که نشانه لطف الهی در
زندگی من است.

سپاس‌گزاری...

شکر شایان نثار ایزد منان که توفیق را رفیق راهم ساخت تا این پایان‌نامه را به پایان برسانم.

سپاس‌گزار کسانی هستم که سراغاز تولد من هستند. از یکی زاده می‌شوم و از دیگری جاودانه. استادی که سپیدی را بر تخته سیاه زندگیم نگاشت و مادر و پدری که تار مویی از آن‌ها به پای من سیاه نماند. از اساتید گرامی سرکار خانم دکتر نگار اقبال و جناب آقای دکتر حسین باغیشنی بسیار سپاس‌گزارم چرا که بدون راهنمایی ایشان تامین این پایان‌نامه بسیار مشکل می‌نمود.

هم‌چنین برخود لازم می‌دانم از اساتید داور، جناب آقای دکتر احمد نزاکتی و جناب آقای دکتر داود شاهسونی و تمام اساتید گروه آمار دانشگاه صنعتی شاهرود دکتر آرشی و دکتر ربیعی تشکر و قدردانی نمایم.

وظیفه خود می‌دانم که از پدر و مادر مهربانم که در سختی‌ها و دشواری‌های زندگی همواره یاری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند صمیمانه سپاس‌گزاری کنم. از همسر عزیزم که قلبی آکنده از عشق و معرفت و محیطی سرشار از سلامت و امنیت و آرامش و آسایش برای من فراهم آورده است بی‌نهایت سپاس‌گزارم. تشکر ویژه دارم از مسئولین بانک مرکزی ایران به خصوص جناب آقای دکتر جلیلی، دکتر اکبری و دکتر ملک‌شعار که در به‌دست آوردن داده‌ها با اینجانب نهایت همکاری را به‌جا آوردند.

هم‌چنین از پروفسور مایرو اولوریا که کدهای شبیه‌سازی این پایان‌نامه را در اختیار من قرار دادند و در زمان تحقیق بسیار مرا راهنمایی نمودند نهایت تشکر و قدردانی را دارم.

با سپاس از خواهر عزیزم که همراه همیشگی من بوده و دوستان و همکلاسی‌های گران‌مایه‌ام که مرا صمیمانه یاری داده‌اند.

مونا عابدی‌مهر

بهمن ۱۳۹۶

تعهد نامه

این جانب **مونا عابدی مهر** دانشجوی کارشناسی ارشد رشته آمار دانشکده علوم ریاضی دانشگاه صنعتی شاهرود، نویسنده پایان نامه با عنوان **مدل رگرسیونی نرخ شفایافتگی آماسیده در صفر و کاربرد آن در تسهیلات بانکی**، تحت راهنمایی دکتر **نگار اقبال** متعهد می شوم:

- تحقیقات در این پایان نامه توسط این جانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش های دیگر پژوهش گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام “دانشگاه صنعتی شاهرود” یا “Shahrood University of Technology” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آوردن نتایج اصلی پایان نامه تاثیرگذار بوده اند، در مقالات مستخرج از پایان نامه رعایت می گردد.
- در تمام مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافت های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

مونا عابدی مهر

بهمن ۱۳۹۶

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود است. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان نامه بدون ذکر منبع مجاز نیست.

چکیده

بانک‌ها و موسسات مالی همواره، تحت شرایطی، به مشتریان خود وام‌هایی را اعطاء می‌کنند. بر اساس عملکرد مشتریان در بازپس دادن وام‌ها، می‌توان آن‌ها را به سه دسته تقسیم کرد. دسته اول مشتریانی هستند که به‌طور کامل ارتباطشان با بانک قطع می‌شود؛ به این افراد متقلب می‌گویند که زمان بقای پرداخت وام آن‌ها برابر صفر است. دسته دوم مشتریانی هستند که به بازپرداخت وام خود اهمیتی نمی‌دهند اما برخلاف متقلبین، این بازپرداخت تا مدتی ادامه پیدا می‌کند و سپس قطع می‌گردد. این دسته از مشتریان زمان بقای بازپرداخت وام مثبت دارند ولی در نهایت ممکن است به متقلب تبدیل شوند. به این نوع مشتریان، قصورکننده می‌گویند. دسته سوم نیز مشتریانی هستند که تمام اقساطشان را به موقع پرداخت کنند.

برای بیشینه‌سازی سود یک بانک، باید نرخ‌های متقلبین و قصورکنندگان آن به کم‌ترین مقدار برسد و مشتریان خوش حساب حداکثر شوند. بنابراین تحلیل طول زمان بقای بازپرداخت وام‌ها در هر بانک می‌تواند حیاتی باشد. تحلیل داده‌های طول زمان تا رخداد یک پدیده (که در این‌جا پدیده مورد نظر بازپرداخت وام است)، در حیطه تحلیل بقا قرار می‌گیرد. در تحلیل داده‌های بقا، مسأله اصلی یافتن یک مدل مناسب برای سنجش تاثیر عوامل مختلف بر طول زمان بقا است. با توجه به ماهیت داده‌های مشتریان بانک‌ها، بقای عدم بازپرداخت وام مشتریان متقلب و قصورکننده به صفر میل نمی‌کند. رده مناسب برای این نوع داده‌های بقا، مدل‌های نرخ شفایافتگی هستند. برای این مدل‌ها دو زیررده کلی مدل‌های آمیخته و غیرآمیخته معرفی شده‌اند. مشکلی که بر سر راه استفاده از مدل‌های نرخ شفایافتگی برای تحلیل رفتار وام‌گیرندگان وجود دارد، وجود دسته اول یعنی متقلبین است که بقای عدم بازپرداخت برابر صفر دارند. این مسأله باعث برآزش نامطلوب مدل نرخ شفایافتگی می‌شود. در چنین مواردی به کارگیری مدل‌های آماسیده در صفر توصیه می‌شود. در این پایان‌نامه، ابتدا مدل نرخ شفایافتگی آماسیده در صفر را معرفی می‌کنیم و سپس از آن برای تحلیل داده‌های بازپرداخت برخی از مشتریان چهار نوع بانک ثبت‌شده در بانک مرکزی ایران، استفاده می‌کنیم.

کلمات کلیدی: نرخ شفایافتگی، تحلیل بقا، آماسیده در صفر، تسهیلات بانکی

لیست مقالات مستخرج از پایان نامه

۱. عابدی مهر م.، اقبال ن. و باغیشنی ح.، (۱۳۹۶)، ”تحلیل تقلب در تسهیلات بانکی با استفاده از مدل های بقا“، یازدهمین سمینار احتمال و فرایندهای تصادفی، صفحه ۷۸-۸۶، قزوین، ایران.

فهرست مطالب

فهرست تصاویر

ف

فهرست جداول

ق

۱	مفاهیم و تعاریف اولیه	۱
۱	۱.۱ رگرسیون خطی	۱.۱
۲	۱.۱.۱ روش درست‌نمایی ماکسیمم	۱.۱.۱
۳	۲.۱ روش نیوتون-رافسون	۲.۱
۴	۱.۲.۱ روش BFGS	۱.۲.۱
۵	۳.۱ رگرسیون خطی تعمیم‌یافته	۳.۱
۶	۱.۳.۱ رگرسیون لجستیک	۱.۳.۱
۸	۲.۳.۱ توزیع وایبل	۲.۳.۱
۸	۴.۱ تحلیل بقا	۴.۱
۱۱	۱.۴.۱ مقدمه‌ای بر مفاهیم بقا	۱.۴.۱
۱۴	۲.۴.۱ برآورد تابع بقا	۲.۴.۱
۱۶	۳.۴.۱ داده‌های سانسور شده	۳.۴.۱
۱۷	۴.۴.۱ انواع برش‌ها	۴.۴.۱
۲۰	۵.۱ مدل‌های بقا	۵.۱
۲۰	۱.۵.۱ مدل‌های پارامتری بقا	۱.۵.۱
۲۱	۲ مدل‌های نرخ شفایافتگی	۲
۲۱	۱.۲ مقدمه	۱.۲
۲۲	۲.۲ تاریخچه	۲.۲
۲۳	۳.۲ مدل نرخ شفایافتگی	۳.۲
۲۴	۱.۳.۲ مدل نرخ شفایافتگی آمیخته	۱.۳.۲
۲۵	۲.۳.۲ مدل نرخ شفایافتگی غیرآمیخته	۲.۳.۲

۲۷	۴.۲	تابع درست‌نمایی مدل‌های نرخ شفایافتگی
۲۷	۱.۴.۲	تابع درست‌نمایی مدل آمیخته
۲۸	۲.۴.۲	تابع درست‌نمایی مدل غیرآمیخته
۳۱	۳	مدل نرخ شفایافتگی آماسیده در صفر برای بقای بازپرداخت وام‌ها
۳۱	۱.۳	مقدمه
۳۳	۲.۳	مدل نرخ شفایافتگی آماسیده در صفر
۳۴	۳.۳	مدل رگرسیونی پیشنهادی
۳۶	۱.۳.۳	مدل پارامتری برای تابع بقای مشتریان خوش‌حساب
۳۷	۴.۳	برازش مدل و استنباط آماری
۴۱	۴	تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی
۴۱	۱.۴	مطالعه شبیه‌سازی
۴۳	۱.۱.۴	الگوریتم شبیه‌سازی
۴۴	۲.۱.۴	نتایج شبیه‌سازی مونت کارلویی
۵۴	۲.۴	مطالعه موردی: تسهیلات بانک مرکزی ایران
۵۷	۱.۲.۴	تحلیل توصیفی داده‌ها
۶۴	۳.۴	نتیجه‌گیری و آینده تحقیق
۶۷		مراجع
۷۱	آ	دستورات نرم‌افزار R
۸۹		واژه‌نامه فارسی به انگلیسی
۹۱		واژه‌نامه انگلیسی به فارسی
۹۳		نمایه

فهرست تصاویر

۱۱	مثال‌هایی از انواع سانسور	۱.۱
۱۹	برش تصادفی برای یک بیمار	۲.۱
۳۲	طول زمان بازپرداخت تسهیلات بانکی	۱.۳
۳۵	تابع بقا مدل نرخ شفایافتگی آماسیده در صفر	۲.۳
	اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی اول انتخاب پارامترها برای ۱۰۰ بار تکرار.	۱.۴
۴۶	اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی اول انتخاب پارامترها	۲.۴
۴۷	برآوردهای ماکسیمم درست‌نمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی اول. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد	۳.۴
۴۸	اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی دوم انتخاب پارامترها برای ۱۰۰ بار تکرار.	۴.۴
۴۹	اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی دوم انتخاب پارامترها	۵.۴
۵۰	برآوردهای ماکسیمم درست‌نمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی دوم. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد	۶.۴
۵۱		

۷.۴	ارایی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درستنمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی شده در سناریوی سوم انتخاب پارامترها برای ۱۰۰ بار تکرار	۵۲
۸.۴	ارایی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درستنمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی شده در سناریوی سوم انتخاب پارامترها	۵۳
۹.۴	برآوردهای ماکسیمم درستنمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی سوم. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد	۵۴
۱۰.۴	تابع بقای برآورد شده به روش کاپلان-مایر به تفکیک گروه‌های جنسیتی	۵۸
۱۱.۴	تابع بقای برآورد شده به روش کاپلان-مایر به تفکیک نوع بانک	۶۰
۱۲.۴	هیستوگرام طول زمان بازپرداخت فقط مشتریان متقلب و قصورکننده	۶۲
۱۳.۴	برآورد توابع بقای وایبل برای مدل نرخ شفایافتگی به تفکیک نوع بانک	۶۴

فهرست جداول

۱۰	نمونه‌ای از داده‌های بقا و سانسورها	۱.۱
۵۸	خلاصه‌ای از داده‌های طول عمر وام بانکی با توجه به جنسیت مشتریان	۱.۴
۵۹	خلاصه‌ای از داده‌های طول عمر وام بانکی بر اساس نوع بانک	۲.۴
۶۱	خلاصه‌ای از داده‌های طول عمر وام بانکی بر اساس نوع وام	۳.۴
۶۱	متغیرهای ظاهری برای انواع بانک‌ها	۴.۴
۶۳	برآورد پارامترهای مدل برای داده‌های تسهیلات بانک مرکزی	۵.۴

فصل ۱

مفاهیم و تعاریف اولیه

در این فصل مفاهیم پایه‌ای و اساسی که در فصل‌های آینده به آن‌ها نیاز داریم، شرح می‌دهیم. ابتدا مدل رگرسیون خطی و روش‌های برآورد پارامتر را بیان می‌کنیم. سپس روش بهینه‌سازی عددی نیوتون-رافسون^۱ را معرفی می‌کنیم. در ادامه، رگرسیون لجستیک که عضوی از رده مدل‌های خطی تعمیم‌یافته^۲ (GLMs) است را تعریف کرده و در پایان مفاهیمی مربوط به تحلیل بقا، مدل‌های بقا و انواع سانسورها و برش‌ها را شرح می‌دهیم.

۱.۱ رگرسیون خطی

ساده‌ترین نوع رگرسیون، رگرسیون خطی است که با توجه به کتاب نیرومند [۶] معادله آن در حالت خاص تنها یک متغیر تبیینی، x ، به صورت

$$y = \beta_0 + \beta_1 x + \varepsilon$$

است که برای بررسی وجود یک رابطه خطی بین متغیر پاسخ y و متغیر تبیینی به کار می‌رود. در این مدل ضرایب رگرسیونی β_0 و β_1 پارامترهای نامعلوم مدل هستند و ε جمله خطای تصادفی است. در صورتی که بیش از یک متغیر تبیینی رگرسیونی، مثلاً k تا، وجود داشته باشند،

^۱Newton-Raphson

^۲Generalized Linear Models

رگرسیون خطی را رگرسیون خطی چندگانه نامیده و به صورت زیر نمایش می‌دهند:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon. \quad (1.1)$$

به عنوان مثال، رابطه بین درآمد یک مشتری با نحوه پرداخت اقساط وام برای همان شخص یک رابطه مستقیم کامل نیست؛ به این معنی که شخصی با درآمد بالاتر لزوماً مبلغ اقساط خود را نسبت به افراد با درآمد پایین‌تر منظم‌تر، در تاریخ‌های تعیین شده، پرداخت نمی‌کند. اما انتظار داریم، با احتمال قابل ملاحظه‌ای، شخصی که درآمد بالاتری دارد تمایل بیشتری به پرداخت اقساط به موقع داشته باشد. در این مثال، پیش‌گویی پرداخت یا عدم پرداخت قسط از روی درآمد شخص و بیان ارتباط بین این دو متغیر با یک مدل رگرسیون خطی امکان‌پذیر است.

یک روش معمول برای برآورد پارامترهای یک مدل رگرسیون (خطی)، روش برآورد درست‌نمایی ماکسیمم^۳ (MLE) است. با توجه به استفاده از این روش برای برآورد پارامترهای مدل‌های مورد علاقه در این پایان‌نامه، در ادامه، ابتدا این روش را به اختصار معرفی می‌کنیم.

۱.۱.۱ روش درست‌نمایی ماکسیمم

روش درست‌نمایی ماکسیمم یکی از قدیمی‌ترین و پراهمیت‌ترین روش‌ها در نظریه برآوردیابی است (پارسیان [۳]). برآورد ML که متداول‌ترین روش در بین استفاده‌کنندگان کاربردی از علم آمار است، اولین بار توسط گاوس [۱۴] به کار گرفته شد و پس از آن به صورت گسترده‌ای توسط فیشر [۱۲] در انتقاد از روش برآوردیابی گشتاوری مورد استفاده قرار گرفت.

فرض کنید بردار متغیرهای تصادفی $X = (X_1, X_2, \dots, X_n)'$ دارای تابع (چگالی) احتمال توأم $f_\theta(x)$ باشد، که در آن مقدار واقعی پارامتر θ عضوی از فضای پارامتر Θ است. یافته‌های این نمونه تصادفی را با $x = (x_1, x_2, \dots, x_n)'$ نشان می‌دهیم. برای هر مقدار داده شده $X = x$ ، تابع درست‌نمایی X را تابع چگالی احتمال توأم (تابع احتمال توأم) X ، یعنی $f_\theta(x)$ ، تعریف می‌کنیم که به صورت تابعی از θ در نظر گرفته می‌شود و آن را با نماد $L(\theta)$ نمایش می‌دهیم؛ به عبارت دیگر

$$L(\theta) = f_\theta(x).$$

ذکر چند نکته در مورد تابع درست‌نمایی ضروری است:

۱. تابع درست‌نمایی $L(\theta)$ لزوماً نسبت به θ مشتق‌پذیر نیست.
۲. تابع درست‌نمایی $L(\theta)$ یک تابع (چگالی) احتمال نیست.
۳. اگر $X_1, \dots, X_n$ یک نمونه تصادفی n تایی از توزیعی از خانواده توابع (چگالی) احتمال

$$\{f_\theta(x) : \theta \in \Theta\}$$

^۳Maximum Likelihood Estimation

باشد، آن گاه

$$L(\theta) = \prod_{i=1}^n f_{\theta}(x_i).$$

برآوردگر ML پارامتر θ نقطه‌ای از فضای پارامتر θ است که تابع درستنمایی به ازای آن ماکسیمم شود؛ به عبارت دیگر

$$\hat{\theta} = \arg \max_{\theta} L(\theta)$$

که در آن منظور از $\hat{\theta}$ همان برآوردگر ML است. خارج از مدل‌های کلاسیک و ساده، برآوردهای ML پارامترها صورت بسته ندارند و نمی‌توان آن‌ها را به‌طور دقیق به‌دست آورد. در این موارد، باید از روش‌های تقریبی برای محاسبه MLE استفاده کرد. برای این کار، روش‌های بهینه‌سازی عددی مختلفی در علوم ریاضی معرفی شده‌اند. یکی از روش‌های معمول بهینه‌سازی عددی، روش نیوتون-رافسون است که در ادامه به‌طور خلاصه آن را معرفی می‌کنیم.

۲.۱ روش نیوتون-رافسون

روش نیوتون که بیش‌تر با نام روش نیوتون-رافسون شناخته می‌شود، روشی عددی برای تعیین ریشه یک تابع است. مبنای این روش استفاده از خط مماس بر منحنی برای یافتن حدس بهبودیافته بعدی و نزدیک‌تر شدن به مقدار ریشه تابع است. دقت داشته باشید که MLE ریشه مشتق تابع درستنمایی است. برای محاسبه شیب منحنی از مشتق منحنی در آن نقطه استفاده می‌شود.

فرض کنید بخواهیم ریشه تابع $f(x)$ را بیابیم. بنابراین باید معادله

$$f(x) = 0$$

را حل کنیم. در روش نیوتون-رافسون، ابتدا نقطه آغازین x_0 را به‌عنوان حدس اولیه انتخاب می‌کنیم و از آن توسط معادله به‌روزکننده نیوتون-رافسون (رابطه (۲.۱)) برای رسیدن به مقدار جدید x_1 استفاده می‌کنیم:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \quad (2.1)$$

با جایگزین کردن x_1 به‌جای x_0 و ادامه همین مراحل، معادله به‌روزکننده در مرحله $n+1$ به صورت زیر حاصل می‌شود:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

هرچه تعداد دفعات تکرار بیش‌تر باشد، مقدار به‌دست آمده به ریشه نزدیک‌تر است. این روش برای توابع چندمتغیره نیز قابل استفاده است و معادله به‌روزکننده یک دستگاه معادلات (بردار) خواهد بود.

این روش تکراری به مقدار حدس اولیه بسیار وابسته است و ممکن است همیشه به جواب واقعی همگرا نشود. اما، در صورت همگرایی، روش بسیار توانمندی است که سرعت همگرایی خوبی نیز دارد. با ترکیب روش نیوتون-رافسون با روش‌هایی که رسیدن به جواب را تضمین می‌کنند، می‌توان به روش مناسب حل معادله دست یافت. یکی از این راهکارها، روش BFGS^۴ [۷] است که نوعی روش شبه‌نیوتون^۵ [۱۶] محسوب می‌شود. در این روش‌ها مشتق دوم تابع نیز محاسبه شده و از آن برای حرکت سریع‌تر به سوی ریشه تابع استفاده می‌شود.

۱.۲.۱ روش BFGS

روش BFGS روشی است برای محاسبات عددی بهینه‌سازی برای برنامه‌سازی غیرخطی بدون قید که جزئیات آن را در [۳۳] می‌توانید مشاهده کنید. این روش، تعمیمی برای روش بهینه‌سازی نیوتون-رافسون است. جهت کاوش، P_k ، در تکرار k ام توسط پاسخ معادله نیوتون-رافسون داده می‌شود که به صورت زیر است:

$$B_k P_k = -\nabla f(x_k)$$

به‌طوری که B_k تقریبی از ماتریس هسیان^۶ است که در هر مرحله به‌روزرسانی می‌شود و $\nabla f(x_k)$ گرادیان تابع است که در x_k مقداردهی می‌شود.

تعریف ۱.۲.۱. ماتریس هسیان، ماتریسی است مربعی که شامل مشتقات جزئی مرتبه دوم یک تابع چندمتغیره است. این ماتریس در واقع بیان‌گر میزان انحنای موضعی تابع مورد نظر به ازای مولفه‌های آن است. تابع n متغیره $f(x_1, x_2, \dots, x_n)$ با دامنه‌ای از اعداد حقیقی را در نظر بگیرید. اگر مشتقات جزئی تابع f موجود باشند، آن‌گاه ماتریس هسیان تابع f عبارت است از

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

بنا به تعریف، گزینه ماتریس هسیان را ماتریس اطلاع مشاهده‌شده^۷ می‌نامند.

برای به‌دست آوردن ریشه x ، با استفاده از روش BFGS، به صورت زیر عمل می‌شود: با شروع از مقدار اولیه x_0 و مقدار تقریبی اولیه B_0 ، مراحل زیر تا همگرایی به x تکرار می‌شوند:

^۴ Broyden-Fletcher-Goldfarb-Shanno

^۵ Quasi Newton

^۶ Hessian matrix

^۷ Observed information matrix

۱. انتخاب جهت P_k با حل $B_k P_k = -\nabla f(x_k)$.

۲. انجام کاوش خطی برای یافتن بهترین اندازه گام α_k در مرحله k ، برای به‌روز کردن مقدار

$$x_{k+1} = x_k + \alpha_k P_k$$

۳. مقداردهی $s_k = \alpha_k P_k$.

۴. محاسبه $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

۵. محاسبه $B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$.

۳.۱ رگرسیون خطی تعمیم‌یافته

مدل خطی تعمیم‌یافته، تعمیمی از مدل رگرسیون خطی است. رده GLMs زمانی مورد توجه است که متغیر پاسخ غیر نرمال، به‌ویژه گسسته یا رسته‌ای، باشد. به‌عنوان نمونه اگر متغیر پاسخ شمارشی یا دودویی باشد، معمولاً به ترتیب از مدل‌های رگرسیونی پواسون و لجستیک، که اعضای رده GLMs هستند، برای تحلیل رگرسیونی آن‌ها استفاده می‌شود.

در رده GLMs توزیع متغیر پاسخ عضوی از خانواده توزیع‌های نمایی است. اگر y_i ، برای $i = 1, 2, \dots, n$ ، مقادیر متغیر پاسخ برای یک نمونه n تایی باشند، یک مدل GLM را می‌توان به صورت

$$g(\mu_i) = g[E(y_i|x_i)] = x_i' \beta \quad (3.1)$$

نمایش داد، که در آن بردار متغیرهای تبیینی رگرسیونی برای مشاهده i ام و β بردار پارامترها یا ضرایب رگرسیونی و هم بعد با بردار متغیرهای تبیینی است. همچنین در مدل (۳.۱) تابع $g(\cdot)$ به تابع پیوند^۸ معروف است که میانگین شرطی توزیع پاسخ به شرط متغیرهای تبیینی را به متغیرهای تبیینی توسط پیش‌گوی خطی^۹ $\eta_i = x_i' \beta$ پیوند می‌دهد. هر مدل خطی تعمیم‌یافته سه جزء دارد:

۱. توزیع متغیر پاسخ که گاهی اوقات ساختار خطا نیز نامیده می‌شود و عضوی از خانواده توزیع‌های نمایی است.

۲. پیش‌گوی خطی η که متغیرهای تبیینی را شامل می‌شود و

۳. تابع پیوند $g(\cdot)$ که پیش‌گوی خطی را به میانگین (شرطی) متغیر پاسخ پیوند می‌دهد.

^۸Link Function

^۹Linear Predictor

مدل‌های رگرسیون خطی، لجستیک، دوجمله‌ای و پواسن، مثال‌هایی از رده مدل‌های خطی تعمیم‌یافته هستند. به عنوان مثال، در مدل رگرسیون خطی (۱.۱)، توزیع پاسخ نرمال است، پیش‌گوی خطی به صورت

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

است و تابع پیوند یک تابع همانی است؛ بنابراین

$$\begin{aligned} E(y|x_1, \dots, x_k) &= \mu \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k. \end{aligned}$$

پس رگرسیون خطی استاندارد، یک مدل خطی تعمیم‌یافته می‌باشد. با توجه به استفاده از رگرسیون لجستیک در این پایان‌نامه، در زیربخش بعدی آن را معرفی می‌کنیم.

۱.۳.۱ رگرسیون لجستیک

رگرسیون لجستیک یکی از مدل‌های رده GLMs است. از این مدل برای تحلیل پاسخ‌های دودویی (که مقدار پاسخ دو سطح دارد)، استفاده می‌شود.

فرض کنید n اجرای مستقل یک آزمایش به یک متغیر پاسخ دودویی y (با مقادیر ۰ و ۱) منتهی شود که به k متغیر تبیینی $x = (x_1, x_2, \dots, x_k)'$ وابسته است. اگر به دلخواه $y = 1$ را یک موفقیت و $y = 0$ را یک شکست در نظر بگیریم، آن‌گاه میانگین پاسخ $p(x_i)$ است، که در واقع $p(x_i)$ همان احتمال موفقیت در توزیع برنولی و x_i متغیرهای تبیینی برای آزمایش i ام است. با توجه به آن‌که میانگین $p(x_i)$ محدود به بازه $(0, 1)$ است، در مدل رگرسیون لجستیک، به کمک تابع پیوند لجیت^{۱۰} با تعریف

$$\text{Logit}(t) = \ln\left(\frac{t}{1-t}\right)$$

این میانگین به صورت زیر در نظر گرفته می‌شود:

$$p(x_i) = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$$

با توجه به آن‌که تعداد موفقیت‌ها در توزیع برنولی دارای توزیع دوجمله‌ای با پارامترهای n و $p(x_i)$ است، و جمله $\binom{n}{y}$ پارامترهای β را در خود ندارد، در نتیجه تابع لگاریتم درست‌نمایی مدل رگرسیون لجستیک به صورت زیر خواهد بود:

$$\ln[L(\beta)] = \sum_{i=1}^n \{y_i \ln\left[\frac{p(x_i)}{1-p(x_i)}\right] + n_i \ln[1-p(x_i)]\}.$$

^{۱۰} Logit

جمله $\ln[\frac{p(x_i)}{1-p(x_i)}]$ همان تابع لجیت است که می توان نوشت

$$\begin{aligned}\ln\left(\frac{p(x_i)}{1-p(x_i)}\right) &= \mathbf{x}_i' \boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} \\ &= \beta_0 + \sum_{j=1}^k x_{ij} \beta_j \quad i = 1, 2, \dots, n.\end{aligned}$$

در نتیجه تابع لگاریتم درستنمایی به صورت زیر قابل بازنویسی است:

$$\ln[L(\boldsymbol{\beta})] = \sum_{i=1}^n \sum_{j=1}^k y_i x_{ij} \beta_j - \sum_{i=1}^n \ln\left[1 + \exp \sum_{j=1}^k x_{ij} \beta_j\right]. \quad (4.1)$$

با نماد ماتریسی، تابع درستنمایی (۴.۱) را می توان به صورت زیر نوشت:

$$\ln[L(\boldsymbol{\beta})] = \boldsymbol{\beta}' X \mathbf{y} - \sum_{i=1}^n \ln[1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})] \quad (5.1)$$

که در آن X ماتریس طرح حاصل از متغیرهای تبیینی و $\mathbf{y}$ بردار پاسخ است. برای محاسبه برآوردهای ML پارامترهای $\boldsymbol{\beta}$ باید تابع درستنمایی (۵.۱) را نسبت به β_j ها ماکسیم کنیم. با مشتق گیری نسبت به $\boldsymbol{\beta}$ نتیجه می شود

$$\frac{\partial \ln[L(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = X' \mathbf{y} - \sum_{j=1}^n \frac{1}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}} e^{\mathbf{x}_i' \boldsymbol{\beta}} \mathbf{x}_i.$$

بنابراین با توجه به تعریف $p(x_i)$ نتیجه می شود

$$\frac{\partial \ln[L(\boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = X' \mathbf{y} - \sum_{j=1}^n p(x_i) \mathbf{x}_i.$$

چون $p(x_i)$ میانگین متغیر تصادفی برنولی است، پس عبارت سمت راست را با نماد ماتریسی می توان به صورت $X'(\mathbf{y} - \boldsymbol{\mu})$ بیان کرد، که در آن

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}$$

و $\mu_i = p(x_i)$. بنابراین، برآوردگر درستنمایی ماکسیمم جواب معادله زیر است:

$$X'(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

که به معادله امتیاز^{۱۱} معروف است.

^{۱۱}Score equation

۲.۳.۱ توزیع وایبل

توزیع وایبل در زمینه‌های کاربردی مختلف مورد استفاده قرار می‌گیرد (عبدی [۵]). این توزیع احتمالی پیوسته به‌ویژه در صنعت به‌عنوان یک مدل مناسب برای مدل‌بندی طول عمر محصولات شناخته می‌شود. اگرچه توزیع وایبل اولین بار توسط دانشمند فرانسوی فرچه [۱۳] ابداع شد و سپس رزین و راملر [۳۲] از آن برای توزیع اندازه ذرات بهره بردند، اما نام آن برگرفته از والودی وایبل [۳۷] است که آن را با جزئیات توصیف کرد.

تابع چگالی احتمال و تابع توزیع تجمعی وایبل به صورت زیر تعریف می‌شوند:

$$f(t, \alpha, \lambda) = \frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha}$$

$$F(t, \alpha, \lambda) = 1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha}$$

که در آن $\lambda > 0$ پارامتر مقیاس و $\alpha > 0$ پارامتر شکل توزیع وایبل هستند. یکی از مهم‌ترین دلایل شهرت و استفاده شایع از این توزیع در برازش داده‌ها، انعطاف بالای آن در تولید شکل‌های گوناگون تابع چگالی است. این ویژگی موجب شده است که از این توزیع برای برازش بسیاری از مجموعه داده‌ها استفاده شود؛ این توزیع، بر اساس پارامترهای خود، دو حالت برای منحنی چگالی که در داده‌های طول عمر واقعی نیز به‌طور ذاتی دیده می‌شوند، ایجاد می‌کند:

الف) به‌طور یکنواخت نزولی باشد.

ب) تک مدی باشد.

شکل تابع چگالی فقط به پارامتر شکل $\alpha > 0$ وابسته است و مستقل از پارامتر مقیاس $\lambda > 0$ عمل می‌کند. توزیع وایبل کاربردهای فراوانی دارد که در این پایان‌نامه از آن در تحلیل بقای بازپرداخت وام‌های بانکی استفاده می‌کنیم. به همین منظور، ابتدا به معرفی مفاهیم پایه‌ای تحلیل بقا می‌پردازیم.

۴.۱ تحلیل بقا

تحلیل بقا^{۱۲} مجموعه‌ای از روش‌های آماری است که در آن تحلیل داده‌ها از زمان مورد علاقه تا زمانی که حادثه مورد نظر رخ دهد، مورد بررسی قرار می‌گیرد (کلین‌بام [۱۹]). زمان رخداد واقعه مورد نظر می‌تواند شامل روزها، هفته‌ها، ماه‌ها یا حتی سال‌ها باشد، که از زمان پیگیری فرد در زمان مورد بررسی تا رخداد واقعه مورد نظر تا زمانی که سن فرد (یا زمان مطالعه) اجازه می‌دهد، به طول می‌انجامد. به‌عنوان مثال‌هایی از واقعه مورد نظر می‌توان به زمان مرگ، بهبود بیماری، بازیابی (مثل برگشت به کار) یا هر تجربه‌ای که مورد علاقه فرد باشد، اشاره کرد.

^{۱۲} Survival Analysis

اگرچه در یک تحقیق ممکن است بیش از یک رویداد، مورد نظر باشد، اما در این جا فرض می‌کنیم رخداد فقط یک رویداد مورد نظر است. در تحلیل بقا معمولاً به زمان مورد نظر اشاره می‌کنیم. مثلاً یک فرد در زمان مطالعه بیش از یک بار از مرگ نجات یابد. معمولاً از این رویداد به عنوان شکست یاد می‌شود، زیرا معمولاً رویداد مورد مطالعه مرگ، بروز بیماری یا مشکلاتی که به نتایجی منفی می‌رسند، هستند. البته زمان بقا می‌تواند نتیجه مثبت هم داشته باشد، مثل طول مدت بهبودی یک فرد بیمار پس از یک جراحی.

مطالعه داده‌های بقا (که طول مدت تا رخداد یک واقعه هستند)، با چالش‌های ویژه‌ای مواجه است که باعث پیدایش این شاخه از آمار شده است. دو چالش عمده که ذکر آن‌ها مهم است، هزینه زمانی گزاف برای دستیابی به این نوع داده‌ها و پدیده سانسور در این داده‌ها هستند. در مورد چالش اول، شاید مثال‌هایی به‌تر آن را توجیه کنند:

- بیمارانی که سرطان خون دارند، هفته‌ها یا حتی سال‌ها طول می‌کشد تا واقعه مرگ برای آن‌ها رخ دهد.
- عدم دسترسی به افراد تحت مطالعه که دچار عارضه قلبی شده‌اند و بیماری آن‌ها در حال توسعه است.
- پیگیری وضعیت مرگ و میر جمعیت سالمندان بالای ۶۰ سال، که به‌طور میانگین در یک محدوده ۱۳ ساله به طول می‌انجامد.
- زمان دستگیری مجدد یک مجرم که ممکن است هیچ‌گاه یا در یک زمان طولانی دوباره مرتکب جرم و دستگیر شود.
- مدت زمان زنده ماندن بیماران پیوند قلبی.

بیش‌تر مطالعات بقا با سانسور همراه هستند. در واقع سانسور زمانی رخ می‌دهد که مطالعه به پایان رسیده، ولی زمان بقا همچنان ادامه دارد. به‌عنوان مثال، در مطالعه بیماران مبتلا به سرطان ممکن است زمان در نظر گرفته‌شده برای انجام مطالعه به پایان برسد ولی برخی از بیماران حاضر در مطالعه همچنان زنده باشند. بنابراین با سانسور مواجه می‌شویم، به این معنی که زمان بقای کامل فرد را نمی‌توانیم مشاهده کنیم. به‌طور کلی سه دلیل برای سانسور وجود دارند:

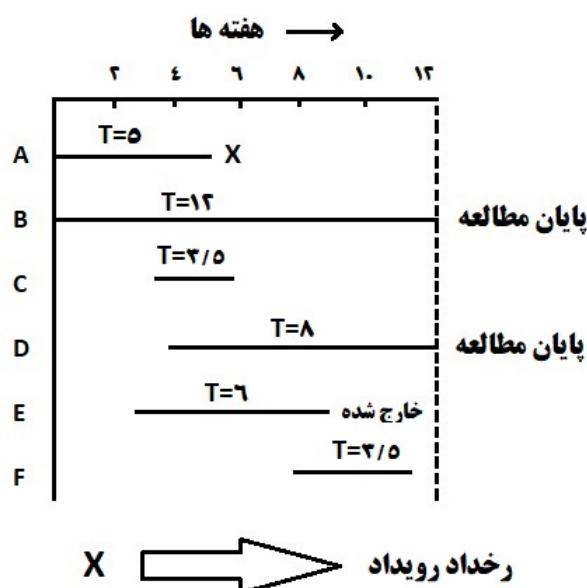
۱. فرد در طول زمان تعیین‌شده برای مطالعه رویداد را تجربه نکند و مطالعه به پایان برسد.
۲. فرد در طی زمان مطالعه از آن خارج شود و پیگیری نشود. مثلاً روند درمان خود را ادامه ندهد.
۳. فرد به دلیلی ناخواسته از مطالعه خارج شود. مثلاً اگر رخداد مورد نظر مرگ بر اثر سرطان باشد، به دلیل دیگری جز سرطان بمیرد.

در جدول ۱.۱ زمان‌های بقا برای شش نفر حاضر در یک مطالعه داده شده است. برای هر شخص زمان بقا و زمان سانسور تا رویداد مورد نظر ثبت شده است. ستون آخر نشان می‌دهد که آیا در زمان تعیین شده رویداد رخ داده است یا زمان سانسور شده است (صفر نشان‌دهنده سانسور و یک بیانگر رخداد رویداد است). به عنوان مثال، اطلاعات جدول برای فرد C شامل زمان بقای ۳/۵ است و صفر بیانگر این است که برای او سانسور رخ داده است؛ در حالی که برای فرد F زمان بقا ۳/۵ است با این تفاوت که برای او رویداد رخ داده است. این جدول یک نمایش از داده‌های بقا و انواع سانسور را ارائه می‌دهد.

جدول ۱.۱: نمونه‌ای از داده‌های بقا و سانسورها

فرد	زمان بقا	سانسور (۰) و شکست (۱)
A	۵	۱
B	۱۲	۰
C	۳/۵	۰
D	۸	۰
E	۶	۰
F	۳/۵	۱

شکل ۱.۱ اطلاعات بقای شش نمونه جدول ۱.۱ را به صورت گرافیکی نمایش می‌دهد. در این شکل برای فرد A رویداد مورد نظر پس از ۵ هفته رخ داده است. بنابراین زمان ثبت‌شده برای او زمان واقعی رخداد رویداد است و سانسوری رخ نداده است. برای شخص B زمان رخداد رویداد از پایان مطالعه، یعنی ۱۲ هفته، فراتر است و مقدار دقیق آن معلوم نیست. تنها می‌دانیم این زمان بیشتر از ۱۲ هفته خواهد بود. در این حالت، زمان رخداد واقعه برای این فرد سانسور شده است که به این نوع سانسور، سانسور از راست می‌گویند و جزئیات بیشتر آن را در بخش‌های بعدی توضیح خواهیم داد. فرد C پس از گذشت سه و نیم هفته از شروع مطالعه وارد مطالعه می‌شود و پس از ۶ هفته از شروع، رویداد مورد نظر برای وی رخ می‌دهد. بنابراین بقای این شخص پس از سه و نیم هفته سانسور می‌شود. فرد D در هفته چهارم وارد مطالعه شده و بدون این که برای او رویداد مورد نظر رخ دهد، زمان مطالعه به اتمام می‌رسد. لذا برای این فرد پس از ۸ هفته سانسور رخ می‌دهد. فرد E پس از سه هفته وارد مطالعه می‌شود و در هفته نهم از مطالعه خارج می‌شود که برای وی سانسور رخ داده است. فرد F نیز در هفته هشتم وارد مطالعه شده اما از آن جا که زمان مطالعه برای این شخص سه هفته و نیم بوده و در هفته ۱۲ رویداد مورد نظر برای او رخ داده است، مانند فرد A برای او سانسوری رخ نمی‌دهد. در این مثال، انواع سانسور و برش مشاهده می‌شوند، که آن‌ها را تشریح خواهیم



شکل ۱.۱: مثال‌هایی از انواع سانسور

کرد.

اکنون می‌توانیم برخی از مفاهیم و اصطلاحات تحلیل بقا را که در این پایان‌نامه به آن‌ها نیاز داریم، مطرح کنیم.

۱.۴.۱ مقدمه‌ای بر مفاهیم بقا

تحلیل بقا شامل روش‌های آماری متنوع برای تحلیل متغیرهای تصادفی است که دارای مقادیر مثبت هستند (آشفته [۱]). مقدار این متغیر تصادفی، کمیت‌هایی مثل زمان شکست یک مؤلفه فیزیکی (مکانیکی یا الکترونیکی) یا زمان مرگ یک واحد زنده (سلول، بیمار، یا حیوان) است. البته ممکن است متغیر تصادفی بقا ربطی به زمان نداشته باشد. برای مثال، متغیر تصادفی مربوط به هزینه پرداختی یک شرکت بیمه به بیمه‌شدگان در یک حالت خاص؛ برای بعضی از افراد ممکن است بیماری بهبود یافته و مبلغ کل پرداختی بیمه آن‌ها معلوم باشد، ولی برای برخی دیگر، بیماری هنوز ادامه داشته و تنها مبلغ پرداختی تا آن زمان، معلوم باشد. در متون مربوط به داده‌های بقا، برای تحلیل بقا نام‌های گوناگونی به کار برده‌اند؛ تحلیل داده‌های طول عمر^{۱۳}، تحلیل قابلیت اعتماد^{۱۴}، و تحلیل زمان تا رخداد پیشامدها^{۱۵} که بسته به نوع کاربرد تحلیل بقا، ممکن است از هر کدام از آن‌ها استفاده شود.

^{۱۳}Life Time Data

^{۱۴}Reliability Analysis

^{۱۵}Time to Event Analysis

برای تعیین زمان بقا، دو نقطه زمانی باید معلوم شوند: (۱) زمان مبدا که می‌تواند زمان شروع مطالعه یا زمان پیدایش مورد تحت مطالعه مثل تولد یک انسان یا شروع بازپرداخت‌های بانکی افراد باشد و (۲) زمان رخداد واقعه یا به اصطلاح شکست، یعنی زمانی که در آن پیشامد نهایی مثل مرگ رخ می‌دهد، یا زمان پایان مطالعه. داده‌های زمان بقا در بسیاری از کاربردها دارای اهمیت هستند. مثلاً شناخت عوامل موثر بر طول فاصله زمانی تا وقوع رویدادی مثل مرگ یک موجود زنده در علوم پزشکی و مطالعات سرطان از اهمیت بالایی برخوردار است.

در بسیاری از کاربردها، شکست یا رخداد حادثه مورد نظر می‌تواند حداکثر یک بار برای هر فرد اتفاق بیافتد. طول عمر یک ماشین صنعتی، اولین زمان مراجعه یک اتومبیل نو به تعمیرگاه، مدت اعتصاب یا زمان بیکاری افراد، زمان بقای یک بیمار پس از درمان یا انجام یک عمل جراحی، و زمان طلاق یک زوج پس از ازدواج، مثال‌هایی از این نوع داده‌ها هستند. در کاربردهای مختلف دیگر نیز رخداد حادثه می‌تواند بیشتر از یک بار رخ دهد؛ به‌عنوان مثال، بازگشت مجدد یک زندانی آزادشده به زندان، روی آوردن دوباره یک معتاد پس از ترک اعتیاد، و عود کردن یک بیماری از آن جمله هستند. به این نوع داده‌ها، داده‌های بازگشتی^{۱۶} می‌گویند. از آنجایی که این روش‌ها در ابتدا غالباً برای مطالعات مرگ و میر به کار برده می‌شد، نام ”تحلیل زمان بقا“ برای آن‌ها مصطلح شد. تحلیل بقا از تحلیل جداول عمر در قرن ۱۷ میلادی شروع شد [۱۵]. شخصی به نام گرانٹ [۱۵] اولین فهرست هفتگی مرگ و میر را در لندن منتشر کرد. تحلیل بقا از نظر علم آمار مجموعه‌ای از روش‌های مختلف آماری در تحلیل متغیرهای تصادفی نامنفی است که مقدار آن می‌تواند زمان شکست یک مولفه فیزیکی یا زمان مرگ یک واحد زنده باشد.

برای مشخص نمودن زمان بقا می‌بایست سه مولفه

۱. مبدا زمان

۲. مقیاس و واحد اندازه‌گیری گذشت زمان

۳. مفهوم شکست یا وقوع حادثه

به‌طور واضح و دقیق تعریف شوند و هیچ‌گونه ابهامی برای آن‌ها وجود نداشته باشد. در مطالعات بالینی زمان بقا به زمان مرگ، پیشرفت یک نشانه خاص یا زمان عود بیماری پس از بهبود آن، اشاره دارد. در داده‌های بقا، نقطه پایان بسیار واضح، روشن و تعریف شده است ولی نقطه شروع به وضوح مشخص نیست، اما بهترین تعریف آن زمانی است که بیماری تشخیص داده می‌شود.

تعریف ۱.۴.۱. متغیر تصادفی بقا یک متغیر تصادفی نامنفی است که معمولاً فاصله زمانی ورود به مطالعه تا وقوع حادثه مورد علاقه تعریف می‌شود.

^{۱۶} Recurrent Event Data

متغیر تصادفی برای زمان بقای شخص را با T نشان می‌دهیم، که همواره نامنفی است؛ یعنی $T \geq 0$. در این پایان‌نامه از حرف t برای نمایش مقدار مشاهده‌شده زمان بقا برای یک آزمودنی استفاده می‌کنیم.

در مطالعات بقا، یکی از اهداف کلیدی پاسخ به این پرسش است که مثلاً احتمال زنده ماندن فردی بیش از ۵ سال چقدر است؟ یا چه عواملی بر این احتمال تاثیرگذار هستند؟ برای پاسخ به این سوال، از کمیت‌های کلیدی مانند تابع بقا^{۱۷}، و تابع مخاطره^{۱۸} استفاده می‌شود که در ادامه آن‌ها را معرفی می‌کنیم.

تعریف ۲.۴.۱ (تابع بقا): این تابع در نقطه زمانی t احتمال زنده ماندن آزمودنی بعد از t است. این تابع را با $S(t)$ نشان می‌دهند و عبارتست از

$$S(t) = P(T > t) \quad 0 \leq t < \infty.$$

یک تابع بقای سره در شرایط زیر صدق می‌کند:

$$S(0) = 1, \quad S(\infty) = 0.$$

با توجه به تعریف تابع بقا، واضح است که $S(t)$ یک تابع ناصعودی و از چپ پیوسته است.

تعریف ۳.۴.۱ (تابع توزیع تجمعی): این تابع را با $F(t)$ نشان داده و احتمال کم‌تر یا مساوی با زمان مشخص t بودن متغیر تصادفی T تعریف می‌شود. به عبارت دیگر

$$F(t) = P(T \leq t).$$

بنا به تعریف تابع بقا، می‌توان نتیجه گرفت $F(t) = 1 - S(t)$.

تعریف ۴.۴.۱ (تابع چگالی احتمال): برای متغیر تصادفی پیوسته بقا، تابع چگالی احتمال، مشتق تابع توزیع تجمعی است و آن را با $f(t)$ نشان می‌دهند. بنابراین

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

تعریف ۵.۴.۱ (تابع مخاطره): این تابع احتمال وقوع رویداد مورد نظر به‌طوری‌آنی است و آن را با نماد $h(t)$ برای واحد زمانی t نشان می‌دهند. به‌طور دقیق‌تر، داریم

$$h(t) = \frac{f(t)}{S(t)}.$$

مشخصه بارز داده‌های بقا که آن را از سایر داده‌ها متمایز می‌کند، وجود پدیده سانسور در آن‌ها است. سانسور شدن زمانی اتفاق می‌افتد که اطلاعات ناقصی درباره زمان بقای آزمودنی‌ها در دست باشد. برای نمایش شکست یا رخداد سانسور در زمان بقا از نماد δ استفاده می‌کنیم؛ هرگاه $\delta = 0$ باشد، به معنی وجود سانسور و هرگاه $\delta = 1$ باشد، به معنی رخداد رویداد مورد نظر (شکست) است.

^{۱۷}Survival Function

^{۱۸}Hazard Function

۲.۴.۱ برآورد تابع بقا

اگر نمونه‌ای از زمان‌های بقای آزمودنی‌ها داشته باشیم، می‌توان نسبتی از جامعه که در طول یک دوره معین و تحت شرایط یکسان، زنده می‌مانند را برآورد کرد. مثلاً با استفاده از اطلاعات مربوط به بیماران سرطانی می‌توان احتمال زنده ماندن یک بیمار جدید در طول یک فاصله زمانی را به‌طور تقریبی، محاسبه کرد.

همان گونه که یک دوره زمانی مشخص را می‌توان به چند فاصله زمانی کوتاه‌تر تقسیم کرد، احتمال بقا و زنده ماندن در آن دوره زمانی مشخص را نیز می‌توان بر اساس حاصل ضرب احتمال فواصل کوتاه‌تر محاسبه کرد. مثلاً احتمال این که شخصی در مدت دو هفته زنده بماند را می‌توان با حاصل ضرب احتمال زنده ماندن در هفته اول و احتمال زنده ماندن در هفته دوم به شرط زنده ماندن در هفته اول به‌دست آورد.

یک روش ناپارامتری معمول برای برآورد تابع بقا، روش کاپلان-مایر^{۱۹} (KM) است [۱۸]. از این روش در اغلب مثال‌های تحلیل بقا استفاده می‌شود و با سایر برآوردهای پارامتری مبتنی بر مدل مقایسه می‌شود.

برآوردگر کاپلان-مایر

برآوردگر کاپلان-مایر، به برآوردگر حاصل ضرب حدی^{۲۰} نیز معروف است. ادوارد. ال. کاپلان^{۲۱} و پل مایر^{۲۲} مقاله‌های مشابهی در مورد این برآوردگر به مجله انجمن آمار آمریکا ارسال کردند، اما سردبیر مجله تصمیم گرفت که کار آن‌ها را در یک مقاله با هم ترکیب کند و به برآوردگر کاپلان-مایر معروف شد.

این روش اغلب در تحقیقات پزشکی برای اندازه‌گیری مدت زمان زنده ماندن بیماران بعد از گذراندن مراحل درمانی مشخص، استفاده می‌شود [۴]. در اقتصاد، از آن برای برآورد طول مدت زمان بیکاری افرادی که شغل خود را از دست داده‌اند، استفاده می‌کنند [۲۵]. در مهندسی این برآوردگر برای محاسبه زمان باقی‌مانده تا خرابی قطعات یک دستگاه مکانیکی مورد استفاده قرار می‌گیرد [۱۷]. در بوم‌شناسی نیز از آن برای برآورد مدت زمان باقی ماندن میوه‌های گوشتی بر روی درختان تا قبل از این که توسط میوه‌خواران از بین بروند، استفاده می‌کنند [۲۳].

این برآوردگر یکی از قدیمی‌ترین روش‌های ناپارامتری برآورد تابع بقا برای برآورد طول عمر بیماران است که اثر داده‌های سانسور از راست را نیز در محاسبات خود در نظر می‌گیرد. این روش برای نمونه‌های کوچک که زمان وقوع حوادث به دقت ثبت و اندازه‌گیری می‌شوند، کارا

^{۱۹} Kaplan-Meier

^{۲۰} Limit-Product

^{۲۱} E.L. Kaplan

^{۲۲} P. Meier

است.

فرض کنید تعداد کل آزمودنی‌های شرکت‌کننده در یک مطالعه که تحت پیگیری و مراقبت قرار دارند برابر با n و $t_1, t_2, \dots, t_n$ زمان‌های بقای متناظر آن‌ها باشند. برای به‌دست آوردن برآوردگر کاپلان-مایر، ابتدا زمان‌های بقا را به‌صورت صعودی مرتب می‌کنیم، به‌طوری که $t_1 < t_2 < \dots < t_n$. تعداد آزمودنی‌هایی که در زمان t_j در معرض خطرند و همچنین آن‌هایی که در همین زمان رویداد مورد علاقه برای آن‌ها رخ داده است را با n_j نشان می‌دهیم. همچنین d_j را برابر تعداد رخدادها در زمان t_j قرار می‌دهیم. با پذیرفتن این که $t_j \leq t_{j+1}$ و استفاده از

$$P(A \cap B) = P(B)P(A|B)$$

می‌توان نوشت

$$\begin{aligned} S(t_j) &= P(T > t_j) \\ &= P(T > t_j, T > t_{j-1}) \\ &= P(T > t_j | T > t_{j-1}) P(T > t_{j-1}) \\ &= P(T > t_j | T > t_{j-1}) P(T > t_{j-1} | T > t_{j-2}) P(T > t_{j-2}) \\ &\vdots \\ &= P(T > t_j | T > t_{j-1}) P(T > t_{j-1} | T > t_{j-2}) \dots P(T > t_0). \end{aligned}$$

با توجه به این که در ابتدای مطالعه تمام افراد نمونه زنده هستند، پس

$$P(T > t_0 = 0) = 1$$

و با استفاده از تعریف احتمال شرطی نتیجه می‌شود

$$P(T > t_j | T > t_{j-1}) = \frac{n_j - d_j}{n_j}.$$

بنابراین برآوردگر کاپلان-مایر به صورت زیر تعریف می‌شود:

$$\hat{S}(t) = \prod_{j: t_j \leq t} \frac{n_j - d_j}{n_j} = \prod_{j: t_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

اگر داده‌ها از راست سانسور شده باشند، در این صورت

$$n_{j+1} = n_j - c_j - d_j \quad j = 0, 1, \dots, n$$

که در آن c_j تعداد سانسورها در زمان t_j می‌باشد. برای دیدن جزئیات بیشتر می‌توانید به کتاب بزرگ‌نیا [۲] مراجعه کنید.

۳.۴.۱ داده‌های سانسور شده

در بسیاری از آزمون‌های طول عمر، تحقیقات زیست‌شناسی و سایر زمینه‌های علمی، با نمونه‌هایی سروکار داریم که در آن‌ها همه مشاهدات نمونه ثبت نشده‌اند یا همه واحدهای نمونه به نتیجه نرسیده‌اند. به عنوان مثال، در یک نمونه به حجم n ممکن است تنها اطلاعات مربوط به n_1 واحد نمونه به طور کامل در دسترس باشند و برای $n - n_1$ واحد دیگر نمونه اطلاعات کاملی وجود نداشته باشند. این محدودیت ممکن است به صورت اختیاری یا از قبل توسط آزمایشگر اعمال شود.

به عنوان نمونه، اگر یکی از واحدهای آزمایشی به طور تصادفی خراب شود یا شخص مورد مطالعه از ادامه همکاری کناره‌گیری کند، در این صورت حذف از نوع غیرعمدی رخ داده است. در مقابل، اگر به علت برخی شرایط ویژه مانند تمام شدن بودجه، طولانی شدن مدت آزمایش، یا فرصت کم برای اعلام نتایج، آزمایشگر آزمایش را قبل از حصول زمان خرابی تمام واحدها متوقف کند، حذف از نوع عمدی اتفاق افتاده است.

همان طور که قبلاً تعریف کردیم به این نوع داده‌ها که واحدهای آزمایشی قبل از مشاهده زمان رخداد واقعه از آزمایش یا مطالعه خارج می‌شوند، داده سانسور شده می‌گویند. بدیهی است با سانسور کردن مشاهدات، مقداری از اطلاعات موجود در نمونه از دست خواهند رفت. با این وجود گاهی اوقات مایل یا مجبور به سانسور کردن هستیم.

سانسورها به صورت‌های مختلفی اعمال می‌شوند. بعضی از انواع سانسورها عبارتند از سانسور از راست، سانسور از چپ، سانسور نوع اول^{۲۳}، سانسور نوع دوم^{۲۴}، سانسور فاصله‌ای^{۲۵}، و سانسور ترکیبی^{۲۶}. برای جزئیات بیش‌تر درباره انواع سانسورها و کاربردهای آن‌ها، می‌توانید به کتاب‌های آرنولد و همکاران (۱۹۹۲)، لاولس (۲۰۰۳، ۱۹۸۲) و نلسن (۲۰۰۴) مراجعه کنید. یکی از اولین تلاش‌ها برای تحلیل مسائل آماری که شامل داده‌های سانسور شده بود، توسط برنولی (۱۷۶۶) انجام گرفت که در تحلیل شیوع بیماری آبله و اطلاعات مربوط به مرگ و میر در اثر این بیماری بود. سانسور از راست یکی از انواع شایع سانسور در مثال‌های واقعی است.

سانسور از راست

همان طور که گفته شد داده‌ها در مطالعات بقا اغلب با سانسور مواجه می‌شوند. معمول‌ترین نوع سانسور، سانسور از راست است که در این حالت زمان رخداد واقعه مورد نظر از مقدار مشاهده شده بزرگ‌تر است. به بیان دیگر، اگر آزمایش پیش از مشاهده رخداد واقعه تمام واحدها پایان یابد و در واقع مشاهدات بزرگ‌تر (آخر) کنار گذاشته شوند، سانسور از راست رخ

^{۲۳}Type-I Censoring

^{۲۴}Type-II Censoring

^{۲۵}Interval Censoring

^{۲۶}Hybrid Censoring

داده است.

از دلایل معمول سانسور از راست، عدم رخداد واقعه (حذف) آزمودنی قبل از پایان مطالعه است. به عنوان مثال، در ارزیابی طول عمر دام‌هایی که در گله‌های گاو شیری زنده هستند، داده سانسور شده به دلیل یکی از عوامل زیر به وجود می‌آید: (۱) حیوان در انتهای دوره مطالعه و جمع‌آوری داده‌ها، هنوز زنده است؛ (۲) هنگامی که حیوان از یک گله به گله دیگر که تحت پوشش مطالعه نیست فروخته می‌شود؛ (۳) زمانی که همه حیوانات موجود در یک گله از برنامه ارزیابی حذف می‌شوند.

سانسور از چپ

اگر در شروع آزمایش، زمان رخداد چند آزمودنی نمونه تحت مطالعه را نداشته باشیم، سانسور از چپ رخ داده است. به عنوان مثال، در مثال گاوهای شیری، سانسور از چپ در حالت حذف حیوان (رخداد واقعه) قبل از شروع مطالعه اتفاق می‌افتد.

۴.۴.۱ انواع برش‌ها

در تحلیل‌های بقا علاوه بر پدیده سانسور، امکان مشاهده برش^{۲۷} نیز وجود دارد. به طور کلی، یک مشاهده بریده شده، تنها قسمتی از اطلاعات مربوط به متغیرهای تصادفی مورد نظر را در بر دارد. در این پایان‌نامه سه نوع برش را معرفی می‌کنیم. ابتدا فرض کنید $T_1, T_2, \dots, T_n$ متغیرهای تصادفی مستقل و هم‌توزیع (i.i.d.) با تابع توزیع F هستند.

برش نوع اول

در برش نوع اول به جای مشاهده T_1 تا T_n (متغیرهای تصادفی مورد توجه)، متغیرهای تصادفی Y_1 تا Y_n مشاهده می‌شوند که برای $i = 1, \dots, n$ به صورت زیر تعریف می‌شوند:

$$Y_i = \begin{cases} T_i & T_i \leq t_c \\ t_c & T_i > t_c \end{cases} \quad (۶.۱)$$

ثابت که t_c را زمان برش نامند. توجه کنید تابع توزیع Y در $y = t_c$ دارای جرم احتمال مثبت $P(T > t_c) > 0$ است.

برش نوع دوم

فرض کنید $r < n$ عددی ثابت و $T_{(1)} < T_{(2)} < \dots < T_{(n)}$ آماره‌های مرتب T_1 تا T_n باشند. در برش نوع دوم زمانی که r امین واقعه رخ دهد، مطالعه پایان می‌پذیرد. بنابراین، در این

^{۲۷}Truncation

حالت فقط متغیرهای T_1 تا T_r قابل مشاهده هستند. در این صورت، متغیرهای قابل مشاهده عبارتند از

$$\begin{aligned} Y_1 &= T_{(1)} \\ Y_2 &= T_{(2)} \\ &\vdots \\ Y_r &= T_{(r)} \\ Y_{r+1} &= T_{(r)} \\ &\vdots \\ Y_n &= T_{(r)}. \end{aligned}$$

هر دو نوع برش اول و دوم در علوم مهندسی به کار گرفته می‌شوند. برای مثال، تعدادی لامپ یا ترانزیستور را در نظر بگیرید که تمام آن‌ها را در زمان $t = 0$ مورد آزمایش قرار می‌دهیم و زمان شکست آن‌ها را ثبت می‌کنیم. ممکن است طول عمر بعضی از ترانزیستورها طولانی باشد و نخواهیم به مدت طولانی منتظر بمانیم تا آزمایش به پایان برسد. بنابراین، امکان دارد آزمایش را در زمان t_c ، که از قبل تعیین شده است، پایان دهیم. در این حالت برش نوع اول انجام شده است. برای برش نوع دوم فرض کنید از قبل ندانیم چه زمانی برای برش مناسب است. لذا، تصمیم بگیریم که در انتظار بمانیم تا نسبت معینی از ترانزیستورها، مانند $\frac{r}{c}$ از کار بیافتند. در این حالت برش نوع دوم رخ داده است.

برش نوع سوم (برش تصادفی)

این نوع برش به برش تصادفی^{۲۸} نیز معروف است. فرض کنید $C_1, \dots, C_n$ متغیرهای تصادفی i.i.d. با تابع توزیع G باشند. مقدار مشاهده‌شده برای C_i زمان برش مربوط به T_i است. در این حالت فقط می‌توانیم $(Y_1, \delta_1), \dots, (Y_n, \delta_n)$ را مشاهده کنیم، که در آن Y_i و δ_i به صورت زیر تعریف می‌شوند:

$$Y_i = \min(T_i, C_i) = T_i \wedge C_i$$

و

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & T_i \leq C_i & \text{یعنی } T_i \text{ بریده نمی‌شود} \\ 0 & T_i > C_i & \text{یعنی } T_i \text{ بریده می‌شود} \end{cases}$$

توجه کنید که $Y_1, \dots, Y_n$ نیز متغیرهای تصادفی i.i.d.، مثلاً با تابع توزیع H ، هستند. علاوه بر این، $\delta_1, \dots, \delta_n$ شامل اطلاعات مربوط به برش هستند.

^{۲۸}Stochastic Truncation

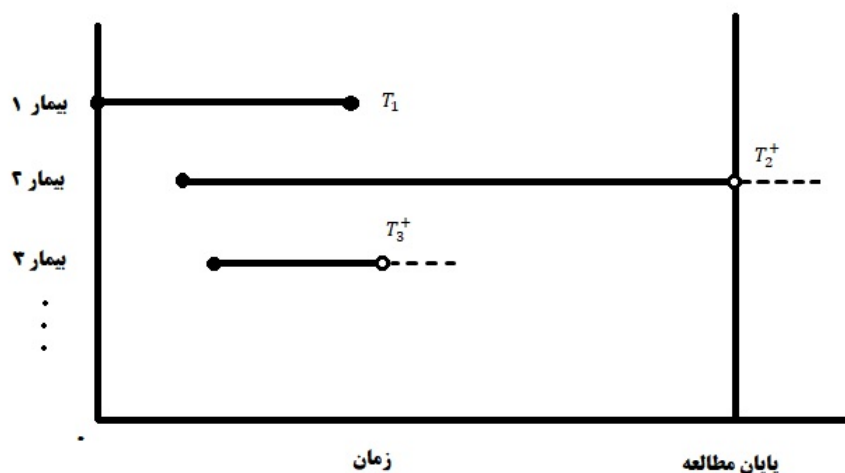
از آن جا که در برش نوع اول و دوم می‌توانستیم ببینیم چه مشاهداتی بریده می‌شوند، تعریف δ_i ها به‌طور واضح ضروری نبود.

برش تصادفی در مطالعات پزشکی و در مورد افراد یا آزمایش‌های بالینی استفاده می‌شود. در یک آزمایش بالینی، ممکن است بیماران برای معالجه در زمان‌های مختلف مراجعه کنند. سپس، هر بیمار به یکی از روش‌های ممکن معالجه شود. حال اگر بخواهیم طول عمر آن‌ها را مطالعه کنیم، عمل برش به یکی از صورت‌های زیر انجام می‌شود:

عدم مراجعه: ممکن است بیمار تصمیم بگیرد به پزشک دیگری مراجعه کند. در نتیجه دیگر او را نخواهیم دید.

قطع معالجه: ممکن است به علت عوارض بد جانبی، درمان را متوقف کنیم.

اتمام بررسی



شکل ۲.۱: برش تصادفی برای یک بیمار

نمودار ۲.۱ نمایشی از برش تصادفی را تشریح می‌کند. در این مثال، بیمار ۱ در زمان $t = 0$ مورد بررسی قرار گرفته و در زمان T_1 فوت شده است. در نتیجه، یک مشاهده کامل (بریده‌نشده) به‌دست آمده است. بیمار ۲ مورد بررسی قرار گرفته و در پایان بررسی هنوز زنده است. در نتیجه یک مشاهده کامل T_2^+ به‌دست آمده است. بیمار ۳ نیز مورد بررسی قرار گرفته و قبل از پایان بررسی، معالجه را قطع کرده است. در نتیجه، یک مشاهده بریده‌شده T_3^+ نتیجه شده است.

یکی از پذیره‌های اساسی برش تصادفی، استقلال متغیرهای تصادفی C_i و T_i است. بدون این پذیره، مساله پیچیده می‌شود. البته معقول به نظر می‌رسد که فرض کنیم مطالعه در زمان‌های تصادفی شروع، به‌طور تصادفی بررسی و قطع می‌شود. با این وجود، اگر دلیلی برای قطع بررسی در طول مطالعه وجود داشته باشد، در این صورت ممکن است بین C_i و T_i ، رابطه‌ای وجود داشته باشد.

در این پایان‌نامه، تنها مشاهدات از راست سانسور شده مورد توجه هستند.

۵.۱ مدل‌های بقا

در تحلیل داده‌های بقا، در کاربردهای مختلف، به دنبال بررسی عوامل موثر بر احتمال بقا (یا تابع مخاطره) هستند. این عوامل بر حسب متغیرهای تبیینی رگرسیونی خود را نشان می‌دهند. برای این کار، در متون تحلیل بقا از رده‌های مختلف مدل‌های پارامتری [۱۹]، ناپارامتری [۱۹] و نیمه‌پارامتری^{۲۹} [۳۶] استفاده می‌شوند. با توجه به استفاده ما از مدل پارامتری وایبل^{۳۰} در این پایان‌نامه، به‌طور مختصر مدل‌های پارامتری را معرفی می‌کنیم.

۱.۵.۱ مدل‌های پارامتری بقا

مدل‌های پارامتری بقا یک زیررده از رده مدل‌های بقا است که در کاربردهای متعدد تحلیل بقا مورد توجه هستند. در مدل‌های پارامتری، فرض می‌شود متغیر تصادفی T از یک خانواده از توزیع‌های پارامتری معلوم پیروی می‌کند. در مدل‌های پارامتری، نتیجه توزیع (به‌عنوان مثال زمان مرگ بیماران مبتلا به سرطان ریه) با توجه به پارامترهای نامعلوم یک توزیع پارامتری معلوم تعیین می‌شود. مدل‌های پارامتری بقا، عملکرد تابع بقا را به صورت پارامتری به مجموعه‌ای از داده‌های بقا متصل می‌کنند.

از مدل‌های پارامتری معمول بقا که در تحلیل داده‌ها مورد استفاده قرار می‌گیرند، می‌توان به مدل‌های نمایی^{۳۱}، وایبل و لگ‌نرمال^{۳۲} اشاره کرد. توزیع وایبل یکی از مدل‌های پرتعداد است که برای مدل‌بندی زمان‌های بقا مورد استفاده قرار می‌گیرد. برای مطالعه بیش‌تر به کتاب کلین‌بام [۱۹] مراجعه کنید.

^{۲۹} Semiparametric Models

^{۳۰} Weibull

^{۳۱} Exponential

^{۳۲} Log Normal

فصل ۲

مدل‌های نرخ شفایافتگی

۱.۲ مقدمه

پیشرفت‌های مدرن علم پزشکی در حوزه‌های مختلف، به‌ویژه سرطان، امکان درمان برخی از بیماری‌ها را فراهم کرده است. به‌عنوان مثال، برای سرطان ریه که قبلاً برای آن درمانی وجود نداشت، با پیشرفت‌های پزشکی صورت گرفته برای درمان آن، امکان درمان یک بیمار مبتلا و عدم رخداد مرگ به واسطه سرطان ریه وجود دارد. بنابراین در تحلیل داده‌های بقایی که زمان رخداد واقعه مورد نظر (مثل مرگ بر اثر سرطان ریه) می‌تواند خیلی بلندمدت باشد، تابع بقا همه آزمودنی‌های تحت مطالعه، در طولانی مدت^۱ ممکن است به صفر میل نکند و به اصطلاح یک تابع بقای ناسره^۲ باشد.

مدل‌های نرخ شفایافتگی^۳ (CRMs) تعمیمی از مدل‌های بقای معمولی هستند که ویژگی شفایافتگی^۴ فقط بخشی از جمعیت مورد مطالعه را در نظر می‌گیرد که رویداد مورد علاقه را هرگز تجربه نمی‌کنند (وحیدپور [۳۵]). در مدل‌های نرخ شفایافتگی، پذیره سره بودن تابع بقا لزوماً برقرار نیست یعنی

$$\lim_{t \rightarrow \infty} S(t) \neq 0.$$

^۱ Long-term

^۲ Improper Survival Function

^۳ Cure Rate Models

^۴ Cure

در یک دسته‌بندی، مدل‌های نرخ شفایافتگی به دو زیررده تقسیم می‌شوند. زیررده اول، مدل‌های نرخ شفایافتگی آمیخته^۵ (برکسون و گیج، [۹]) هستند که به‌طور گسترده‌ای برای مدل‌بندی داده‌های بقا با نسبت شفایافتگی مورد استفاده قرار می‌گیرند. زیررده دوم، مدل‌های نرخ شفایافتگی غیرآمیخته^۶ (چن و همکاران، [۱۱]) هستند که کم‌تر مورد توجه قرار گرفته‌اند. در مدل‌های نرخ شفایافتگی، زمان بقای افراد درمان‌شده ممکن است در پایان مطالعه سانسور شود. به عبارتی اگر مطالعه ادامه داشته باشد، ممکن است افراد مورد مطالعه درمان شوند و واقع هرگز رخ ندهد.

به‌طور کلی، هرچند افراد شفایافته شامل زیرمجموعه‌ای از افراد سانسور شده هستند، اما جنبه چالش برانگیز در مدل‌های نرخ شفایافتگی، مدل‌بندی نسبت شفایافتگی است که چندان واضح نیست. برای همین توصیه می‌شود قبل از تحلیل داده‌های بقا، ابتدا وجود نسبت شفایافتگی بررسی شود و در صورت تایید، از مدل‌هایی که این شفایافتگی را لحاظ می‌کنند، استفاده شود.

۲.۲ تاریخچه

مقاله برکسون و گیج [۹] مهم‌ترین مرجع برای مدل‌های نرخ شفایافتگی است. ایده مدل‌های نرخ شفایافتگی آمیخته از مقایسه بین منحنی بقا^۷ برای دو گروه درمان و کنترل در یک مطالعه پزشکی برای مدل‌بندی روند پیشرفت یک بیماری، آغاز شد. در آن مطالعه، تابع بقای دو گروه، بر حسب لگاریتم زمان بقای آن‌ها، روی یک شکل رسم شدند و مورد مقایسه قرار گرفتند. تفاوت بقای دو گروه از روی اختلاف دو منحنی در جهت محور عمودی قابل محاسبه و مقایسه است. وضعیت دو منحنی بعد از گذشت زمان مشخصی به صورت موازی با محور زمان قرار گرفتند که به معنی نرخ مخاطره مساوی برای دو گروه بود. به عبارت دیگر از جایی به بعد، مخاطره‌های دو گروه کنترل و درمان یکسان شدند و این بدان معنی بود که برخی از گروه درمان، شفا یافته‌اند.

نسبت شفایافتگی^۸ نسبت بیمارانی است که دارای نرخ شکست نرمال (نرخ بقای غیرصفر به واسطه بیماری) در گروه کنترل هستند. این تعریف ابتدا توسط برکسون و گیج [۹] پیشنهاد شد. برکسون و گیج [۹] جمعیت را به دو گروه فرضی تقسیم کردند: یک گروه با نرخ شکست نرمال که بخش شفایافته نامیده می‌شود و گروه دیگر با نرخ شکست خاصی که می‌تواند نرخ شکست برای یک بیماری تحت مطالعه باشد. این دو نرخ شکست توسط $1 - p_0$ ، برای نرخ شکست نرمال و p_0 برای نرخ شکست حاصل از بیماری خاص نشان داده می‌شوند. این دو نرخ

^۵Mixture Cure Rate Models

^۶Non-mixture Cure Rate Models

^۷Survival Curves

^۸Cured Fraction

شکست به طور همزمان و مستقل عمل می کنند. احتمال بقا برای بخش شفایافته با نسبت p_0 ، برابر یک است و احتمال بقا برای بخش غیرشفایافته با نسبت $1 - p_0$ ، از تابع بقای $S(t)$ پیروی می کند. افرادی که درمان شده اند بخشی از کسانی هستند که در پایان آزمایش زنده مانده اند و افرادی که درمان نشده اند افرادی هستند که با رویداد مورد بررسی در حال درمان هستند یا مرگ برای آن ها رخ داده است.

مطالعات بقا و به دنبال آن بررسی نرخ شفایافتگی در طی دو سه دهه گذشته همواره مورد بحث و بررسی قرار گرفته است. اولین بار برکسون و گیج [۹] مدل نرخ شفایافتگی را معرفی کردند. چن و همکاران [۱۱] یک نسخه بیزی از این مدل را ارائه دادند. بر روی این مدل ها پژوهش های زیادی طی سال های اخیر انجام شده اند.

باریگا و همکاران [۸] از این مدل ها برای تحقیقات بانکی بهره برده اند. هم چنین در زمینه مدل بندی ریسک اعتباری، اولوریا و لوزادا [۲۷] مدلی ارائه دادند که زمان اولین پیگیری برای وام های غیرفعال را مدل می کند. اولوریا و لوزادا [۲۶] زمان دریافت وام تا زمان بازپرداخت کامل وام های بانکی را با استفاده از مدل های نرخ شفایافتگی مورد تحلیل و بررسی قرار دادند. لوزادا و همکاران [۲۵] یک مدل نرخ شفایافتگی آماسیده در صفر^۹ را که در فصل سوم به طور کامل در مورد آن صحبت می کنیم، برای برآورد نرخ تقلب در وام های بانکی معرفی کردند. مدل پیشنهادی آن ها غیرآمیخته بود. یک نسخه آمیخته مشابه آن، توسط لوزادا و همکاران [۲۱] معرفی شد.

۳.۲ مدل نرخ شفایافتگی

در صورتی که تمام افراد جامعه در معرض خطر رویداد مورد نظر نباشند، برای مدل بندی طول زمان رخداد رویداد مذکور می توان از مدل های نرخ شفایافتگی استفاده کرد. به عبارتی، اگر در مطالعات مربوط به بررسی زمان تا رخداد یک پیشامد درصدی از افراد جامعه در مقابل آن پیشامد مصونیت داشته باشند از مدل نرخ شفایافتگی استفاده می کنیم. یکی از نکات مهم که باید در مدل های نرخ شفایافتگی به آن توجه کرد، این است که از این مدل فقط در مطالعاتی می توان استفاده نمود که طول مدت مطالعه به قدر کافی طولانی باشد تا تمام افراد مشکوک پیشامد مورد نظر را تجربه کنند.

مدل های نرخ شفایافتگی را می توان به دو خانواده مدل های آمیخته و غیرآمیخته رده بندی کرد. در ادامه، هر یک از این دو زیررده را تشریح می کنیم.

^۹Zero-inflated Cure Rate Model

۱.۳.۲ مدل نرخ شفایافتگی آمیخته

در برخی از مطالعات و پژوهش‌های پزشکی، بیماران مبتلا به بیماری‌هایی هستند که پس از درمان به‌طور کامل بهبود می‌یابند و به آن‌ها شفایافته می‌گویند. زمانی که تحلیل زمان بقای افراد مبتلا به این بیماری‌ها مورد توجه قرار می‌گیرد، رویداد مورد نظر (مثل مرگ بر اثر بیماری) برای بخش شفایافته اتفاق نخواهد افتاد. اولین مقاله در این رابطه برای برآورد نسبت بیماران سرطانی که به درمان پاسخ داده‌اند توسط بوگ [۱۰] ارائه گردید.

برکسون و گیج [۹] مدل نرخ شفایافتگی آمیخته را پیشنهاد کردند که به وسیله آن هم بیماران شفایافته و هم بیمارانی که با شکست مواجه شده‌اند، در مدل لحاظ می‌شوند. در مدل‌های نرخ شفایافتگی آمیخته فرض بر این است که جامعه به صورت ناهمگن از دو زیرجامعه، شامل افراد در معرض خطر^{۱۰} و افراد شفایافته یا مصون، تشکیل شده است. در این مدل‌ها فرض می‌شود هر فرد با احتمال $1-p$ در معرض خطر برای تجربه پیشامد مورد نظر باشد و با احتمال p شفایافته باشد. در این مدل‌ها، تابع بقا و تابع چگالی همه افراد به صورت زیر تعریف می‌شوند:

$$S(t) = p + (1-p)S^*(t) \quad t \geq 0 \quad (1.2)$$

$$f(t) = (1-p)f^*(t) \quad t \geq 0 \quad (2.2)$$

که در آن t زمان وقوع رخداد برای افراد تعریف می‌شود، و توابع $S^*(t)$ و $f^*(t)$ به ترتیب تابع بقا و تابع چگالی احتمال طول عمر افراد در معرض خطر هستند. همچنین p نسبت افراد شفایافته است.

برای به‌دست آوردن تابع چگالی (۲.۲) با استفاده از تابع بقا (۱.۲)، با توجه به این‌که می‌دانیم $S(t) = 1 - F(t)$ و

$$S'(t) = \frac{dS(t)}{dt} = \frac{-dF(t)}{dt} = -f(t).$$

پس

$$S^{*'}(t) = -f^*(t).$$

بنابراین می‌توان نوشت

$$S'(t) = (1-p) * (-f^*(t)) = -(1-p)f^*(t)$$

$$-f(t) = -(1-p)f^*(t)$$

$$f(t) = (1-p)f^*(t).$$

^{۱۰}Uncured

در رابطه (۱.۲) تابع S برخلاف تابع S^* یک تابع بقای ناسره است؛ یعنی

$$\lim_{t \rightarrow +\infty} S(t) = p > 0. \quad (3.2)$$

برای آن که درستی رابطه (۳.۲) را نشان دهیم، کافی است به جای $S(t)$ رابطه (۱.۲) را جاگذاری کنیم. داریم

$$\begin{aligned} \lim_{t \rightarrow +\infty} S(t) &= \lim_{t \rightarrow +\infty} (p + (1-p)S^*(t)) \\ &= \lim_{t \rightarrow +\infty} p + \lim_{t \rightarrow +\infty} (1-p)S^*(t) \end{aligned}$$

با توجه به آن که $\lim_{t \rightarrow +\infty} S^*(t) = 0$ ، پس

$$\lim_{t \rightarrow +\infty} S(t) = p > 0.$$

مزیتی که مدل نرخ شفایافتگی نسبت به سایر مدل‌های بقا دارد آن است که اجازه می‌دهد متغیرهای تبیینی رگرسیونی به هر دو قسمت نسبت شفایافتگی، یعنی p ، و تابع بقای بخش در معرض خطر، یعنی $S^*(t)$ ، مرتبط شوند. به عبارت دیگر، این مدل اجازه می‌دهد متغیرهای تبیینی تأثیرات مختلفی روی بیماران درمان‌شده و درمان‌نشده داشته باشند.

مدل‌های شفایافته آمیخته که توسط برکسون و گیج [۹] ارایه شدند، ابتدا در پزشکی برای مدل‌بندی بقای عمر بیماران مبتلا به سرطان مورد استفاده قرار گرفت. در این زمینه محققان دیگری مانند اودوس و همکاران [۲۸] و تانگ و همکاران [۳۴] فعالیت داشته‌اند.

۲.۳.۲ مدل نرخ شفایافتگی غیرآمیخته

مدل‌های نرخ شفایافتگی غیرآمیخته اولین بار توسط یاکولف [۳۸] ارایه شدند. سپس چن و همکاران [۱۱] مدل را بر اساس سلول‌های سرطانی معرفی کردند که به مدل نرخ شفایافتگی زمان پیشرفت^{۱۱} نیز معروف است. مزیت این رده از مدل‌های نرخ شفایافتگی نسبت به نسخه آمیخته آن، تعبیر ساده و قابل فهم‌تر آن از شفایافتگی بر اساس مکانیسم پیشرفت بیماری سرطان بر حسب افزایش سلول‌های سرطانی است. رودریگس و همکاران [۳۱] نیز این مدل را مورد بررسی قرار دادند. برای تشریح مدل از مثال سلول‌های سرطانی یک بافت زنده استفاده می‌کنیم.

فرض کنید N تعداد سلول‌های سرطانی فعال در بافت یک فرد پس از درمان باشد که از توزیع پواسن با میانگین θ پیروی می‌کند. زمان‌های تصادفی شروع فعالیت سلول سرطانی i ام را با Z_i ، $i = 0, 1, \dots, N$ ، نشان می‌دهیم و فرض می‌کنیم مستقل و هم‌توزیع با تابع توزیع $F(t) = 1 - S(t)$ و مستقل از N هستند. دقت داشته باشید $P(N = 0) \neq 0$ و برابر $e^{-\theta}$ است. اگر $P(Z_0 = \infty) = 1$ و متغیر تصادفی $0 \leq i \leq N$ ، $T = \min Z_i$ زمان بازگشت سرطان تعریف

^{۱۱}Promotion Time Cure Rate Model

شود، تابع بقای متغیر تصادفی T برای زمان بازگشت مجدد بیماری به صورت زیر به دست می‌آید:

$$\begin{aligned}
 S_p(t) &= P(T > t, N \geq \circ) \\
 &= P(N = \circ) + P(Z_1 > t, \dots, Z_N > t, N \geq 1) \\
 &= \frac{e^{-\theta} \theta^\circ}{\circ!} + P(Z_1 > t, \dots, Z_N > t | N \geq 1) P(N \geq 1) \\
 &= \exp(-\theta) + \sum_{k=1}^{\infty} S(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\
 &= \sum_{k=\circ}^{\infty} S(t)^k \frac{\theta^k}{k!} \exp(-\theta) \\
 &= \exp(-\theta) \exp(\theta S(t)) = \exp(-\theta F(t)). \quad (4.2)
 \end{aligned}$$

تابع چگالی احتمال مرتبط با مدل غیرآمیخته نیز به صورت زیر خواهد بود:

$$f_p(t) = \theta f(t) \exp(-\theta F(t)).$$

با توجه به آن که $F(\circ) = \circ$ این مدل شرط $S_p(\circ) = 1$ را داراست. اما با توجه به آن که $F(\infty) = 1$ ، واضح است که

$$\lim_{t \rightarrow +\infty} S_p(t) = \exp(-\theta \lim_{t \rightarrow +\infty} F(t)) = \exp(-\theta F(\infty)) = \exp(-\theta) \neq \circ.$$

بنابراین تابع بقای $S_p(t)$ یک تابع بقای ناسره است. کمیت $\exp(-\theta)$ نسبت شفایافتگی مدل محسوب می‌شود. یعنی نسبت افرادی که مستعد ابتلای مجدد به سرطان نیستند و در واقع درمان شده‌اند. زمانی که $\theta \rightarrow \infty$ نسبت شفایافتگی به سمت صفر و زمانی که $\theta \rightarrow \circ$ نسبت شفایافتگی به یک میل می‌کند.

در این مدل، افراد مبتلا به سرطان (شفانیافته یا در معرض خطر) دارای سلول‌های سرطانی اولیه بیش‌تر از صفر هستند. پس این افراد حداقل یک سلول سرطانی دارند. تابع بقای این افراد به صورت زیر محاسبه می‌شود:

$$S_p^*(t) = P(T > t, N \geq 1) = \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}. \quad (5.2)$$

برای اثبات رابطه (۵.۲)، می‌دانیم

$$S_p(t) = P(T > t, N \geq \circ) = \exp(-\theta F(t)). \quad (6.2)$$

بنابراین

$$\begin{aligned}
 S_p(t) &= P(T > t, N \geq \circ) \\
 &= P(N = \circ) + P(T > t, N \geq 1) \\
 &= P(N = \circ) + S_p^*(t). \quad (7.2)
 \end{aligned}$$

چون تابع بقا مقدار منفی اختیار نمی‌کند، پس

$$\begin{aligned} P(T > t, N \geq 1) &= P(N \geq 1)P(T > t | N \geq 1) \\ &= (1 - P(N = 0))P(T > t | N \geq 1) \end{aligned} \quad (۸.۲)$$

از (۷.۲) و (۸.۲) می‌توان نتیجه گرفت

$$\begin{aligned} S_p(t) &= P(N = 0) + (1 - P(N = 0))S_p^*(t) \\ S_p^*(t) &= \frac{S_p(t) - P(N = 0)}{1 - P(N = 0)} \end{aligned} \quad (۹.۲)$$

با جاگذاری رابطه (۶.۲) در (۹.۲) می‌توان نوشت

$$\begin{aligned} S_p^*(t) &= \frac{\exp(-\theta F(t)) - P(N = 0)}{(1 - P(N = 0))} \\ &= \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)}. \end{aligned}$$

تابع بقای (۵.۲) یک تابع بقای سره است، به این معنی که

$$S_p^*(0) = 1, \quad S_p^*(\infty) = 0.$$

تابع چگالی احتمال متناظر با تابع بقای (۵.۲) نیز به شکل

$$f^*(t) = -\frac{d}{dt}S^*(t) = \left(\frac{\exp(-\theta F(t))}{1 - \exp(-\theta)} \right) \theta f(t) \quad (۱۰.۲)$$

نوشته می‌شود.

در مساله مدل‌بندی بقای بازپرداخت وام‌های بانکی که در دو فصل ۳ و ۴ به آن می‌پردازیم، N را می‌توان تعداد اقساط پرداخت‌شده توسط مشتری تعبیر کرد. بنابراین $N = 0$ به معنی عدم پرداخت هیچ‌کدام از اقساط است که نتیجه آن وجود یک مشتری متقلب خواهد بود.

۴.۲ تابع درست‌نمایی مدل‌های نرخ شفایافتگی

در این بخش نحوه تشکیل و محاسبه تابع درست‌نمایی دو مدل معرفی‌شده را تشریح می‌کنیم

۱.۴.۲ تابع درست‌نمایی مدل آمیخته

فرض کنید در مدل نرخ شفایافتگی آمیخته (۱.۲)، θ بردار پارامترهای مدل است که می‌تواند پارامترهای تعریف‌شده برای نسبت شفایافتگی p و تابع بقای افراد در معرض خطر $S^*(t)$ را شامل شود. همچنین فرض کنید زمان‌های $t = (t_1, t_2, \dots, t_n)'$ زمان‌های رخداد پیشامد

مورد نظر یا سانسور راست برای همه افراد تحت مطالعه باشند و δ_i تابع نشانگر سانسور باشد که به صورت زیر تعریف می‌شود:

$$\delta_i = \begin{cases} 1 & \text{اگر فرد } i \text{ ام در معرض شکست باشد} \\ 0 & \text{اگر فرد } i \text{ ام در معرض سانسور باشد} \end{cases}$$

تابع درستنمایی برای این مدل با توجه به نوع سانسور از راست، به صورت زیر محاسبه می‌شود:

$$\begin{aligned} L(\theta|t_i) &= \prod_{i=1}^n f(t_i|\theta) \\ &= \prod_{i=1}^n f(t_i|\theta)^{\delta_i} S(t_i|\theta)^{1-\delta_i} \\ &= \prod_{i=1}^n \{(1-p)f^*(t)\}^{\delta_i} \{p + (1-p)S^*(t)\}^{1-\delta_i} \end{aligned}$$

که در آن، توابع $S^*(t)$ و $f^*(t)$ و پارامتر p می‌توانند به بردار پارامترهای θ وابسته باشند. احتمال $(1-p)f^*(t)$ سهم افرادی است که برای آن‌ها در زمان t رخداد مورد نظر اتفاق افتاده است و احتمال $p + (1-p)S^*(t)$ احتمال بقا برای افرادی است که در زمان t سانسور شده‌اند.

۲.۴.۲ تابع درستنمایی مدل غیرآمیخته

مشابه مدل آمیخته، بردار مشاهدات t و نشانگر سانسور $\delta = (\delta_1, \dots, \delta_n)'$ را در نظر بگیرید. مجموعه داده‌های مشاهده‌شده را با $D_{obs} = \{n, t, \delta\}$ و داده‌های کامل را با $D = \{n, t, \delta, N\}$ نمایش می‌دهیم، به طوری که N متغیر پنهان و دارای توزیع پواسن با پارامتر v است. همان‌طور که قبلاً ذکر شد در مدل شفایافتگی غیرآمیخته متغیر N_i تعداد سلول‌های سرطانی برای فرد i ام است و یک متغیر غیرقابل مشاهده می‌باشد. همچنین $Z_i, i = 1, 2, \dots, N$ ، متغیرهای تصادفی پنهانی هستند که مستقل و هم‌توزیع با $F(t|\gamma)$ و مستقل از N می‌باشند. تابع درستنمایی کامل بر اساس پارامترهای (v, γ, β) به صورت زیر به دست می‌آید:

$$\begin{aligned} L(v, \gamma, \beta|D_{obs}) &= f(D_{obs}|v, \gamma, \beta) = f(n, t, \delta, N|v, \gamma, \beta) \\ &= \prod_{i=1}^n f(t_i, \delta_i, N_i|v, \gamma, \beta) \\ &= \prod_{i=1}^n f(t_i|\delta_i, \gamma, N_i) P(N_i|v, \beta) \\ &= \prod_{i=1}^n f_p(t_i|\gamma, N_i)^{\delta_i} S_p(t_i|\gamma, N_i)^{1-\delta_i} \prod_{i=1}^n \frac{e^{-v} v^{N_i}}{N_i!} \end{aligned}$$

به‌طوری که

$$\begin{aligned} S_p(t_i|\gamma, N_i) &= P(T > t_i|\gamma, N_i) = P(Z_{i1} > t_i, Z_{i2} > t_i, \dots, Z_{iN} > t_i|\gamma, N_i) \\ &= P(Z_{i1} > t_i|\gamma, N_i)P(Z_{i2} > t_i|\gamma, N_i) \dots P(Z_{iN} > t_i|\gamma, N_i) \\ &= S(t_i|\gamma, N_i)^{N_i} \end{aligned}$$

9

$$f_p(t_i|\gamma, N_i) = -\frac{dS_p(t_i|\gamma, N_i)}{dt_i} = N_i S(t_i|\gamma, N_i)^{N_i-1} f(t_i|\gamma, N_i).$$

در نتیجه

$$\begin{aligned} L(v, \gamma, \beta|D_{obs}) &= \{(N_i S(t_i|\gamma)^{N_i-1} f(t_i|\gamma))^{\delta_i} (S(t_i|\gamma)^{N_i})^{1-\delta_i}\} \prod_{i=1}^n \frac{e^{-v} v^{N_i}}{N_i!} \\ &= \{(N_i f(t_i|\gamma))^{\delta_i} (S(t_i|\gamma))^{N_i \delta_i}\} \prod_{i=1}^n \frac{e^{-v} v^{N_i}}{N_i!} \\ &= \{S(t_i|\gamma)^{N_i-\delta_i} (N_i f(t_i|\gamma))^{\delta_i}\} \exp\left\{\sum_{i=1}^n ((N_i \log v) - \log(N_i!) - v)\right\}. \end{aligned}$$

فصل ۳

مدل نرخ شفایافتگی آماسیده در صفر برای بقای بازپرداخت وام‌ها

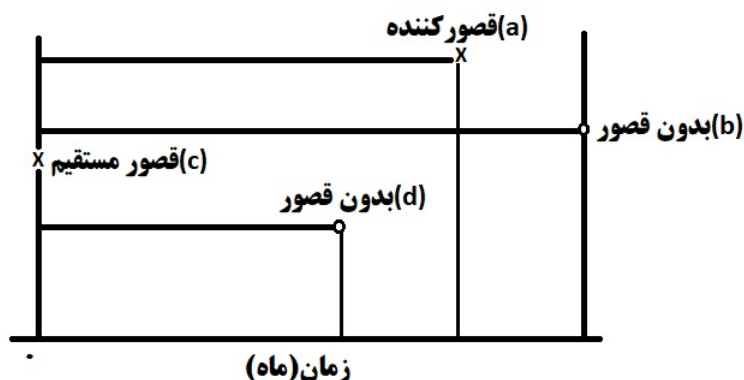
۱.۳ مقدمه

اغلب بانک‌ها و موسسات مالی وام‌هایی را به مشتریان خود اعطا می‌کنند که در مواردی ارتباط مشتری به‌طور کامل با بانک قطع شده و کل مبلغ اعطاشده از سوی بانک از دست می‌رود. مسلماً این گروه از وام‌گیرندگان برای بانک ضررهایی را تحمیل می‌کنند. به این دسته از مشتریان، متقلب گفته می‌شود. در واقع این مشتریان، تعداد بازپرداخت‌های برابر صفر دارند و به عبارت دیگر، زمان بقای بازپرداخت وام آن‌ها صفر است.

دسته دیگری از مشتریان وجود دارند که قادر به بازپرداخت بدهی‌های خود به بانک نیستند، و برای بانک مشکل‌ساز می‌شوند. البته این مشتریان برخلاف مشتریان متقلب مبلغی از اقساط خود را (نه به‌طور کامل) پرداخت کرده‌اند. به این نوع مشتریان قصورکننده می‌گویند. اغلب اوقات این دسته از مشتریان به دلیل مشکلات مالی موفق به پرداخت باقی‌مانده اقساط خود نمی‌شوند و در معرض خطر تقلب قرار می‌گیرند.

خوشبختانه برای بقای بانک‌ها و موسسات مالی، دسته سوم مشتریان که تمام اقساط خود را بازپرداخت می‌کنند، بیشترین فراوانی را دارا هستند. این مشتریان را با عنوان خوش حساب می‌شناسیم. این دسته از مشتریان به تعهدات خود وفادارند و قصور برای آن‌ها وجود ندارد.

در واقع این افراد نسبت به عدم بازپرداخت (تقلب) مصون (یا شفایافته) هستند. برای تشریح سه دسته مورد نظر، شکل ۱.۳ را در نظر بگیرید، مقدار c زمان بقا برای یک مشتری متقلب و برابر صفر است؛ زمان a برای قصورکننده‌ها به‌طور پیش‌فرض مثبت است. زمان d یک بازپرداخت زود هنگام را نشان می‌دهد، در صورتی که مدت زمان بازپرداخت وام برابر b در پایان دوره پرداخت گردیده است.



شکل ۱.۳: طول زمان بازپرداخت تسهیلات بانکی

بانک‌ها برای افزایش سود خود باید تعداد مشتریان متقلب و قصورکننده را به حداقل رسانده و تعداد مشتریان خوش حساب خود را افزایش دهد. از آنجایی که اطلاعات مدیران بانک‌ها و موسسات مالی در مورد مشتریان خود محدود است، نمی‌توانند به‌طور قطعی پیش‌بینی کنند کدام یک از مشتریان و چه نسبتی از آن‌ها جزو کدام یک از سه دسته مذکور هستند. در واقع هدف مورد نظر در این پایان‌نامه، شناسایی افرادی است که به بازپرداخت وام خود اهمیتی نمی‌دهند. بانک‌ها و موسسات مالی باید از دادن وام به این دسته از افراد اجتناب کنند. مدیران بانکی همواره امیدوارند که اطلاعات مشتریانی که به بازپرداخت وام خود اهمیتی نمی‌دهند در آینده‌ای نزدیک ثبت شود و به آن‌ها وامی اعطا نشود. همان‌طور که در مقاله لوزادا و همکاران [۲۲] به آن اشاره شده است، یکی از دلایل عمده استفاده از تحلیل بقا در مدل‌بندی ریسک اعتبارات به‌جای سایر تکنیک‌های مدل‌بندی، این است که علاوه بر نظارت بر ریسک اعتبار اسناد وام در طول زمان، می‌توان داده‌های سانسور شده را نیز وارد تحلیل و مشتریان را به دسته‌های خوب و بد تقسیم کرد. باریگا و همکارانش [۸] با تعریف اصطلاحات جدیدی از مدل‌های نرخ شفایافتگی، این مطالعات را در مدل‌بندی اعطای تسهیلات بانکی به کار بردند. اصطلاح شفایافتگی در مدل نرخ شفایافتگی، برای مساله مورد نظر ما، به عدم قصور در بازپرداخت تسهیلات اعتباری برمی‌گردد و از آن به مدل نرخ عدم قصور نیز یاد می‌کنند.

تمام تسهیلات بانکی (وام‌ها) از بدو تأمین مالی مورد کنترل و نظارت قرار می‌گیرند. بنابراین زمان اولیه $t = 0$ تاریخ شروع وام را نشان می‌دهد. اما با توجه به این که ثبت عدم بازپرداخت تسهیلات معمولاً حداقل سه ماه متوالی طول می‌کشد، به این منظور حداقل سه ماه عدم بازپرداخت در شروع وام برای ثبت به‌عنوان $t = 0$ لازم است. این به آن معنی است که اگر در داده‌های بقای بازپرداخت تسهیلات $t \leq 3$ باشد، آن را برابر $t = 0$ در نظر می‌گیریم. همان‌طور که در مدل‌های بقا و نرخ شفایافتگی مطرح شد، تابع بقا در نقطه شروع زمان، یعنی $t = 0$ ، برابر یک است. به عبارت دیگر

$$S(0) = 1 - F(0) = 1.$$

اما برای داده‌های بانکی که تشریح شدند، احتمال بقا در $t = 0$ کمتر از یک است. یا به عبارت دیگر

$$F(0) > 0.$$

بنابراین مدل نرخ شفایافتگی آمیخته یا غیرآمیخته مناسب مدل‌بندی طول زمان عدم بازپرداخت تسهیلات بانکی نیست. در واقع در این داده‌ها طول زمان برابر صفر نیز وجود دارد و به عبارتی داده‌ها در صفر می‌توانند آماسیده^۱ باشند. از این رو باید مدل نرخ شفایافتگی برای در نظر گرفتن این آماسیدگی در صفر تعمیم می‌یافت. مدل نرخ شفایافتگی غیرآمیخته آماسیده در صفر توسط لوزادا و همکاران [۲۰] و نسخه آمیخته آن توسط لوزادا و همکاران [۲۱] معرفی شدند و از آن‌ها برای برآورد نرخ تقلب در بازپرداخت تسهیلات بانکی استفاده کردند. در ادامه این مدل و جزییات نظری مورد نیاز را ارایه می‌کنیم.

۲.۳ مدل نرخ شفایافتگی آماسیده در صفر

اشاره کردیم مشکلی که مدل نرخ شفایافتگی دارد، آن است که با حضور صفرها (مشتریان متقلب) سازگار نیست. برای رفع این مشکل، لوزادا و همکارانش [۲۰] مدل نرخ شفایافتگی آماسیده در صفر را پیشنهاد دادند که بهبودیافته مدل نرخ شفایافتگی است. مدل‌های آماسیده در صفر مدل‌های آمیخته‌ای هستند که یک مولفه شمارشی و جرم نقطه‌ای در صفر را ترکیب می‌کنند و تابع بقای آن‌ها به صورت زیر تعریف می‌شود:

$$S(t) = \gamma_1 + (1 - \gamma_0 - \gamma_1)S^*(t) \quad t \geq 0.$$

با توجه به این که

$$\begin{aligned} S(t) &= 1 - F(t) \\ S'(t) &= \frac{dS(t)}{dt} = \frac{-dF(t)}{dt} = -f(t) \\ S^{*'}(t) &= -f^*(t). \end{aligned}$$

^۱Inflated

بنابراین

$$\begin{aligned} S'(t) &= (1 - \gamma_0 - \gamma_1)S^{*'}(t) \\ -f(t) &= -(1 - \gamma_0 - \gamma_1)f^*(t). \end{aligned}$$

پس

$$f(t) = (1 - \gamma_0 - \gamma_1)f^*(t).$$

از آن جا که $S^*(\infty) = 0$ ، بنابراین

$$\begin{aligned} \lim_{t \rightarrow \infty} S(t) &= \lim_{t \rightarrow \infty} (\gamma_1 + (1 - \gamma_0 - \gamma_1)S^*(t)) \\ &= \lim_{t \rightarrow \infty} \gamma_1 + \lim_{t \rightarrow \infty} (1 - \gamma_0 - \gamma_1)S^*(t) \\ &= \lim_{t \rightarrow \infty} \gamma_1 + (1 - \gamma_0 - \gamma_1) \lim_{t \rightarrow \infty} S^*(t) \\ &= \gamma_1 > 0. \end{aligned}$$

هم‌چنین به دلیل این که $S^*(0) = 1$ ، پس

$$\begin{aligned} S(0) &= \gamma_1 + (1 - \gamma_0 - \gamma_1)S^*(0) \\ &= \gamma_1 + 1 - \gamma_0 - \gamma_1 \\ &= 1 - \gamma_0 < 1. \end{aligned}$$

بنابراین γ_1 نسبت مشتریان قصورکننده و γ_0 نسبت مشتریان متقلب است. توجه داشته باشید اگر $\gamma_0 = 0$ باشد، مدل مذکور به همان مدل نرخ شفایافتگی معمولی تبدیل می‌شود. اکنون از روی این مدل به سادگی می‌توان دریافت که اگر در بین مشتریان، متقلب داشته باشیم، در مدل معمولی نرخ شفایافتگی چون فراوانی صفرها در نظر گرفته نمی‌شود، نتایج استنباط معتبر نبوده و می‌توانند گمراه‌کننده باشند؛ مثلاً خطای استاندارد برآوردگرها بیش برآورد می‌شود.

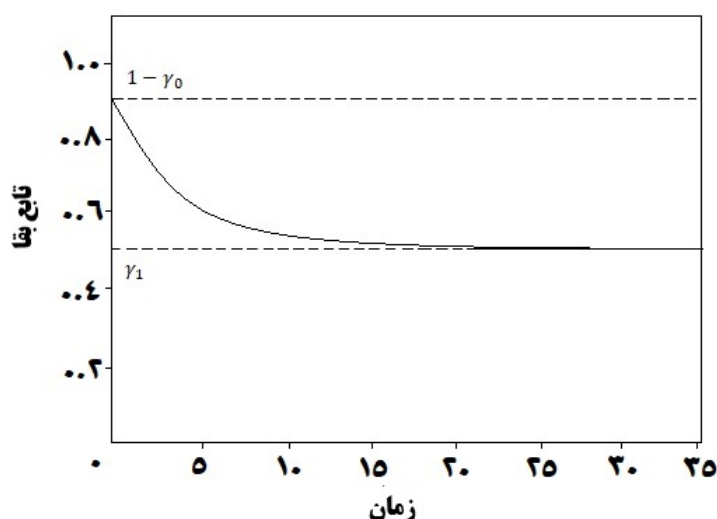
۳.۳ مدل رگرسیونی پیشنهادی

با توجه به مدل نرخ شفایافتگی آمیخته (۱.۲) و مدل غیرآمیخته (۴.۲) و با متحد کردن این دو، می‌توان مدل غیرآمیخته را به صورت یک مدل آمیخته به صورت

$$S(t) = \exp(-\theta) + (1 - \exp(-\theta))S^*(t) \quad t \geq 0 \quad (1.3)$$

بازنویسی کرد، که در آن $S^*(t)$ تابع بقای (۵.۲) است. چون هدف مطالعه نسبت مشتریان متقلب در کنار قصورکننده است، باید مدل (۱.۳) به نحوی تعمیم داده شود که نسبت مشتریان متقلب و قصورکننده هر دو در مدل لحاظ شوند. برای این منظور، مدل

$$S(t) = \gamma_1 + (1 - \gamma_0 - \gamma_1)S^*(t) \quad t \geq 0 \quad (2.3)$$



شکل ۲.۳: تابع بقا مدل نرخ شفایافتگی آماسیده در صفر

توسط لوزادا و همکاران [۲۰] معرفی شد، که در آن $\gamma_0 = \exp(-\kappa)$ و $\gamma_1 = \exp(-\theta)$. مدل معرفی شده (۲.۳) به مدل نرخ شفایافتگی آماسیده در صفر معروف شد و برای داده‌های تسهیلات بانکی مناسب است. برای آن که تاثیر برداری از متغیرهای تبیینی مانند سن مشتری، میزان هر قسط، تعداد اقساط، و نوع وام را بر متقلب یا قصورکننده بودن مشتری بررسی کنیم، باید از چارچوب مدل‌های رگرسیونی استفاده کنیم. با توجه به آن که دامنه تغییرات γ_0 و γ_1 بازه $(0, 1)$ است، باید از چارچوب مدل‌های خطی تعمیم‌یافته برای مدل‌بندی آن‌ها استفاده کنیم. در این مدل‌ها برای پارامترهایی که در $(0, 1)$ قرار می‌گیرند، معمولاً از تابع پیوند لجیت استفاده می‌شود. فرض کنید $X = (X'_1, X'_2)'$ بردار متغیرهای تبیینی با بعد $p = q_1 + q_2$ باشد که به دو زیربردار X_1 با بعد q_1 و X_2 با بعد q_2 قابل تجزیه است. برای اطمینان از آن که γ_0, γ_1 و همچنین $(1 - \gamma_0 - \gamma_1)$ در $(0, 1)$ قرار می‌گیرند، برای هر مشاهده $i = 1, \dots, n$ و با استفاده از تابع پیوند لجیت، تعریف می‌کنیم

$$\kappa_i = -\log\left(\frac{e^{X'_{1i}\beta_\kappa}}{1 + e^{X'_{1i}\beta_\kappa} + e^{X'_{2i}\beta_\theta}}\right) \quad (3.3)$$

$$\theta_i = -\log\left(\frac{e^{X'_{2i}\beta_\theta}}{1 + e^{X'_{1i}\beta_\kappa} + e^{X'_{2i}\beta_\theta}}\right) \quad (4.3)$$

که در آن $\beta_\kappa \in \mathbb{R}^{q_1}$ یک بردار q_1 بعدی از ضرایب رگرسیونی است که تاثیر متغیرهای تبیینی بر نرخ مشتریان متقلب را نشان می‌دهد و $\beta_\theta \in \mathbb{R}^{q_2}$ یک بردار q_2 بعدی از ضرایب رگرسیونی است که اثر متغیرهای تبیینی را بر نسبت مشتریان قصورکننده بیان می‌کند.

می‌دانیم $\gamma_0 = \exp(-\kappa)$. با توجه به رابطه (۳.۳)، می‌توان مقدار γ_0 را به صورت زیر به دست آورد:

$$\begin{aligned}\gamma_0 &= \exp(-(-\log(\frac{e^{X'_1\beta_\kappa}}{1 + e^{X'_1\beta_\kappa} + e^{X'_1\beta_\theta}}))) \\ &= \frac{e^{X'_1\beta_\kappa}}{1 + e^{X'_1\beta_\kappa} + e^{X'_1\beta_\theta}}.\end{aligned}\quad (۵.۳)$$

به همین ترتیب با توجه به $\gamma_1 = \exp(-\theta)$ و رابطه (۴.۳)، γ_1 به صورت زیر محاسبه می‌شود:

$$\begin{aligned}\gamma_1 &= \exp(-(-\log(\frac{e^{X'_2\beta_\theta}}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}}))) \\ &= \frac{e^{X'_2\beta_\theta}}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}}.\end{aligned}\quad (۶.۳)$$

چون γ_0 و γ_1 بر اساس تابع پیوند لجیت تعریف شده‌اند، واضح است که همواره در بازه $(0, 1)$ قرار می‌گیرند. دلیل تعریف کردن γ_0 و γ_1 در (۵.۳) و (۶.۳)، تضمین قرار گرفتن $(1 - \gamma_0 - \gamma_1)$ در بازه $(0, 1)$ است. برای نمایش این ضمانت، با توجه به روابط (۵.۳) و (۶.۳)، می‌توان نشان داد

$$\begin{aligned}1 - \gamma_0 - \gamma_1 &= 1 - \frac{e^{X'_1\beta_\kappa}}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}} - \frac{e^{X'_2\beta_\theta}}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}} \\ &= \frac{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta} - e^{X'_1\beta_\kappa} - e^{X'_2\beta_\theta}}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}} \\ &= \frac{1}{1 + e^{X'_1\beta_\kappa} + e^{X'_2\beta_\theta}}.\end{aligned}$$

بنابراین $(1 - \gamma_0 - \gamma_1)$ نیز در $(0, 1)$ قرار دارد.

۱.۳.۳ مدل پارامتری برای تابع بقای مشتریان خوش حساب

برای تکمیل مدل (۲.۳) باید تابع بقای مشتریان خوش حساب، $S^*(t)$ ، را مشخص کنیم. در واقع در این جا، مشتریان خوش حساب، افرادی هستند که نسبت به رویداد مورد علاقه، یعنی عدم بازپرداخت تسهیلات، مصون یا شفایافته نیستند. برای این منظور از یک مدل پارامتری وایبل استفاده می‌کنیم.

فرض کنید $f(t)$ و $F(t)$ به ترتیب تابع چگالی احتمال و تابع توزیع تجمعی وایبل باشند:

$$\begin{aligned}F(t) &= 1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha} \\ f(t) &= \frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha}.\end{aligned}$$

ویژگی‌های توزیع وایبل با دو پارامتر شکل $\alpha > 0$ و مقیاس $\lambda > 0$ ، آن را به یکی از مدل‌های پارامتری مناسب برای تحلیل بقا و قابلیت اطمینان تبدیل کرده‌اند. بنابراین فرض می‌کنیم

$S^*(t)$ تابع بقای توزیع وایبل است. بنابراین، با جمع‌بندی همه جزئیات مطرح‌شده، مدل نرخ شفایافتگی آماسیده در صفر به صورت زیر ارایه می‌شود:

$$S(t) = \exp(-\theta) + (1 - \exp(-\kappa) - \exp(-\theta))S^*(t) \quad (۷.۳)$$

که در آن

$$\begin{aligned} S^*(t) &= \frac{\exp(-\theta F(t)) - \exp(-\theta)}{1 - \exp(-\theta)} \\ f^*(t) &= \left(\frac{\exp(-\theta F(t))}{1 - \exp(-\theta)} \right) \theta f(t) \\ F(t) &= 1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha} \\ f(t) &= \frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha} \end{aligned}$$

به‌طوری که κ و θ در (۳.۳) و (۴.۳) معرفی شده‌اند. در مدل (۷.۳)، بردار پارامترهای نامعلوم را با $\vartheta = (\alpha, \lambda, \beta_\kappa, \beta_\theta)$ نمایش می‌دهیم.

۴.۳ برازش مدل و استنباط آماری

برای برازش مدل و برآورد پارامترهای آن از رهیافت مبتنی بر درست‌نمایی استفاده می‌کنیم. با توجه به بازه تحت مطالعه، ممکن است برخی از مشاهدات سانسور شده (از راست) باشند. اگر زمان بقای مشتری i ام را با t_i مشخص کنیم، آن‌گاه سهم هر مشاهده در تابع درست‌نمایی به سه بخش تقسیم می‌شود:

- اگر $t_i = 0$ آن‌گاه سهم مشاهده i در تابع درست‌نمایی برابر است با γ_{i0} .
- اگر مشاهده t_i سانسور شده نباشد، آن‌گاه سهم آن در تابع درست‌نمایی برابر است با $(1 - \gamma_{i0} - \gamma_{i1})f^*(t_i)$.
- و اگر مشاهده t_i از راست سانسور شده باشد، سهم آن در تابع درست‌نمایی برابر است با $\gamma_{i1} + (1 - \gamma_{i0} - \gamma_{i1})S^*(t_i)$.

فرض کنید $D = \{t_i, \delta_i, x_{1i}, x_{2i}\}_{i=1}^n$ مجموعه داده‌ها باشد که در آن δ_i متغیر تصادفی نشانگر سانسور است، به‌طوری که $\delta_i = 1$ اگر t_i زمان رخداد مشاهده‌شده (تقلب یا قصور) باشد و $\delta_i = 0$ اگر t_i از راست سانسور شده باشد.

برای نوشتن تابع درست‌نمایی، $I(t_i > 0)$ را تابع نشانگر تعریف می‌کنیم به این معنی که برای $t_i > 0$ برابر یک و در سایر جاها صفر است. بنابراین تابع درست‌نمایی برای همه مشاهدات به صورت

$$L(\vartheta|D) \propto \prod_{i=1}^n \{ (1 - I(t_i > 0))\gamma_{i0} + I(t_i > 0)[(1 - \gamma_{i0} - \gamma_{i1})f^*(t_i)]^{\delta_i} [\gamma_{i1} + (1 - \gamma_{i0} - \gamma_{i1})S^*(t_i)]^{1-\delta_i} \}$$

است. این تابع را می‌توان به صورت زیر نیز نوشت:

$$L(\boldsymbol{\vartheta}|D) = \prod_{i:t_i=0} \gamma_0 \prod_{i:t_i>0} \{[(1 - \gamma_0 - \gamma_1)f_w^*(t_i)]^{\delta_i} [\gamma_1 + (1 - \gamma_0 - \gamma_1)S_w^*(t_i)]^{1-\delta_i}\}.$$

بنابراین با لگاریتم گرفتن از تابع درستنمایی، می‌توان نوشت

$$\begin{aligned} \log\{L(\boldsymbol{\vartheta}|D)\} &= \sum_{i:t_i=0} \log(\gamma_0) + \sum_{i:t_i>0} \log\{[(1 - \gamma_0 - \gamma_1)f_w^*(t_i)]^{\delta_i}\} \\ &+ \sum_{i:t_i>0} \log\{[\gamma_1 + (1 - \gamma_0 - \gamma_1)S_w^*(t_i)]^{1-\delta_i}\} \\ &= \sum_{i:t_i=0} \log(\gamma_0) + \sum_{i:t_i>0} \delta_i \log\{(1 - \gamma_0 - \gamma_1)\} + \sum_{i:t_i>0} \delta_i \log[f_w^*(t_i)] \\ &+ \sum_{i:t_i>0} (1 - \delta_i) \log\{[\gamma_1 + (1 - \gamma_0 - \gamma_1)S_w^*(t_i)]\}. \end{aligned}$$

از طرفی با توجه به روابط

$$\begin{aligned} f_w^*(t_i) &= \frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha} \\ F_w^*(t_i) &= 1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha} \\ S_w^*(t_i) &= 1 - F_w^*(t_i) \\ &= 1 - (1 - e^{-\left(\frac{t}{\lambda}\right)^\alpha}) \\ &= e^{-\left(\frac{t}{\lambda}\right)^\alpha} \end{aligned}$$

می‌توان نتیجه گرفت

$$\begin{aligned} \log\{L(\boldsymbol{\vartheta}|D)\} &= \sum_{i:t_i=0} \log(\gamma_0) + \sum_{i:t_i>0} \delta_i \log\{(1 - \gamma_0 - \gamma_1)\} \\ &+ \sum_{i:t_i>0} \delta_i \log\left[\frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^\alpha}\right] \\ &+ \sum_{i:t_i>0} (1 - \delta_i) \log\{[\gamma_1 + (1 - \gamma_0 - \gamma_1)e^{-\left(\frac{t}{\lambda}\right)^\alpha}]\}. \end{aligned}$$

اکنون با توجه به معادله‌های

$$\begin{aligned} \gamma_0 &= \frac{e^{\mathbf{x}'_1 \beta_\kappa}}{1 + e^{\mathbf{x}'_1 \beta_\kappa} + e^{\mathbf{x}'_1 \beta_\theta}} \\ \gamma_1 &= \frac{e^{\mathbf{x}'_1 \beta_\theta}}{1 + e^{\mathbf{x}'_1 \beta_\kappa} + e^{\mathbf{x}'_1 \beta_\theta}} \\ 1 - \gamma_0 - \gamma_1 &= \frac{1}{1 + e^{\mathbf{x}'_1 \beta_\kappa} + e^{\mathbf{x}'_1 \beta_\theta}} \end{aligned}$$

و جایگذاری آن‌ها در تابع درستنمایی، تابع لگاریتم درستنمایی به صورت زیر به دست می‌آید:

$$\begin{aligned} \log\{L(\boldsymbol{\vartheta}|D)\} &= \sum_{i:t_i=0} \log\left(\frac{e^{x'_{i1}\beta_{\kappa}}}{1 + e^{x'_{i1}\beta_{\kappa}} + e^{x'_{i1}\beta_{\theta}}}\right) + \sum_{i:t_i>0} \log\left\{\frac{1}{1 + e^{x'_{i1}\beta_{\kappa}} + e^{x'_{i1}\beta_{\theta}}}\right\} \\ &+ \sum_{i:t_i>0} \delta_i \log\left[\frac{\alpha}{\lambda} \left(\frac{t}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}\right] \\ &+ \sum_{i:t_i>0} (1 - \delta_i) \log\left\{\left[\frac{e^{x'_{i1}\beta_{\theta}}}{1 + e^{x'_{i1}\beta_{\kappa}} + e^{x'_{i1}\beta_{\theta}}} + \left(\frac{1}{1 + e^{x'_{i1}\beta_{\kappa}} + e^{x'_{i1}\beta_{\theta}}}\right) e^{-\left(\frac{t}{\lambda}\right)^{\alpha}}\right]\right\}. \end{aligned}$$

برآوردهای ماکسیمم درستنمایی $\boldsymbol{\vartheta}$ ، که با $\hat{\boldsymbol{\vartheta}}$ نمایش می‌دهیم، با ماکسیمم کردن تابع درستنمایی $L(\boldsymbol{\vartheta}|D)$ در فضای پارامتر به دست می‌آیند. با توجه به پیچیده بودن تابع درستنمایی مذکور، برای $\hat{\boldsymbol{\vartheta}}$ شکل بسته وجود ندارد و باید آن را به صورت تقریبی و با روش‌های عددی مثل BFGS محاسبه کرد.

توزیع مجانبی برآوردهای ماکسیمم درستنمایی $\hat{\boldsymbol{\vartheta}}$ ، تحت شرایطی، نرمال با میانگین $\boldsymbol{\vartheta}$ و ماتریس واریانس $I^{-1}(\hat{\boldsymbol{\vartheta}})$ است، که در آن $I(\boldsymbol{\vartheta})$ ماتریس اطلاع فشر مدل است [۲۴]. ماتریس I را می‌توان با معکوس ماتریس هسیان $I(\boldsymbol{\vartheta}) = \left\{\frac{-\partial^2 \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'}\right\}^{-1}$ مشاهده شده که در $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$ مقداردهی شده است برآورد کرد، که در آن $\ell(\boldsymbol{\vartheta})$ تابع لگاریتم درستنمایی مدل است. با استفاده از توزیع مجانبی نرمال برآوردهای درستنمایی ماکسیمم، می‌توان استنباط آماری تقریبی، شامل محاسبه فواصل اطمینان، آزمون فرضیه‌ها، و انتخاب مدل، را برای مدل برازش شده انجام داد. مهم‌تر از این، نسبت‌های مشتریان خوش حساب، قصورکننده و متقلب به ازای هر سطح از اطلاعات متغیرهای تبیینی قابل برآورد هستند که می‌تواند برای یک بانک اطلاعات ارزشمندی محسوب شود. افزون بر این تابع بقای مشتریان خوش حساب به عنوان بخش کلیدی تحلیل داده‌های تسهیلات بانکی، قابل برآورد و استفاده است.

فصل ۴

تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

در این فصل، عملکرد مدل نرخ شفایافتگی آماسیده در صفر را با یک مطالعه شبیه‌سازی گسترده ارزیابی می‌کنیم. همچنین کاربست مدل را برای تحلیل یک مجموعه داده واقعی در مورد طول زمان بازپرداخت تسهیلات اعطاشده توسط چهار بانک داخل کشور و شناسایی عوامل موثر بر رفتار بازپرداخت مشتریان، نشان می‌دهیم. در پایان نیز نتیجه‌گیری نهایی پایان‌نامه و آینده تحقیق را مطرح می‌کنیم.

۱.۴ مطالعه شبیه‌سازی

هدف از مطالعه شبیه‌سازی، بررسی رفتار برآوردهای ML در مدل نرخ شفایافتگی آماسیده در صفر (معرفی‌شده در فصل ۳) است. منظور از رفتار برآوردهای ML پارامترهای مدل، که به‌طور عددی به‌دست می‌آیند، سازگاری و نزدیک بودن تقریب توزیع مجانبی نرمال به توزیع واقعی برآوردها است. در این مطالعه

۴۲ تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

- برای بررسی سازگاری برآوردهای ML پارامترهای مدل، ریشه میانگین توان دوم خطا^۱ (RMSE) را محاسبه و همگرایی آن به صفر را، با افزایش حجم نمونه، بررسی کردیم.
 - برای بررسی دقت توزیع مجانبی نرمال و اعتبارسنجی استنباط‌های ممکن حاصل از مدل برازش‌شده، نرخ پوشش تجربی^۲ فواصل اطمینان نتیجه‌شده از تقریب نرمال را برای هر کدام از پارامترهای مدل محاسبه و نزدیکی آن به سطح اطمینان اسمی^۳ را ارزیابی کردیم. در تمام شبیه‌سازی‌ها سطح اطمینان اسمی را برابر ۹۵٪ در نظر گرفتیم.
- در این مطالعه، برای همه تنظیمات شبیه‌سازی، از مدل نرخ شفایافتگی آماسیده در صفر ۱۰۰۰ مجموعه داده برای هر کدام از حجم‌های نمونه $n = ۱۰۰, ۲۵۰, ۵۰۰, ۷۵۰, ۱۰۰۰$ تولید کردیم. برآوردهای ML پارامترها و مقادیر خطای استاندارد متناظر آن‌ها نیز به روش عددی BFGS به‌دست آمدند. این روش توسط تابع `optim()` در نرم‌افزار R قابل پیاده‌سازی است.
- سه مثال شبیه‌سازی مختلف را برای مدل نرخ شفایافتگی (بدون قصور) آماسیده در صفر با تابع بقای پارامتری وایبل، در نظر گرفتیم. در هر سه مثال یک متغیر تبیینی رگرسیونی x را از یک توزیع برنولی با پارامتر موفقیت ۰/۵ تولید کردیم که از آن می‌توان به‌عنوان ویژگی‌های مشتریان، مثل جنسیت آن‌ها یا گرفتن یا نگرفتن تسهیلات دیگر از سایر موسسات مالی، تعبیر کرد.

پارامترهای مدل را به صورت

$$\begin{aligned}\gamma_{0i} &= \frac{e^{\beta_{10} + x_i \beta_{11}}}{1 + e^{\beta_{10} + x_i \beta_{11}} + e^{\beta_{20} + x_i \beta_{21}}} \\ \gamma_{1i} &= \frac{e^{\beta_{20} + x_i \beta_{21}}}{1 + e^{\beta_{10} + x_i \beta_{11}} + e^{\beta_{20} + x_i \beta_{21}}} \\ \alpha_i &= e^{\beta_{30} + x_i \beta_{31}} \\ \lambda_i &= e^{\beta_{40} + x_i \beta_{41}}\end{aligned}\quad (۱.۴)$$

در نظر گرفتیم. با استفاده از پارامترهای تعیین‌شده در (۱.۴)، سه سناریوی مختلف برای مقادیر واقعی پارامترها در هر سه مثال شبیه‌سازی تعریف کردیم:

۱. برای سناریوی اول، $\beta_{10} = -۳$ و $\beta_{11} = ۱$ تعیین شدند. همچنین β_{20} برابر ۲/۵- و β_{21} برابر ۰/۳ انتخاب شدند. با توجه به این که میانگین x برابر ۰/۵ است، با تعیین این مقادیر برای پارامترهای $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}$ ، مقادیر γ_0 و γ_1 در مدل نرخ شفایافتگی آماسیده در صفر به ترتیب به‌طور متوسط برابر ۰/۰۶۹۷ و ۰/۰۸۰۹ نتیجه شدند. در مقایسه با سناریوهای ۲ و ۳، این سناریو دارای نرخ قصور و بدون قصور کوچک‌تری است که به ترتیب برابر ۶/۹۷٪ و ۸/۰۹٪ به‌دست می‌آیند. برای پارامترهای رگرسیونی موجود در

^۱ Root Mean Squared Error

^۲ Empirical Coverage Rate

^۳ Nominal Confidence Level

پارامترهای تابع بقای وایبل نیز مقادیر واقعی $(\beta_{30}, \beta_{31}, \beta_{40}, \beta_{41}) = (0.5, 0.5, 1.5, 2)$ انتخاب شدند. با این انتخاب و بنا به مقدار میانگین 0.5 برای متغیر تبیینی x ، پارامترهای α و λ در توزیع وایبل به‌طور متوسط به ترتیب برابر $2/11$ و $2/18$ نتیجه می‌شوند.

۲. برای سناریوی دوم، $\beta_{10} = -2$ و $\beta_{11} = 2$ انتخاب شدند. همچنین β_{20} برابر $1/5$ و β_{21} برابر $1/5$ در نظر گرفته شدند. با این انتخاب‌ها، مقادیر γ_0 و γ_1 به ترتیب به‌طور متوسط برابر 0.1999 و 0.2566 نتیجه شدند. بنابراین، در این سناریو نرخ قصور و بدون قصور به ترتیب برابر 19.99% و 25.66% به‌دست می‌آیند. برای پارامترهای رگرسیونی موجود در پارامترهای تابع بقای وایبل نیز مقادیر واقعی $(\beta_{30}, \beta_{31}, \beta_{40}, \beta_{41}) = (-0.5, 1.5, -0.5, 3)$ انتخاب شدند. با این انتخاب‌ها، پارامترهای α و λ در توزیع وایبل به‌طور متوسط به ترتیب برابر $1/28$ و $2/71$ نتیجه می‌شوند.

۳. برای سناریوی سوم، $\beta_{10} = -0.5$ و $\beta_{11} = 0.75$ انتخاب شدند. همچنین β_{20} برابر -0.35 و β_{21} برابر $1/75$ در نظر گرفته شدند. با این انتخاب‌ها، مقادیر γ_0 و γ_1 به ترتیب به‌طور متوسط برابر 0.2469 و 0.7832 به‌دست آمدند. بنابراین، در این سناریو نرخ قصور و بدون قصور به ترتیب برابر 24.69% و 78.32% به‌دست می‌آیند که در مقایسه با دو سناریوی قبلی، دارای نرخ‌های بزرگ‌تری است. برای پارامترهای رگرسیونی موجود در پارامترهای تابع بقای وایبل نیز مقادیر واقعی $(\beta_{30}, \beta_{31}, \beta_{40}, \beta_{41}) = (-1, 2, 1/25, 3/5)$ انتخاب شدند. با این انتخاب‌ها، پارامترهای α و λ در توزیع وایبل به‌طور متوسط به ترتیب برابر 1 و $20/8$ نتیجه می‌شوند.

در نظر گرفتن این سه سناریو برای لحاظ کردن انواع شرایط تقلب و قصوری است که ممکن است در واقعیت دیده شوند.

۱.۱.۴ الگوریتم شبیه‌سازی

برای شبیه‌سازی یک مجموعه داده به حجم n از مدل نرخ شفاف‌افتگی آماسیده در صفر با تنظیمات مطرح‌شده، از الگوریتم زیر استفاده می‌کنیم. تاکید می‌کنیم در این مدل γ_0 نسبت آماسیدگی در صفر (یا همان نسبت مشتریان متقلب)، γ_1 نسبت مشتریان بدون قصور و $1 - \gamma_1 - \gamma_0$ نسبت قصورکننده‌ها با تابع بقای وایبل با پارامترهای α و λ است.

الگوریتم تولید نمونه از مدل نرخ شفاف‌افتگی آماسیده در صفر

ضرایب رگرسیونی β_{10} و β_{11} مربوط به زمان آماسیدگی در صفر، برای γ_0 هستند؛ ضرایب رگرسیونی β_{20} و β_{21} نرخ قصور، یعنی γ_1 ، را کنترل می‌کنند. همچنین ضرایب β_{30} و β_{31} مربوط به پارامتر α در توزیع وایبل هستند و ضرایب رگرسیونی β_{40} و β_{41} مربوط به پارامتر λ هستند.

۴۴ تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

برای شبیه‌سازی داده‌ها باید مراحل الگوریتم زیر به تعداد n بار تکرار شوند. برای مرحله $i = 1, \dots, n$

۱. مشاهده x_i را از توزیع برنولی با پارامتر موفقیت γ_i تولید کنید. با تعریف پارامترهای شبیه‌سازی، مقادیر γ_i ، α_i و λ_i را محاسبه کنید.

۲. مقدار u_i را از توزیع یکنواخت استاندارد تولید کنید.

۳. اگر $u_i \leq \gamma_i$ ، آن‌گاه قرار دهید $s_i = 0$.

۴. اگر $u_i > 1 - \gamma_i$ ، آن‌گاه قرار دهید $s_i = \infty$.

۵. اگر $\gamma_i < u_i \leq 1 - \gamma_i$ ، آن‌گاه v_i را از توزیع یکنواخت $Unif(\gamma_i, 1 - \gamma_i)$ تولید کنید و s_i را به‌عنوان ریشه $F(s_i) - v_i = 0$ محاسبه کنید، که در آن F تابع توزیع وایبل با پارامترهای α_i و λ_i است.

۶. مقدار w_i را از توزیع یکنواخت $Unif(0, \max(s_i))$ تولید کن.

۷. تعریف کنید $t_i = \min(s_i, w_i)$. اگر $t_i < w_i$ ، قرار دهید $\delta_i = 1$ و در غیر این‌صورت قرار دهید $\delta_i = 0$.

۸. مشاهدات (x_i, t_i, δ_i) را به‌عنوان خروجی گزارش کنید.

برای شناسایی‌پذیر^۴ شدن مدل، در شبیه‌سازی‌ها از یک توزیع سانسور یکنواخت با دامنه محدود استفاده کردیم برای اطلاعات بیش‌تر در این مورد مقاله روچا و همکاران [۳۰] را ببینید.

۲.۱.۴ نتایج شبیه‌سازی مونت کارلویی

شکل‌های ۱.۴، ۲.۴ و ۳.۴ نتایج شبیه‌سازی را برای سناریوی اول انتخاب پارامترهای مدل نشان می‌دهند. به‌طور مشابه شکل‌های ۴.۴، ۵.۴ و ۶.۴ نتایج مربوط به سناریوی دوم انتخاب پارامترها هستند و شکل‌های ۷.۴، ۸.۴ و ۹.۴ نتایج مربوط به سناریوی سوم انتخاب پارامترها هستند. این نتایج برای حجم‌های نمونه $n = 100, 250, 500, 750, 1000$ و براساس تولید ۱۰۰۰ مجموعه داده برای هر سناریو به‌دست آمده‌اند. با توجه به نتایج حاصل می‌توان مدعی شد

۱. برای تمام برآوردهای ML پارامترهای مدل، آریبی و RMSE با افزایش حجم نمونه به صفر همگرا می‌شوند. بنابراین سازگاری برآوردها قابل نتیجه‌گیری است. همچنین نرخ پوشش تجربی به سطح اطمینان اسمی ۹۵٪ نزدیک می‌شوند که حاکی از تقریب خوب توزیع مجانبی نرمال به‌عنوان توزیع برآوردهای ML است.

^۴Identifiability

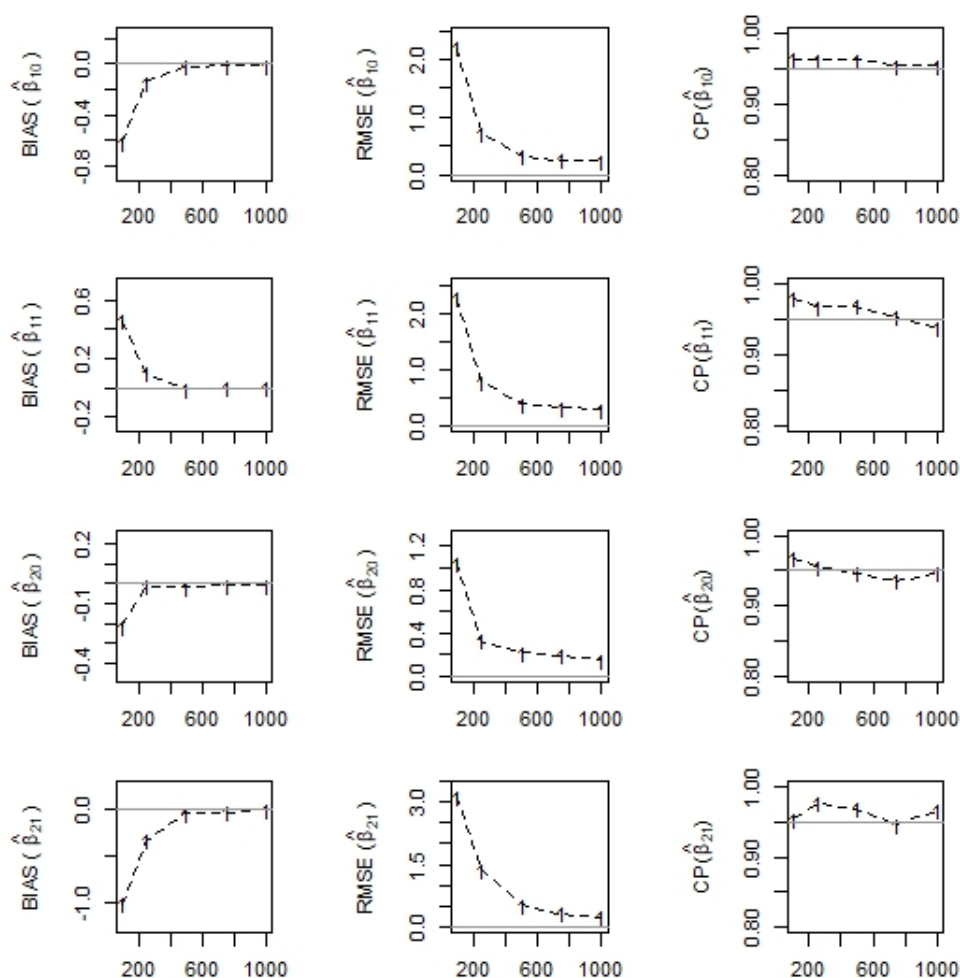
۲. برآوردهای ماکسیمم درست‌نمایی ضرایب رگرسیونی پارامترهای تابع بقای وایبل در هر سه سناریو و به ازای حجم‌های نمونه در نظر گرفته‌شده به مقادیر واقعی پارامترهای انتخاب‌شده بسیار نزدیک هستند. خطای استاندارد برآوردها نیز نشان از دقت خوب برآوردها برای تمام حجم‌های نمونه دارد.

با این حال، در سناریوهای اول و دوم، پارامترهای $\hat{\beta}_{11}$ و $\hat{\beta}_{21}$ برای دستیابی به همگرایی سریع‌تر نیاز به حجم نمونه بزرگتری دارند.

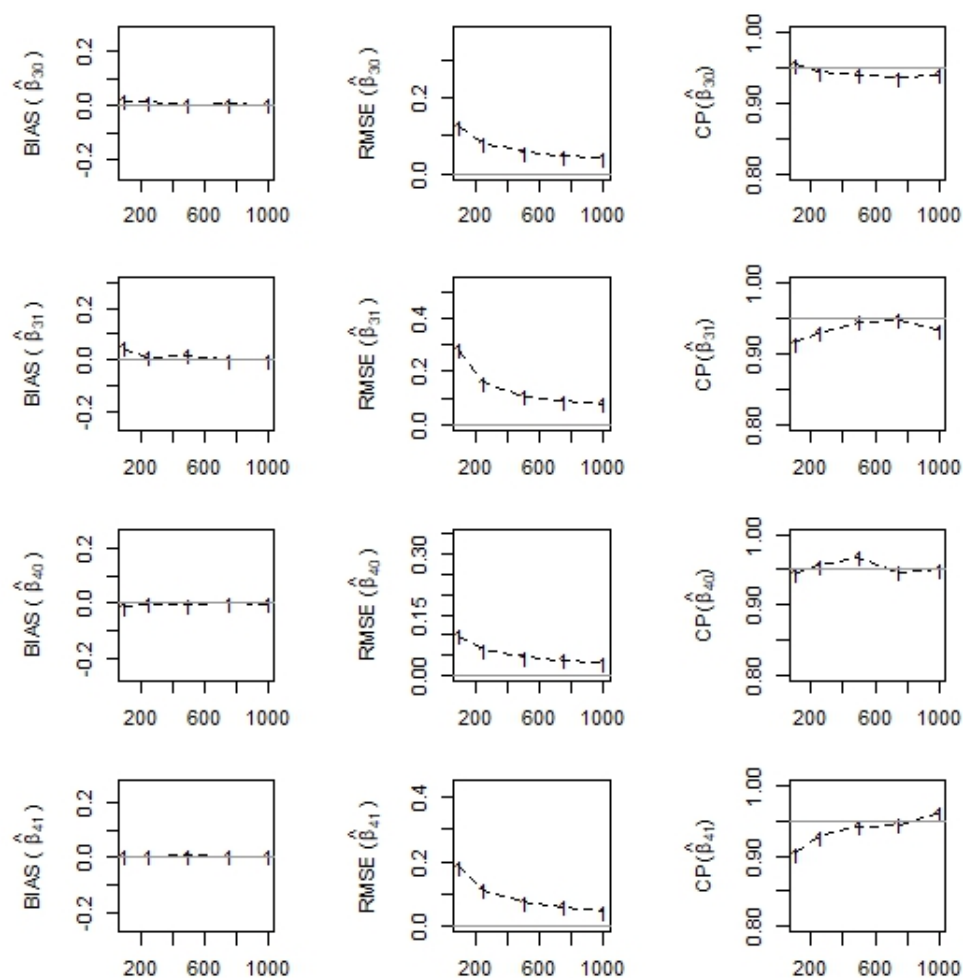
۳. در سناریوهایی که نرخ تقلب و بدون قصور بزرگ‌تر هستند، یعنی سناریوی دوم (با نرخ‌های متوسط) و سناریوی سوم (با نرخ بالا)، وضعیت شاخص‌های اندازه‌گیری شامل نرخ پوشش، اریبی و RMSE نسبت به سناریوی اول با نرخ‌های کوچک، بهتر است.

۴۶ تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

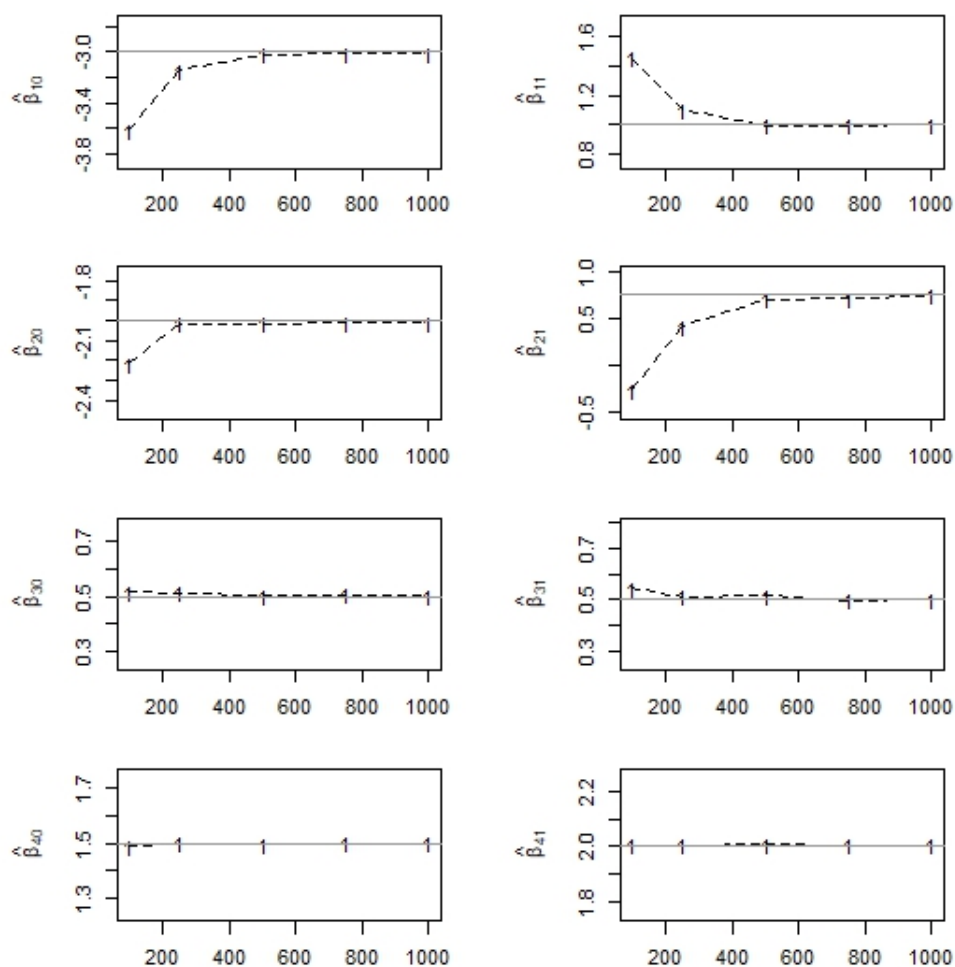
با توجه به نتایج حاصل از مطالعه شبیه‌سازی، می‌توان کارایی روش برآورد ML در مدل مورد نظر این پایان‌نامه را نتیجه گرفت.



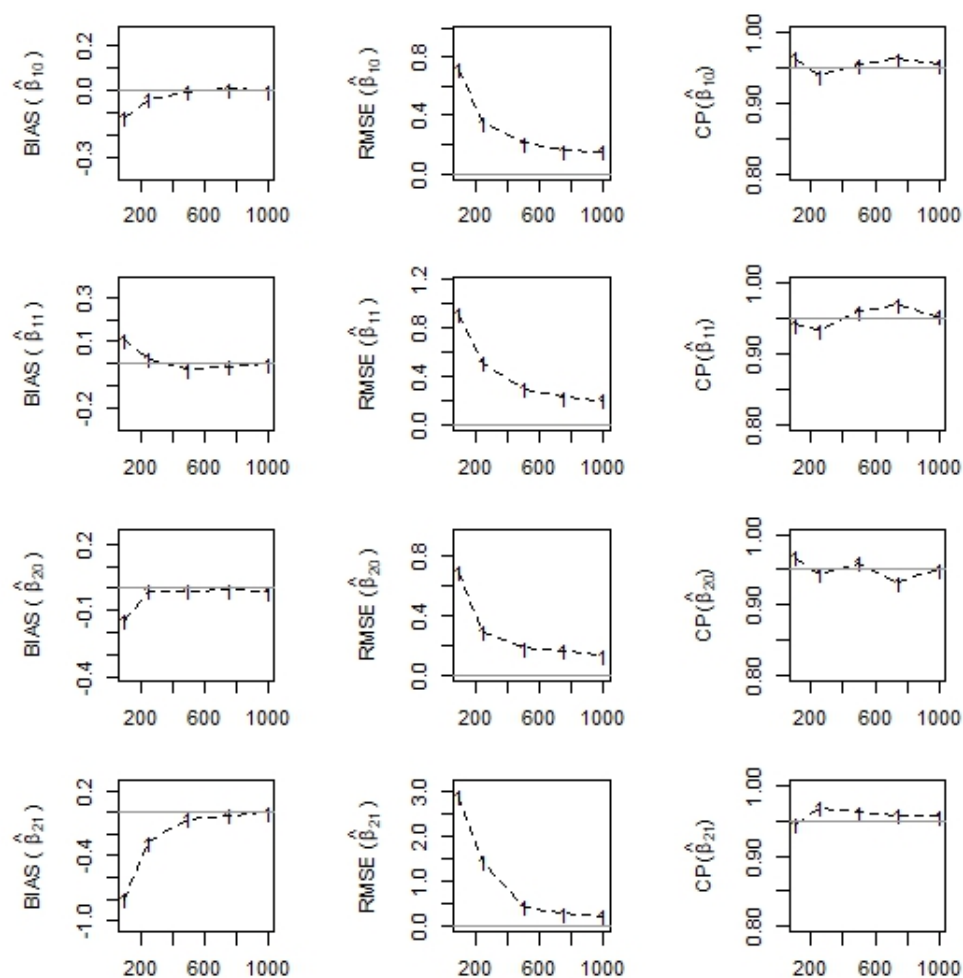
شکل ۱.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی اول انتخاب پارامترها برای ۱۰۰ بار تکرار.



شکل ۲.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی شده در سناریوی اول انتخاب پارامترها

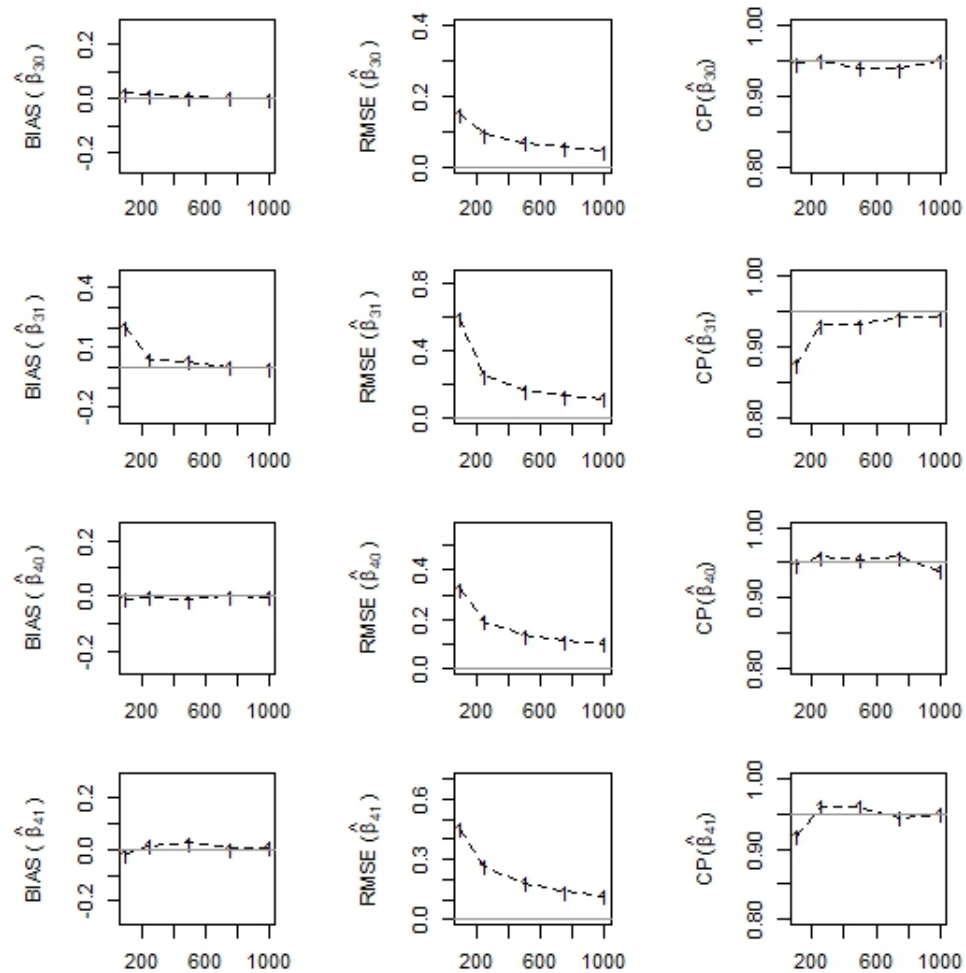


شکل ۳.۴: برآوردهای ماکسیمم درست‌نمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی اول. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد

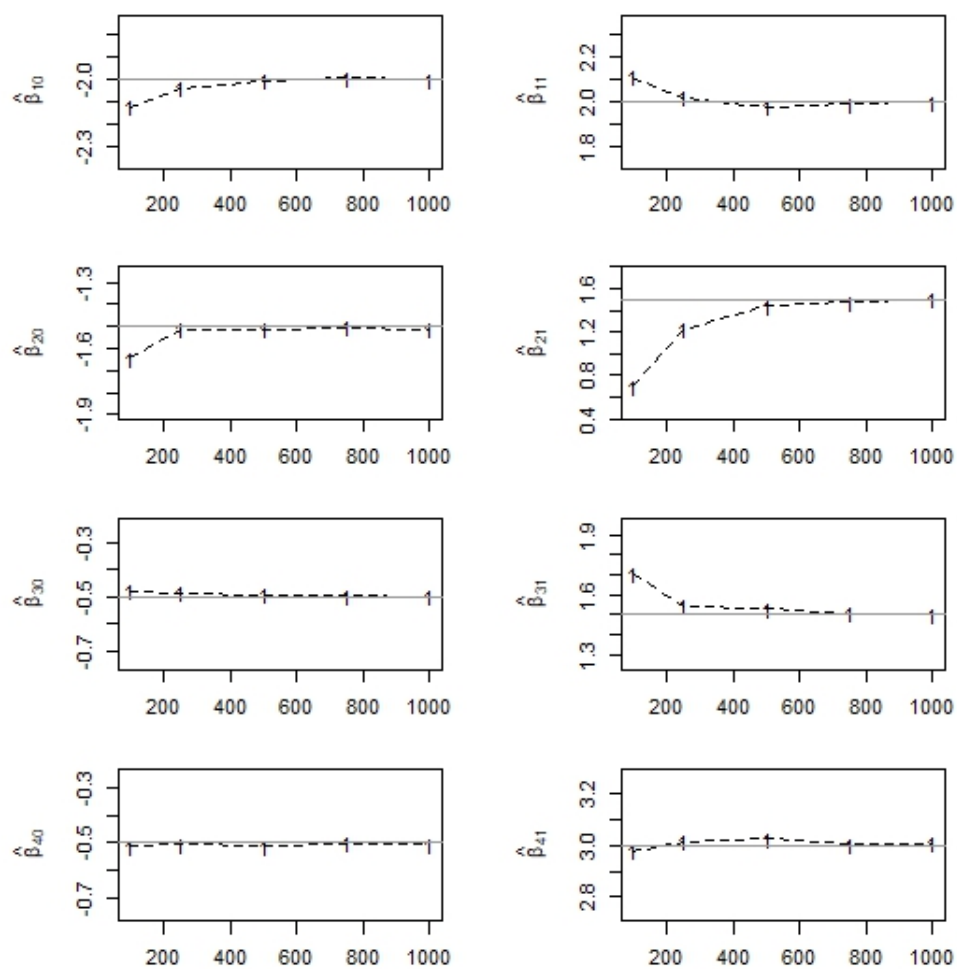


شکل ۴.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درستنمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی شده در سناریوی دوم انتخاب پارامترها برای ۱۰۰ بار تکرار

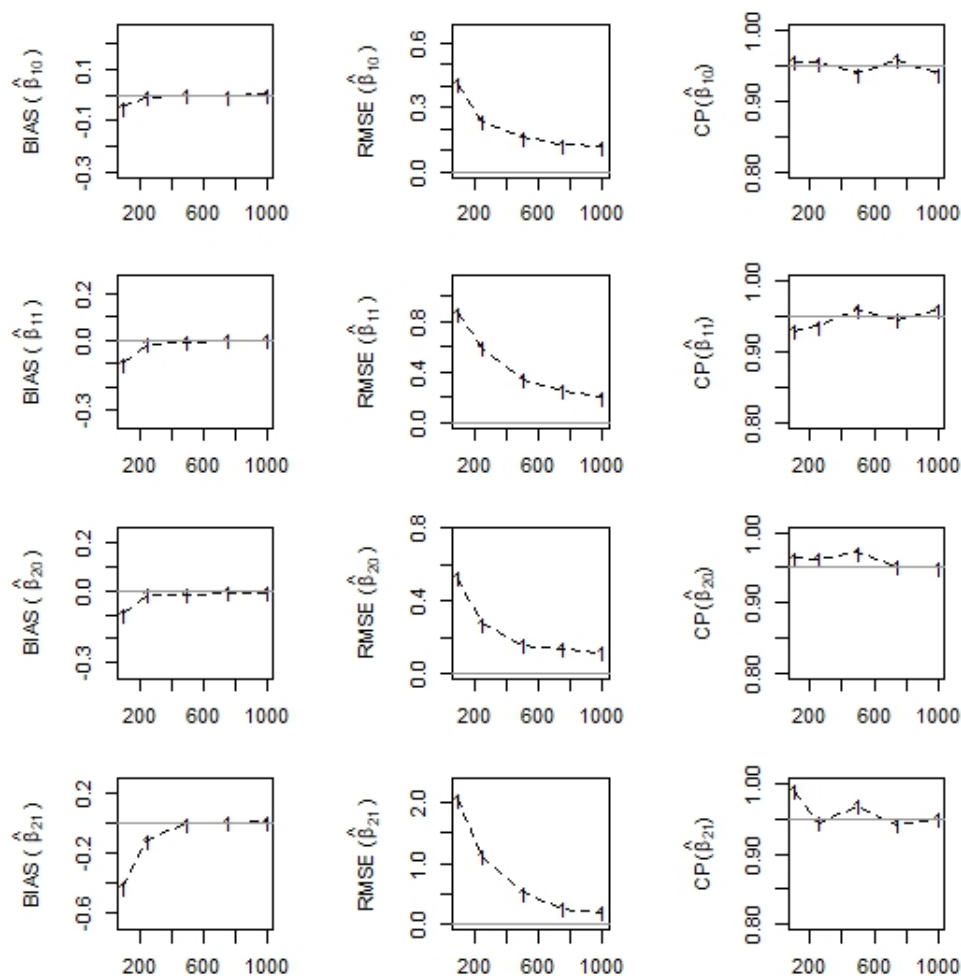
۵. تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی



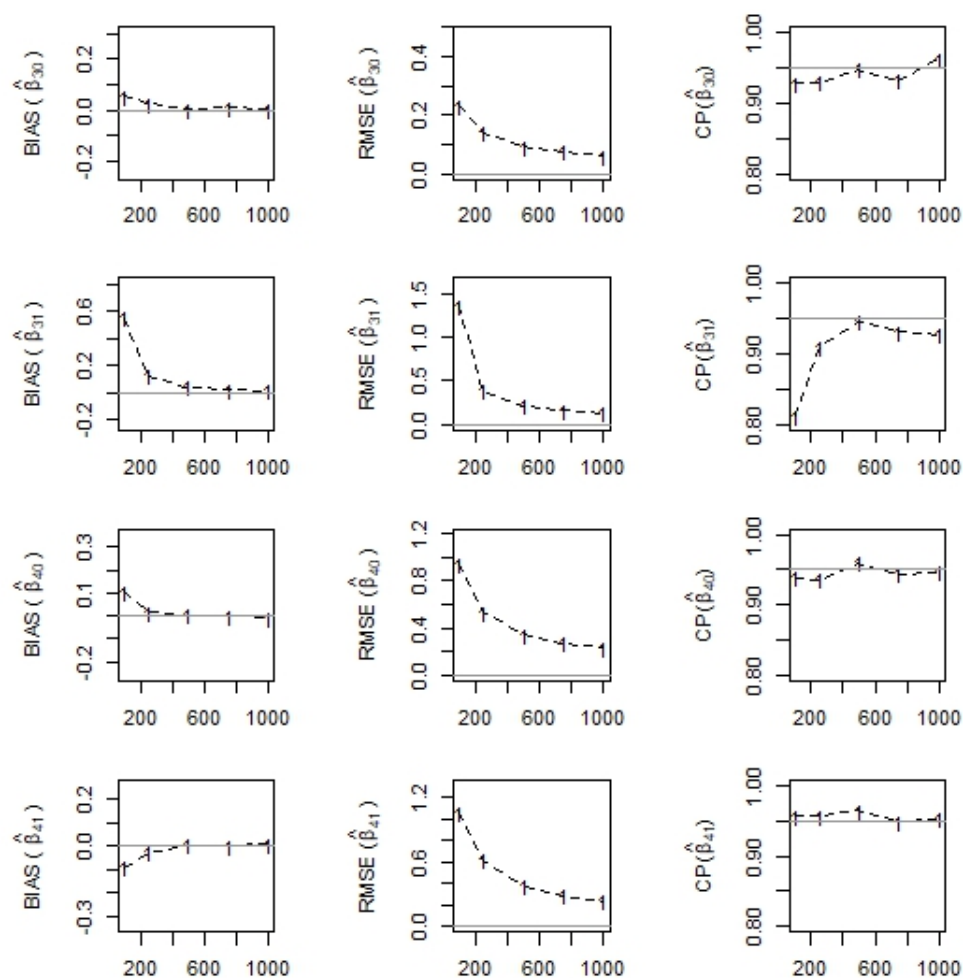
شکل ۵.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی شده در سناریوی دوم انتخاب پارامترها



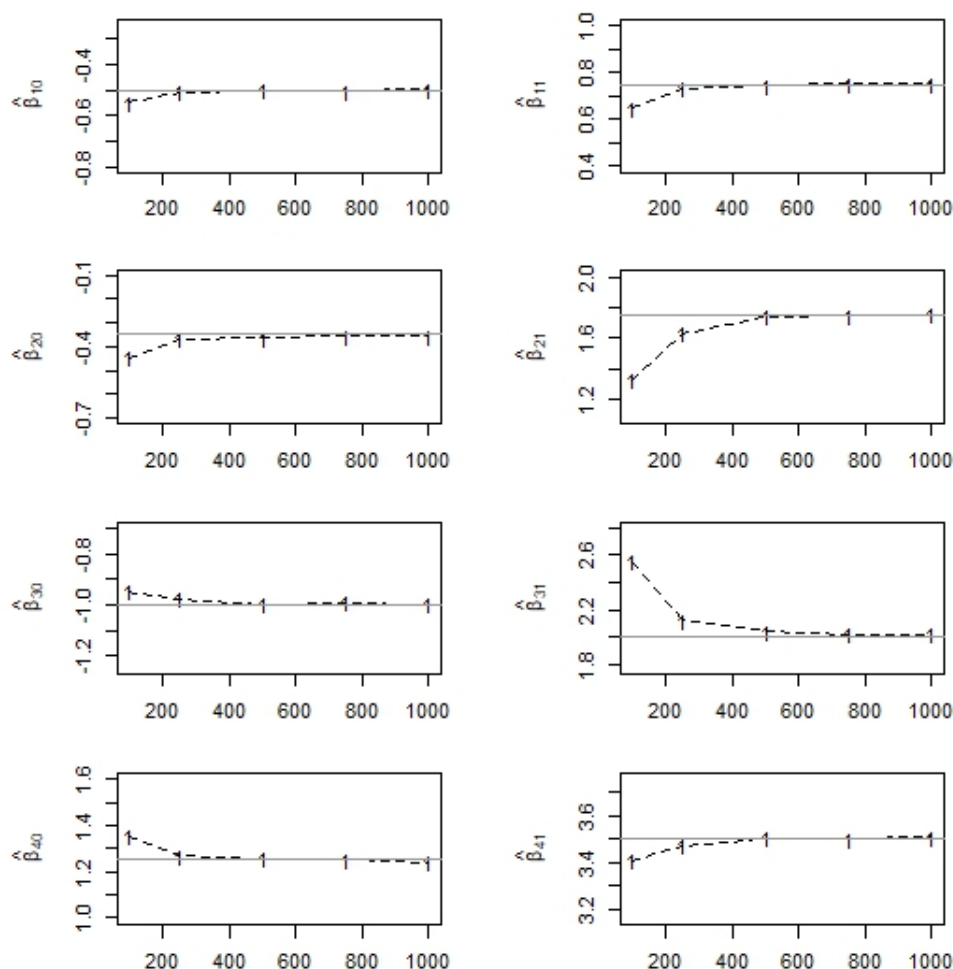
شکل ۶.۴: برآوردهای ماکسیمم درست‌نمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی دوم. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد



شکل ۷.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای نرخ‌های تقلب و قصور مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی سوم انتخاب پارامترها برای ۱۰۰ بار تکرار



شکل ۸.۴: اریبی (BIAS)، ریشه میانگین توان دوم خطا (RMSE) و نرخ پوشش تجربی (CP) برای برآوردهای ماکسیمم درست‌نمایی پارامترهای توزیع وایبل مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های شبیه‌سازی‌شده در سناریوی سوم انتخاب پارامترها



شکل ۹.۴: برآوردهای ماکسیمم درست‌نمایی و خطای استاندارد متناظر آن‌ها برای حجم‌های نمونه مختلف در سناریوی سوم. خط افقی در نمودارها، مقدار واقعی پارامتر را نشان می‌دهد

۲.۴ مطالعه موردی: تسهیلات بانک مرکزی ایران

انواع تسهیلات بانکی که در نظام بانکی ایران تعریف شده‌اند، عبارتند از

۱. فروش اقساطی: بانک‌ها می‌توانند به منظور ایجاد تسهیلات لازم جهت استفاده در امور تولیدی و خدماتی، سه دسته از کالاهای زیر را تهیه کنند و از طریق فروش اقساطی در اختیار متقاضیان قرار دهند:

- مواد اولیه، لوازم یدکی و ابزار کار
- ماشین‌آلات، تأسیسات و تجهیزات تولیدی
- فروش اقساطی مسکن

منظور از فروش اقساطی، واگذاری عین به بهای معلوم، به غیر، به ترتیبی است که تمام یا قسمتی از بهای مزبور به اقساط مساوی یا غیرمساوی در سررسید یا سررسیدهای معین دریافت گردد.

۲. جعاله: یکی دیگر از تسهیلات بانکی که در نظام بانکداری اسلامی به عنوان یک ابزار مورد استفاده قرار می گیرد، جعاله است که به موجب آن جاعل یا کارفرما در مقابل انجام عمل معین (طبق قرارداد) ملزم به ادای مبلغی اجرت معلوم می شود. طرفی که عمل یا کار را انجام می دهد، عامل یا پیمان کار نامیده می شود.

با استفاده از عقد جعاله در به کارگیری تسهیلات بانکی، برای بانکها امکان گسترش و توسعه در امور مربوط به تولیدات صنعتی و کشاورزی، بازرگانی و خدماتی با تنظیم قرارداد به عنوان عامل یا عندالقتضا به عنوان جاعل فراهم می گردد. بنابراین، بانکها قادرند با تنظیم قرارداد جعاله، تقاضای مشتریانی را که نمی توانند نیازشان را از طریق سایر تسهیلات برآورده سازند، مورد اجابت قرار دهند.

۳. قرض الحسنه: در نظام بانکداری ربوی، وجه نقد به صورت وام یا انواع دیگر تسهیلات ربوی در اختیار مشتری قرار می گیرد تا او به هر نحو که مایل است آن را به مصرف برساند. ولی در بانکداری بدون ربا، نیاز مشتریان مستقیمان و در قالب تسهیلات اعطایی توسط بانک تأمین و برطرف می شود.

عقد قرض الحسنه به منظور تأمین نیازهای ضروری مردم مانند ازدواج، و تعمیر مسکن مورد استفاده قرار می گیرد.

۴. مشارکت حقوقی: مشارکت حقوقی عبارت است از تأمین قسمتی از سرمایه شرکت های سهامی جدید یا خرید قسمتی از سهام شرکت های سهامی موجود. طبق تعریف، مشارکت بانکها در سرمایه شرکت های سهامی تولیدی (صنعتی، معدنی، کشاورزی و ساختمانی) و نیز شرکت های سهامی بازرگانی و خدماتی به منظور تأمین کمبود منابع مالی مورد نیاز این گونه شرکت ها، تحت عنوان مشارکت حقوقی صورت می گیرد. به این ترتیب امکان تأمین قسمتی از منابع مالی بلندمدت که برای اجرای طرح مورد نظر شرکت های سهامی ضروری است، از محل منابع سیستم بانکی به وجود می آید.

۵. مشارکت مدنی: از جمله روش هایی که بر اساس موازین شرع، بانک می تواند با اشخاص (حقیقی و حقوقی) وارد معامله شود و نیازهای مالی این گونه اشخاص را برطرف کند، مشارکت مدنی است.

۶. ضمانت نامه: بانکها به اشخاص حقیقی و حقوقی برای ضمانت حسن انجام کار می دهند. در صورتی که شخص از پیمان کار خود رضایت نداشته باشد آن را به اجرا گذاشته و به عنوان خسارت دریافت می کند. در این حالت فردی که درخواست ضمانت کرده است

۵۶ تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

موظف است آن مبلغ را به بانک برگرداند و اگر از کار پیمانکار رضایت داشت ضمانت‌نامه باطل می‌شود.

۷. مضاربه: عقدی است که به موجب آن، بر اساس یک قرارداد بین بانک و شخص (حقیقی و حقوقی) سرمایه و کار لازم برای اقدام به یک امر تجاری (خرید و فروش کالا) فراهم می‌شود. در این قرارداد، بانک (مضارب) تأمین‌کننده وجه مورد لزوم (سرمایه مضاربه) و طرف دیگر قرارداد (عامل) عهده‌دار انجام کلیه امور مربوط به موضوع قرارداد مضاربه است. سود حاصل از انجام معامله مورد نظر، در پایان کار بین بانک و عامل تقسیم خواهد شد. نسبت این تقسیم بر اساس توافق اولیه خواهد بود.

۸. خرید دین: از دیگر ابزارهای کوتاه‌مدت که بانک‌ها به وسیله آن می‌توانند تسهیلات در اختیار مشتریان خود قرار دهند، خرید دین است. قبل از اجرای قانون بانک‌داری بدون ربا، در سیستم بانکی از ابزار خرید دین در تنزیل سفته‌های تجاری استفاده می‌شد. اسناد و اوراق تجاری که از طرف بانک‌ها خریداری می‌شود، باید حقیقی و ناشی از معاملات تجاری بوده و دارنده سفته نیز معتبر باشد تا لطمه‌ای به اصل مبلغ و سود مورد انتشار قابل برگشت وارد نیاید.

بانک‌ها مجاز به خریداری اسناد و اوراق تجاری هستند که مدت سررسید آن‌ها از یک سال تجاوز نکند و در موقع خرید اسناد و اوراق تجاری ناشی از عملیات بازرگانی مختارند آن‌ها را به قیمت کمتر از مبلغ اسمی خریداری کنند.

۹. اجاره به شرط تملیک: یکی از روش‌های بسیار قابل استفاده توسط آن دسته از مشتریان بانک‌ها که در فعالیتهای تولیدی و خدماتی اشتغال دارند یا متقاضی دریافت تسهیلات بانکی بابت مسکن می‌باشند، اجاره به شرط تملیک است که جایگزین موارد عدیدهای از وام‌ها و اعتبارات ربوی معمول در گذشته شده است.

روش اجاره به شرط تملیک مبتنی بر عقد اجاره‌ای است که در آن شرط می‌شود مستأجر در پایان مدت اجاره، در صورت عمل به شرایط مندرج در قرارداد، عین مستأجره (کالا یا اموال موضوع اجاره) را مالک شود.

با استفاده از این روش، بانک‌ها می‌توانند به منظور ایجاد تسهیلات لازم برای گسترش بخش‌های مختلف اقتصاد اعم از امور خدماتی، کشاورزی، صنعتی، معدنی و نیز مسکن، به عنوان موجر، مبادرت به معاملات اجاره به شرط تملیک کنند.

۱۰. سایر: به دلیل محدودیت در تعداد وام‌های باقی‌مانده دسته‌ای از وام‌ها به‌عنوان سایر در نظر گرفته شده‌اند که شامل وام‌های سلف، عقود مبادله‌ای، و عقود مشارکت می‌شوند.

در این مثال واقعی، به کاربست و برازش مدل نرخ شفایافتگی آماسیده در صفر برای داده‌های تسهیلات بانکی ثبت‌شده از چهار بانک در بانک مرکزی ایران، می‌پردازیم.

داده‌های این مثال، یک نمونه به حجم ۱۷۴۷ از مشتریان چهار بانک هستند که از آن بانک‌ها تسهیلات دریافت کرده‌اند. این مجموعه داده، اطلاعات ۲۸ مشتری متقلب، ۴۴۸ مشتری قصورکننده و ۱۲۷۱ مشتری خوش حساب را در خود دارد.

با توجه به محرمانه بودن اطلاعات، چهار بانک مورد نظر را با کدهای A، B، C و D معرفی می‌کنیم که زیر نظر بانک مرکزی ایران در سراسر کشور به مشتریان خود تسهیلاتی را ارائه کرده‌اند. تاریخ تسهیلات اعطاشده در بازه ۱۳۸۳/۱۰/۰۹ تا ۱۳۹۰/۰۷/۰۶ هستند. تاریخ شروع بازپرداخت این مشتریان نیز در دامنه زمانی ۱۳۸۸/۰۴/۳۰ تا ۱۳۹۵/۱۲/۲۹ متغیر است. برای تحلیل این داده‌ها، ابتدا نوع هر مشتری را با توجه به طول زمان بازپرداخت‌هایشان، یعنی فاصله زمانی دریافت وام تا زمانی که بازپرداخت کرده‌اند را تعیین کردیم. مشتری شامل سه دسته متقلب، قصورکننده و خوش حساب هستند. مشتریان متقلب شامل افرادی هستند که حداکثر سه بازپرداخت برای آن‌ها ثبت شده است. مشتریان قصورکننده به آن دسته از مشتریان اطلاق می‌گردد که تا تاریخی معین پرداخت داشته‌اند اما بعد از آن دیگر پرداختی نکرده‌اند. برای این دسته از مشتریان سانسور رخ داده است. مشتریان خوش حساب نیز شامل دو دسته هستند: اول افرادی که پرداختشان به اتمام رسیده است و بدهی آن‌ها صفر است و دوم مشتریانی هستند که تا تاریخ ۱۳۹۵/۱۲/۲۹ تمام اقساطشان را به موقع پرداخت کرده‌اند که برای این دسته مشتریان خوش حساب، سانسور رخ داده است.

سایر اطلاعاتی که در مجموعه داده‌ها موجود هستند، جنسیت مشتریان، نوع بانکی که از آن تسهیلات دریافت کرده‌اند، میزان تسهیلات دریافتی (به ریال)، و نوع وام دریافتی از انواع ۱۰ گانه‌ای که اسم بردیم، می‌باشند.

۱.۲.۴ تحلیل توصیفی داده‌ها

متغیر پاسخ زمان آخرین بازپرداخت مشتریان، t ، است. برای مشتری متقلب، آخرین زمان بازپرداخت $t = 0$ ثبت شده است، زیرا هیچ‌گونه پرداختی نداشتند یا حداکثر ۳ بازپرداخت داشته‌اند. مشتری قصورکننده تا تاریخی که پرداخت نموده، ثبت گردیده است و زمان t به عنوان رخداد واقع مورد علاقه تلقی می‌شود. مشتری خوش حساب نیز با توجه به بازه زمانی داده‌شده یا هم‌چنان در حال پرداخت هستند و تاریخ بازپرداخت آن‌ها هنوز به اتمام نرسیده است که زمان t برای آن‌ها زمان رخداد محسوب می‌شود یا بازپرداخت آن‌ها تمام شده است که می‌توان آن را سانسور شده فرض کرد.

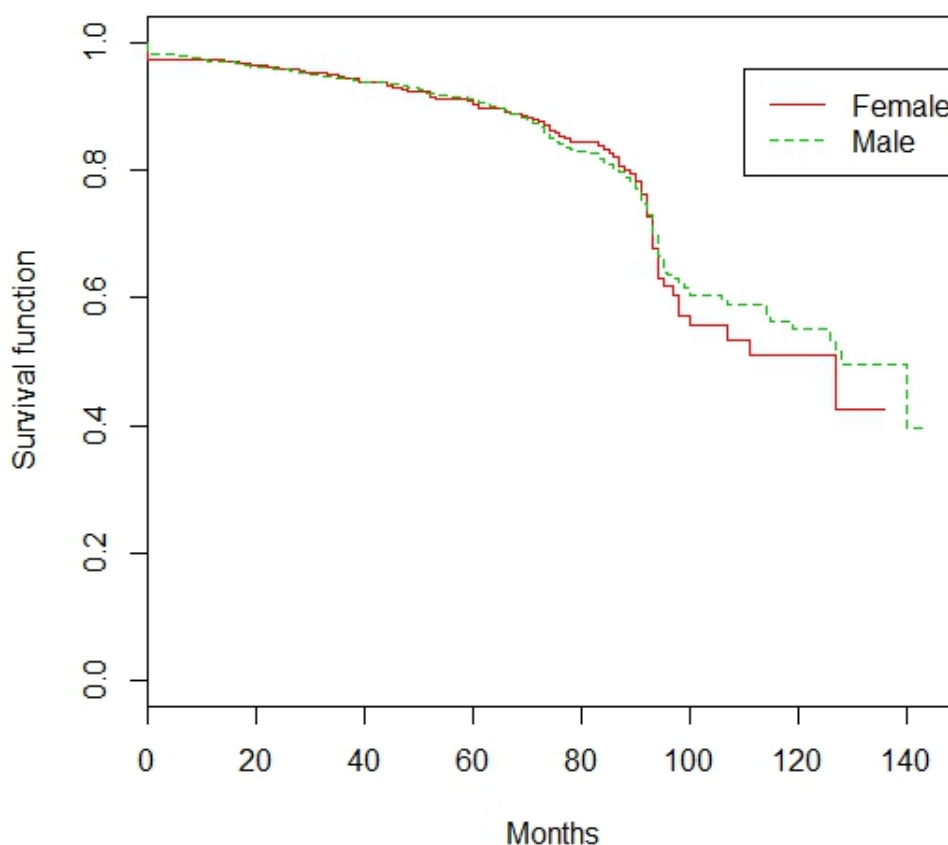
جدول ۱.۴ تفکیک نوع مشتریان را بر حسب جنسیت آن‌ها نمایش می‌دهد. تعداد کل مشتریان موجود در داده‌ها برای گروه مردها، برابر ۱۲۵۵ و برای گروه زن‌ها، برابر ۴۹۲ نفر است. همچنین برای هر گروه تعداد متقلبان، قصورکنندگان و خوش حساب‌ها نیز قابل مشاهده است.

برآورد ناپارامتری تابع بقای طول زمان بازپرداخت‌ها به روش کاپلان-مایر را برای این

۵۸ تحلیل زمان بقای بازپرداخت وام‌های اعطایی: مطالعه شبیه‌سازی و کاربرد واقعی

جدول ۱۰.۴: خلاصه‌ای از داده‌های طول عمر وام بانکی با توجه به جنسیت مشتریان

جنسیت	تعداد مشتریان	تعداد متقلبان	تعداد قصورکنندگان	تعداد خوش حساب‌ها
مرد	۱۲۵۵	۱۸ (۱/۴۳۴۲٪)	۲۹۹ (۲۳/۸۲۴۷٪)	۹۳۸ (۷۴/۷۴۱۰٪)
زن	۴۹۲	۱۰ (۲/۰۳۲۵٪)	۱۴۹ (۳۰/۲۸۴۵٪)	۳۳۳ (۶۸/۶۷۲۹٪)
جمع	۱۷۴۷	۲۸ (۱/۶۰۲۸٪)	۴۴۸ (۲۵/۶۴۳۹٪)	۱۲۷۱ (۷۲/۷۵۳۳٪)



شکل ۱۰.۴: تابع بقای برآوردشده به روش کاپلان-مایر به تفکیک گروه‌های جنسیتی

داده‌ها به تفکیک مردها و زن‌ها محاسبه کرده و در شکل ۱۰.۴ نمایش داده‌ایم. همان‌طور که از منحنی‌های برآوردشده تابع بقا برای دو گروه مرد و زن مشاهده می‌کنید،

• برآوردها برای هر دو گروه جنسیتی مرد و زن تقریباً شبیه هم هستند و اختلاف ناچیزی برای مواردی که طول زمان آن‌ها بیش‌تر از ۱۰۰ ماه است، دیده می‌شود.

• تابع بقا برای هر دو گروه در نقطه زمانی $t = 0$ از ۱ شروع نمی‌شود و نسبت کوچکی از تقلب در هر دو گروه دیده می‌شود. در واقع تابع بقا برای هر دو گروه ناسره است. البته نرخ متقلبان چندان قابل ملاحظه نیست.

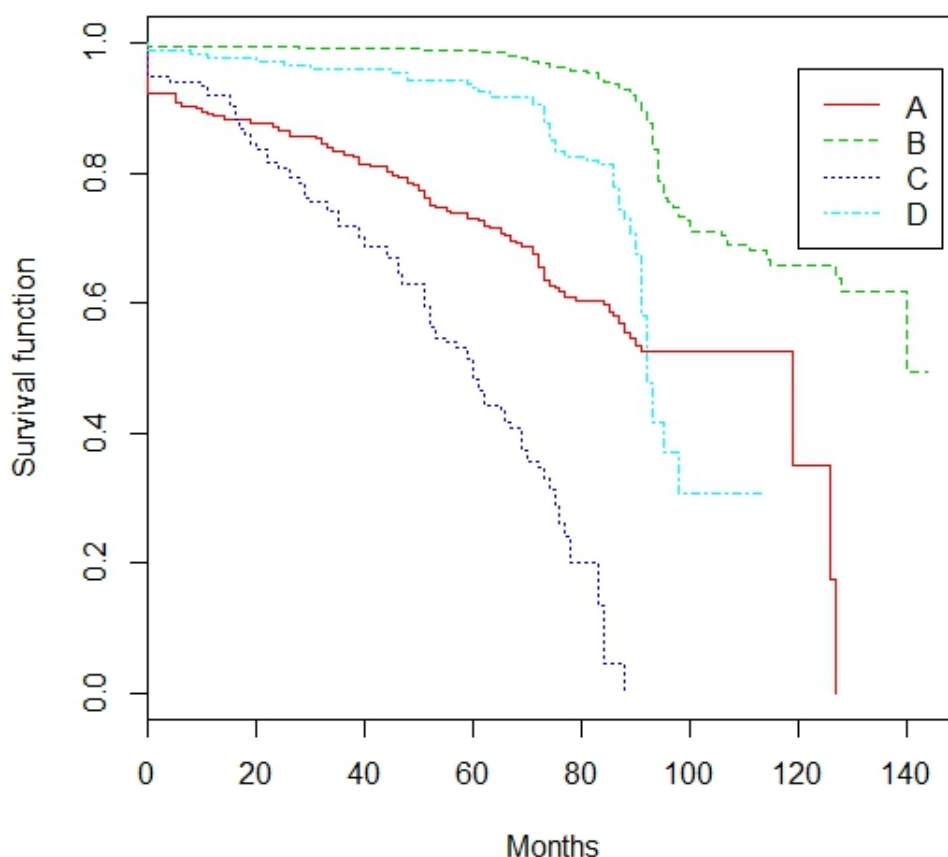
جدول ۲.۴ اطلاعات مشابه جدول ۱.۴ را به تفکیک نوع بانک اعطاکننده وام نشان می‌دهد. بیشترین مشتریان برای بانک B ثبت شده‌اند. در عین حال، بالاترین متقلبین با توجه به تعداد مشتری را بانک A دارد.

جدول ۲.۴: خلاصه‌ای از داده‌های طول عمر وام بانکی بر اساس نوع بانک

نوع بانک	تعداد مشتریان	تعداد متقلبان	تعداد قصورکنندگان	تعداد خوش حساب‌ها
A	۲۵۸	۱۵ (۵/۸۱۳۹٪)	۱۰۰ (۳۸/۷۵۹۷٪)	۱۴۳ (۵۵/۴۲۶۳٪)
B	۱۱۸۱	۶ (۰/۵۰۸۰٪)	۱۶۷ (۱۴/۱۴۰۵٪)	۱۰۰۸ (۸۵/۳۵۱۴٪)
C	۱۳۵	۵ (۳/۷۰۳۷٪)	۱۰۸ (۸٪)	۲۲ (۱۶/۲۹۶۳٪)
D	۱۷۳	۲ (۱/۱۵۶۰٪)	۷۳ (۴۲/۱۹۶۵٪)	۹۸ (۵۶/۶۴۷۴٪)
جمع	۱۷۴۷	۲۸ (۱/۶۰۲۸٪)	۴۴۸ (۲۵/۶۴۳۹٪)	۱۲۷۱ (۷۲/۷۵۳۳٪)

شکل ۱۱.۴ منحنی‌های تابع بقای برآوردشده کاپلان-مایر را به تفکیک نوع بانک برای داده‌های طول بازپرداخت‌ها نشان می‌دهد. واضح است که تابع بقای برآوردشده برای هر چهار نوع بانک ناسره است، به این معنی که در زمان $t = 0$ مقدار تابع بقا کمتر از ۱ است. با این‌که بیشترین مشتری برای بانک B ثبت شده‌اند، اما مشتریان این بانک نرخ تقلب بسیار کوچکی را دارند. بیشترین نرخ تقلب نیز همان‌طور که در جدول ۲.۴ نشان داده شده است، مربوط به بانک A است. بانک C نیز نرخ تقلب قابل ملاحظه‌ای دارد؛ اما نکته‌ای که اشاره به آن ارزشمند است، نزول سریع تابع بقای مدت بازپرداخت‌ها برای بانک C در مقایسه با بانک B است. بانک D نیز تابع بقای کاهشی سریع دارد. به‌طور کلی نرخ شفایافتگی (بدون قصور) برای بانک B از سایر بانک‌ها بیشتر است و بازپرداخت‌های بیشتری نسبت به بقیه بانک‌ها را ثبت کرده‌اند. رتبه دوم مربوط به بانک A هست. بنابراین می‌توان گفت سیاست‌گذاری‌های بانک B در اعطای تسهیلات بانکی، کارایی بیشتری نسبت به سایر بانک‌ها برای دوری از تقلب و قصور داراست.

جدول ۳.۴ نیز اطلاعات توصیفی بر حسب نوع وام‌ها را نشان می‌دهد. در این جدول انواع



شکل ۱۱.۴: تابع بقای برآوردشده به روش کاپلان-مایر به تفکیک نوع بانک

وام‌ها با کدهای ۱، ۲، ...، ۱۰ به ترتیب برای فروش اقساطی، جعاله، قرض الحسنه، مشارکت حقوقی، مشارکت مدنی، ضمانت‌نامه، مضاربه، خرید دین، اجاره به شرط تملیک و سایر، مشخص شده‌اند. به دلیل تعداد دسته‌های زیاد این متغیر تبیینی و اطلاعات بسیار کم برای شش نوع وام از ده نوع، برآوردهای کاپلان-مایر به تفکیک این متغیر را محاسبه و گزارش نکردیم.

شکل ۱۲.۴ هیستوگرام مشاهدات مربوط به مشتریان متقلب و قصورکننده را نشان می‌دهد. با توجه به این نمودار می‌توان ایده‌ای از تغییرات آخرین بازپرداخت‌های قصورکننده‌ها دریافت که نشان می‌دهد حتی بعد از بازپرداخت‌های درازمدت، احتمال قصور صفر نخواهد بود.

برای برازش مدل نرخ شفایافتگی آماسیده در صفر بر روی این داده‌ها، پس از بررسی معنی‌داری متغیرهای تبیینی مختلف شامل جنسیت، نوع بانک، و میزان تسهیلات، تنها نوع بانک زمانی که برای پارامترهای تابع بقای وایبل در نظر گرفته شود، معنی‌دار تشخیص داده

جدول ۳.۴: خلاصه‌ای از داده‌های طول عمر وام بانکی بر اساس نوع وام

نوع وام	تعداد کل	تعداد متقلبان	تعداد قصورکنندگان	تعداد سانسورها
فروش اقساطی	۱۱۹۴	۵(۰/۴۱۸۸٪)	۱۸۹(۱۵/۸۲۹۱٪)	۱۰۰۰(۸۳/۷۵۲۱٪)
جعاله	۵۳	۰(۰٪)	۴۸(۹۰/۵۶۶۰٪)	۵(۹/۴۳۴۰٪)
قرض الحسنه	۸۷	۸(۹/۱۹۵۴٪)	۷۳(۸۳/۹۰۸۰٪)	۶(۶/۸۹۶۵٪)
مشارکت حقوقی	۱۹	۱(۵/۲۶۳۱٪)	۱۴(۷۳/۶۸۴۲٪)	۴(۲۱/۰۵۲۶٪)
مشارکت مدنی	۴۱	۰(۰٪)	۱۷(۴۱/۴۶۳۴٪)	۲۴(۵۸/۵۳۶۶٪)
ضمانت‌نامه	۱	۱(۱۰۰٪)	۰(۰٪)	۰(۰٪)
مضاربه	۶	۰(۰٪)	۶(۱۰۰٪)	۰(۰٪)
خرید دین	۱	۰(۰٪)	۰(۰٪)	۱(۱۰۰٪)
اجاره به شرط تملیک	۲	۰(۰٪)	۰(۰٪)	۲(۱۰۰٪)
سایر	۳۴۳	۱۳(۳/۷۹۰۱٪)	۱۰۱(۲۹/۴۴۶۱٪)	۲۲۹(۶۶/۷۶۳۸٪)
جمع	۱۷۴۷	۲۸(۱/۶۰۲۸٪)	۴۴۸(۲۵/۶۴۳۹٪)	۱۲۷۱(۷۲/۷۵۳۳٪)

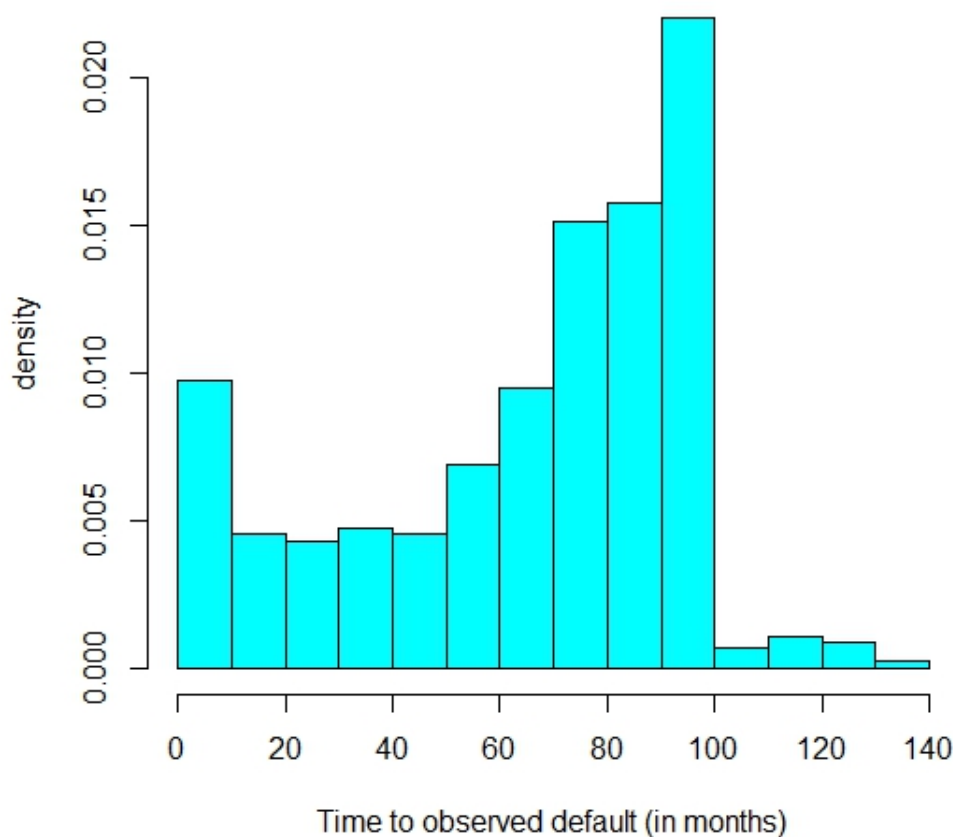
شد. لازم به ذکر است با توجه به رسته‌ای بودن متغیر نوع بانک، برای وارد کردن آن در مدل از متغیرهای ظاهری^۵ دودویی برای معرفی چهار نوع بانک استفاده کردیم. بنابراین چون چهار نوع بانک داشتیم، سه متغیر ظاهری z_1 ، z_2 و z_3 را تعریف کردیم، به‌طوری که در جدول ۴.۴ مشاهده می‌کنید.

جدول ۴.۴: متغیرهای ظاهری برای انواع بانکها

	z_1	z_2	z_3
A	۰	۰	۰
B	۱	۰	۰
C	۰	۱	۰
D	۰	۰	۱

با این توضیح، مدل نرخ شفایافتگی آماسیده در صفر را با پارامترهای مختلف به صورت

^۵Dummy Variables



شکل ۱۲.۴: هیستوگرام طول زمان بازپرداخت فقط مشتریان متقلب و قصورکننده

زیر در نظر گرفتیم:

$$\begin{aligned}\gamma_{\kappa^0 i} &= \frac{e^{\beta_{\kappa^0}}}{1 + e^{\beta_{\kappa^0}} + e^{\beta_{\theta^0}}} \\ \gamma_{\theta^0 i} &= \frac{e^{\beta_{\theta^0}}}{1 + e^{\beta_{\kappa^0}} + e^{\beta_{\theta^0}}} \\ \alpha_i &= e^{\beta_{\alpha^0} + z_{1i}\beta_{\alpha 1} + z_{2i}\beta_{\alpha 2} + z_{3i}\beta_{\alpha 3}} \\ \lambda_i &= e^{\beta_{\lambda^0} + z_{1i}\beta_{\lambda 1} + z_{2i}\beta_{\lambda 2} + z_{3i}\beta_{\lambda 3}}\end{aligned}$$

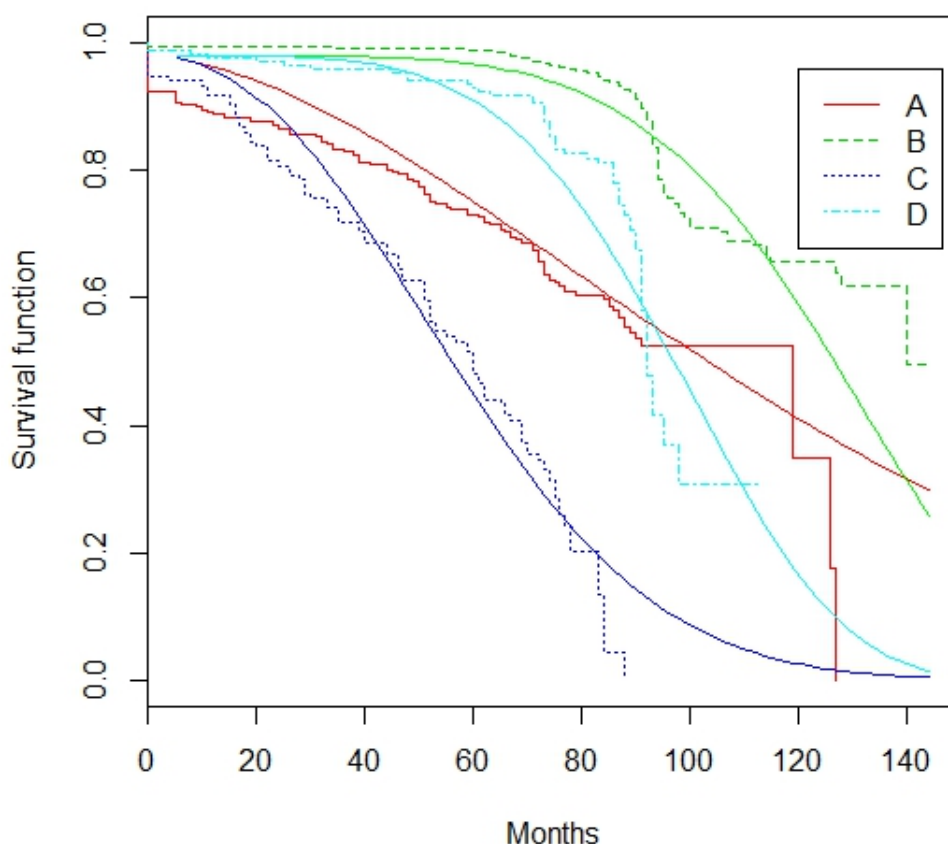
نتایج برازش مدل بر روی داده‌ها در جدول ۵.۴ گزارش شده‌اند. با توجه به نتایج به‌دست آمده بر حسب مقدار آماره $t = \frac{\hat{\beta}}{se(\hat{\beta})}$ ، می‌توان مشاهده کرد تنها ضریب رگرسیونی پارامتر بانک B، یعنی $\beta_{\lambda 1}$ ، معنی‌دار نیست.

با استفاده از مدل برازش‌شده نرخ تقلب برای این داده‌ها، به‌طور کلی، مقدار ۲۱٪ برآورد می‌شود. برآورد توابع بقای طول زمان بازپرداخت‌ها به تفکیک هر بانک از روی مدل در

جدول ۵.۴: برآورد پارامترهای مدل برای داده‌های تسهیلات بانک مرکزی

مقدار t	خطای معیار	برآورد پارامتر	پارامتر
-۲۲/۹۳	۰/۱۷	-۳/۸۶	β_{κ^0}
-۰/۰۰۸	۲۹۶۵	-۲۵/۰۰	β_{θ^0}
۵/۵۶	۰/۰۹۷	۰/۵۴	β_{α^0}
۹/۹۷	۰/۱۱	۱/۱۲	β_{α^1}
۱/۹۷	۰/۱۳۱	۰/۲۶	β_{α^2}
۶/۷۶	۰/۱۴۶	۰/۹۸	β_{α^3}
۶۴/۶۶	۰/۰۷۵	۴/۸۷	β_{λ^0}
۰/۶۰	۰/۰۷۹	۰/۰۴۷	β_{λ^1}
-۷/۵۲	۰/۰۸۸	-۰/۶۶	β_{λ^2}
-۲/۵۴	۰/۰۸۲	-۰/۲۰۷	β_{λ^3}

شکل ۱۳.۴ نمایش داده شده‌اند. برآوردهای تابع بقا برای هر بانک در مدل پارامتری در نظر گرفته‌شده با برآوردهای ناپارامتری حاصل از روش کاپلان-مایر توافق زیادی دارند. در عین حال، برای زمان‌های طولانی، برآورد ناپارامتری کاپلان-مایر به دلیل کم بودن افراد در معرض خطر چندان دقیق نیست. بنابراین برآوردهای حاصل از مدل می‌تواند به‌عنوان جانشین‌های کاراتری مورد استفاده قرار گیرند. البته برآوردهای نرخ تقلب توسط مدل با برآوردهای ناپارامتری اختلاف دارند که دلیل آن هم عدم معنی‌داری متغیرهای تبیینی در مدل‌بندی نرخ‌های تقلب و قصور است. وجود اطلاعات کمکی بیشتر می‌تواند به برآوردهای دقیق‌تر کمک کند.



شکل ۱۳.۴: برآورد توابع بقای وایبل برای مدل نرخ شفایافتگی به تفکیک نوع بانک

۳.۴ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه، تعمیمی از یک مدل نرخ شفایافتگی بقا معرفی شد که برای مدل‌بندی نسبت متقلبین در مسایل اعطای تسهیلات بانکی می‌تواند مفید باشد و مورد استفاده قرار گیرد. برای استنباط در این مدل، دیدگاه مبتنی بر درست‌نمایی را معرفی کردیم و برآوردهای ماکسیمم درست‌نمایی مدل را با روش عددی BFGS به‌دست آوردیم. همچنین نشان دادیم برآوردهای حاصل از ماکسیمم‌سازی عددی تابع درست‌نمایی ویژگی‌های بهینه‌نظری برآوردهای ML، شامل سازگاری، کارایی و نرمال‌جانبی، را دارا هستند.

با مثال‌های شبیه‌سازی و یک مجموعه داده واقعی مربوط به داده‌های تسهیلات چهار نوع بانکی داخلی، عملکرد مدل پیشنهادی را مورد ارزیابی قرار دادیم. در مجموع نتایج حاصل از مثال‌ها بر کارایی مناسب مدل پیشنهادی تاکید دارند.

با توجه به مطالب ارایه‌شده در این پایان‌نامه، می‌توان موارد زیر را به‌عنوان آینده تحقیق مطرح کرد:

- استنباط مبتنی بر دیدگاه بیزی در مدل پیشنهادی می‌تواند مورد توجه باشد. با توجه به این‌که رفتار وام‌گیرندگان در ایران در گذشته برای بانک‌های مختلف تا حدودی مشخص است، می‌توان از آن اطلاعات برای تعریف توزیع‌های پیشین مناسب برای پارامترهای مدل استفاده کرد. بنابراین انتظار می‌رود اگر چنین توزیع‌های پیشینی را بتوان تعریف کرد، استنباط‌های کاراتری به‌دست آورد.
- در مدل پیشنهادی پایان‌نامه، اثر متغیرهای تبیینی خطی در نظر گرفته شد. مدل‌بندی ناپارامتری اثر این متغیرها در بخش‌های مختلف مدل شامل نرخ تقلب و قصور و همچنین پارامترهای توزیع بقا، با استفاده از روش‌هایی مثل اسپلاین‌ها، می‌تواند جالب و موضوعی جذاب برای تحقیق باشد.
- برای مدل‌بندی توزیع بقا از مدل پارامتری وایبل استفاده کردیم. به‌کارگیری روش‌های ناپارامتری بقا مثل مدل مخاطره متناسب کاکس^۶ نیز قابل بررسی است. همچنین، از سایر توزیع‌های پارامتری مثل لگ‌نرمال می‌توان بهره گرفت.

^۶ Cox Proportional Hazard Model

مراجع

- [۱] آشفته ا، بزرگ‌نیا ا، (۱۳۸۰)، ”تحلیل بقا”، مجله اندیشه آماری، شماره ۲، سال ۷، ۳-۱۰
- [۲] بزرگ‌نیا ا. رضایی پژند ح. (۱۳۸۰)، ”تحلیل بقا” چاپ اول، انتشارات دانشگاه فردوسی مشهد
- [۳] پاریسیان ا. (۱۳۹۱)، ”مبانی آمار ریاضی”، چاپ دهم، انتشارات دانشگاه صنعتی اصفهان
- [۴] حقیقی ف. و جان‌بابایی ق. و صادقی س. (۱۳۹۰)، ”بکارگیری تحلیل بقا در برآورد طول عمر بیماران مبتلا به سرطان معده”، مجله دانشگاه علوم پزشکی مازندران، شماره ۸۵، دوره ۲۱، صفحه ۲۸-۳۶
- [۵] عبدی م، (۱۳۸۸)، پایان نامه ارشد: ”استنباط برای توزیع وایبل براساس نمونه‌های کامل و سانسور شده”، دانشکده علوم پایه، دانشگاه مازندران
- [۶] نیرومند ح.ع. (۱۳۸۴)، ”الگوهای خطی تعمیم‌یافته: با کاربردهای آن در علوم و مهندسی”، انتشارات دانشگاه فردوسی مشهد
- [7] Avriel M. (2003), ” *Nonlinear programming: analysis and methods* ” Courier Corporation
- [8] Barriga G.D.C. and Cancho V.G. and Louzada F. (2015), ” *A non-default rate regression model for credit scoring* ” Applied Stochastic Models in Business and Industry **31**, pp 846–861
- [9] Berkson J. and Gage R.P. (1952), ” *Survival curve for cancer patients following treatment* ” Journal of the American Statistical Association **47**, pp 501–515
- [10] Boag J.W. (1949), ” *Maximum likelihood estimates of the proportion of patients cured by cancer therapy* ”, Journal of the Royal Statistical Society, **11(1)**, pp 15-53
- [11] Chen M.H. and Ibrahim J.G. and Sinha D. (1999), ” *A new Bayesian model for survival data with a surviving fraction* ” Journal of the American Statistical Association **94**, pp 909–919

-
- [12] Fisher A. R. (1925), " *Theory of statistical estimation* ", Mathematical Proceedings of the Cambridge Philosophical Society, **22(5)**, pp 700-725
- [13] Frechet M. (1928), " *Sur la loi de probabilit  de lecart maximum* " Annales de la societ  polonaise de mathematique , pp. 93-116
- [14] Gauss C.F. (1821), " *Theory of the combination of observations which leads to the smallest errors* ", Gauss Werke **4**, pp 1-93
- [15] Graunt J.(1977). " *Natural and political observations mentioned in a following index, and made upon bills of mortality* ", In: Mathematical Demography. Biomathematics, vol 6. Springer, Berlin, Heidelberg
- [16] Hansen E. (1978), " *Interval forms of Newtons method* ", Computing, **20(2)**, pp 153-163
- [17] Johansson P. and Ekstrand-Tobin A. and Svensson T. and Bok G. (2012). " *Laboratory study to determine the critical moisture level for mould growth on building materials* " , International Biodeterioration and Biodegradation Vol 73, pp 23-32
- [18] Kaplan E.L. and Meier P. (1958), " *Nonparametric Estimation from Incomplete Observations* " . Journal of the American Statistical Association Vol 53,**282**, pp 457-481
- [19] Kleinbaum D.G. and Klein M. (2010), " *Survival analysis* " Springer, Vol 3
- [20] Louzada F. and Oliveira Jr M.R.d. and Moreira F.F. (2015), " *The zero-inflated cure rate regression model: Applications to fraud detection in bank loan portfolios* ", arXiv:1509.05244
- [21] Louzada F. and Oliveira Jr M.R.d. and Moreira F.F. (2016), " *Zero-inflated cure rate regression models for time-to-default with applications* ", arXiv:1509.05244
- [22] Louzada F. and Oliveira Jr M.R.d. and Moreira F.F. (2017), " *A zero-inflated non-default rate regression model for credit scoring data* ", Communications in Statistics-Theory and Methods, DOI: 10.1080/03610926.2017.1346803
- [23] McCarty J.P. and Levey D.J. and Greenberg C.H. and Sargent S. (2002), " *Spatial and temporal variation in fruit use by wildlife in a forested landscape* ", Forest Ecology and Management Vol 164, **1**, pp. 277-291
- [24] Migon H.S. and Gamerman D. and Louzada F. (2014), " *Statistical inference: an integrated approach* " . CRCpress.

- [25] Najafi F. and Ahmadijouybari T. and Moradinazar M. and Ataie M. and Hatami M. and Almasi A. and others (2013), " *The survival rate of self-immolators in Kermanshah Province 2010-2011* ", Journal of Kermanshah University of Medical Sciences Vol. 17, **9**, pp 563-571
- [26] Oliveira M.R. and Louzada F. (2014), " *An evidence of link between default and loss of bank loans from the modeling of competing risks* ", Singaporean Journal of Business Economics and Management Studies **3**, pp 30-37
- [27] Oliveira M.R. and Louzada F. (2014), " *Recovery risk: Application of the latent competing risks model to non performing loans* ", Tecnologia de Credito **88**, pp 43-53
- [28] Othus M. and Barlogie B. and LeBlanc M.L. and Crowley J.J. (2012), " *Cure models as a useful statistical tool for analyzing survival* ", Clinical Cancer Research **18**, **14**, pp. 3731-3736.
- [29] Raphson J. (1702), " *Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, and expedita, ex nova infinitarum serierum methodo, deducta ac demonstrata* "
- [30] Rocha R. and Nadarajah S. and Tomazella V. and Louzada F. and Eudes A. (2017). " *New defective models based on the kumaraswamy family of distributions with application to cancer data sets* ". Statistical Methods in Medical Research , **26(4)**, pp 1737-1755
- [31] Rodrigues J. and Cancho V.G. and Castro M. and Louzada-Neto F. (2009), " *On the uni cation of long-term survival models* ", Statistics and Probability Letters **79**, **6**, pp 753-759
- [32] Rosin P. (1933), " *The laws governing the fineness of powdered coal* " J. Inst. Fuel **7**, pp 29-36
- [33] Team R.C. (2015), " *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014* "
- [34] Tong E.N. and Mues C. and Thomas L.C. (2012), " *Mixture cure models in credit scoring: If and when borrowers default* ". European Journal of Operational Research **218**, **1**, pp. 132-139.
- [35] Vahidpour M. (2016), " *Cure Rate Models* " Ecole Polytechnique, Montreal (Canada)
- [36] Vittinghoff E. and Glidden D.V. and Shiboski S.C. and McCulloch C.E. (2011), " *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models* " Springer Science and Business Media

-
- [37] Weibull W. (1961), "*Fatigue Testing and Analysis of Results: Publ. for and on Behalf of Group for Aeronautical Research and Development, North Atlantic Treaty Organisation*" Pergamon Press
- [38] Yakovlev A.Y. and Tsodikov A.D. (1996), "*Stochastic Models of Tumor Latency and their Biostatistical Applications* ". New Jersey: World Scientific

پیوست آ

دستورات نرم افزار R

در این پیوست، دستورات R مربوط برای بازتولید نتایج مثال شبیه سازی در فصل ۴ ارائه شده اند.

```
rm(list=ls(all=TRUE))  
library(parallel)    #Library for PARALLEL Processing  
  
simFunc<-function(N){  
  B<-500              #Number of Estimates.  
  set.seed(2016)  
  
  #Set the covariates.  
  #Set the first parameter scenario.  
  beta1=0.5  
  beta2=0.5  
  beta3=1.5  
  beta4=2  
  beta5 =  -3
```

```

beta6 = 1
beta7 = -2
beta8 = 0.75
varp=c(beta1,beta2,beta3,beta4,beta5,beta6,beta7,beta8)
lime = 10

# Declaring vectors and counters.
emle <-matrix(nrow=B,ncol=8)
old = options(digits=10)
o<-1
ite<-1
pc1<-rep(0,times=B)
pc2<-rep(0,times=B)
pc3<-rep(0,times=B)
pc4<-rep(0,times=B)
pc5<-rep(0,times=B)
pc6<-rep(0,times=B)
pc7<-rep(0,times=B)
pc8<-rep(0,times=B)
censura = matrix(numeric(B),B,1)

model <- function(N,varp){
a      = exp(varp[1])*exp(y*varp[2])
lam     = exp(varp[3])*exp(y*varp[4])

# Parameters - WEIGHT IN ZEROS AND INFS: GAMMA ZERO AND GAMMA ONES WITH A COVARIABLE
gamm_zer = (exp(varp[5])*exp(y*varp[6]))/(1+exp(varp[5])*exp(y*varp[6]))+
exp(varp[7])*exp(y*varp[8]))
gamm_inf = (exp(varp[7])*exp(y*varp[8]))/(1+exp(varp[5])*exp(y*varp[6]))+
exp(varp[7])*exp(y*varp[8]))

U=runif(N)
V=runif(N,gamm_zer,1-gamm_inf)
y=numeric(N)

```

```

for(i in 1:N){
  if(U[i]<=gamm_zer[i]){
    y[i] = 0
  }else
    if(U[i]>gamm_zer[i] & U[i] <= 1-gamm_inf[i]){
      #y[i]=qweibull(V[i], a, lam)
      y[i] = lam[i]*((-log(1 - (V[i]-gamm_zer[i])/(1-gamm_zer[i]-gamm_inf[i]))))^(1/a[i]))
    }else {
      y[i]=+Inf
    }
  }
  return(y)
}

```

```

GPE=function(varp)
{
  a      = exp(varp[1])*exp(y*varp[2])
  lam    = exp(varp[3])*exp(y*varp[4])
  gamm_zer = (exp(varp[5])*exp(y*varp[6]))/(1+exp(varp[5])*exp(y*varp[6])+
  exp(varp[7])*exp(y*varp[8]))
  gamm_inf = (exp(varp[7])*exp(y*varp[8]))/(1+exp(varp[5])*exp(y*varp[6])+
  exp(varp[7])*exp(y*varp[8]))
  f0      = gamm_zer
  fw      = (exp(log(a)-log(lam)))*((exp(log(time)-log(lam)))^(a-1))*
  (exp(-(exp(log(time)-log(lam)))^a))
  Sw      = (exp(-(exp(log(time)-log(lam)))^a))
  fpop     = (1-gamm_zer-gamm_inf)*fw
  Spop     = gamm_inf + (1-gamm_zer-gamm_inf)*Sw
  f        = (fpop^delta)*(Spop^(1-delta))
  g        <- ifelse(time==0, f0, f)
  adFunc   = sum(log(g))
  return(adFunc)
}

```

```
while(o<=B){
y=rbinom(N,1,0.5)
tp = model(N,varp)
tM = max(tp[tp<+Inf])
Z=runif(N,0,tM)
time=numeric(N)
for (i in 1:N)
{time[i] = min(tp[i],Z[i])
}
delta=numeric(N)
for (i in 1:N) {
if (time[i] < Z[i]){
delta[i]=1
}
}

t<-time
d<-delta
x<-y

fit=try(optim(varp,GPE,method="BFGS",hessian=TRUE,control=list(fnscale=-1)))
estc = try(fit$par)
Hc=try(fit$hessian); varic=try(-solve(Hc))

if ( is.finite(estc[1]) & is.finite(estc[2]) & is.finite(estc[3]) &
is.finite(estc[4]) & is.finite(estc[5]) & is.finite(estc[6]) &
is.finite(estc[7]) & is.finite(estc[8])) {
if ( is.finite(varic[1,1]) & is.finite(varic[2,2]) &
is.finite(varic[3,3]) & is.finite(varic[4,4]) &
is.finite(varic[5,5]) & is.finite(varic[6,6]) & is.finite(varic[7,7]) &
is.finite(varic[8,8])){
if ( varic[1,1] > 0 & varic[2,2] > 0 & varic[3,3] > 0 & varic[4,4] > 0 &
varic[5,5] > 0 & varic[6,6] > 0 & varic[7,7] > 0 & varic[8,8] > 0 &
```

```
estc[1] < lime & estc[2] < lime & estc[3] < lime & estc[4] < lime &
estc[5] < lime & estc[6] < lime & estc[7] < lime & estc[8] < lime ){

var.varp1 =varic[1,1]; Lvarp1 =estc[1]-1.96*sqrt(var.varp1);
  Uvarp1=estc[1]+1.96*sqrt(var.varp1)
var.varp2 =varic[2,2]; Lvarp2 =estc[2]-1.96*sqrt(var.varp2);
  Uvarp2=estc[2]+1.96*sqrt(var.varp2)
var.varp3 =varic[3,3]; Lvarp3 =estc[3]-1.96*sqrt(var.varp3);
  Uvarp3=estc[3]+1.96*sqrt(var.varp3)
var.varp4 =varic[4,4]; Lvarp4 =estc[4]-1.96*sqrt(var.varp4);
  Uvarp4=estc[4]+1.96*sqrt(var.varp4)
var.varp5 =varic[5,5]; Lvarp5 =estc[5]-1.96*sqrt(var.varp5);
  Uvarp5=estc[5]+1.96*sqrt(var.varp5)
var.varp6 =varic[6,6]; Lvarp6 =estc[6]-1.96*sqrt(var.varp6);
  Uvarp6=estc[6]+1.96*sqrt(var.varp6)
var.varp7 =varic[7,7]; Lvarp7 =estc[7]-1.96*sqrt(var.varp7);
  Uvarp7=estc[7]+1.96*sqrt(var.varp7)
var.varp8 =varic[8,8]; Lvarp8 =estc[8]-1.96*sqrt(var.varp8);
  Uvarp8=estc[8]+1.96*sqrt(var.varp8)

emle[o,1] <- estc[1]; emle[o,2] <- estc[2]; emle[o,3] <- estc[3]; emle[o,4] <- estc[4];
emle[o,5] <- estc[5]; emle[o,6] <- estc[6]; emle[o,7] <- estc[7]; emle[o,8] <- estc[8];
{
if (beta1 >= Lvarp1 && beta1 <= Uvarp1) pc1[o]<-1
if (beta2 >= Lvarp2 && beta2 <= Uvarp2) pc2[o]<-1
if (beta3 >= Lvarp3 && beta3 <= Uvarp3) pc3[o]<-1
if (beta4 >= Lvarp4 && beta4 <= Uvarp4) pc4[o]<-1
if (beta5 >= Lvarp5 && beta5 <= Uvarp5) pc5[o]<-1
if (beta6 >= Lvarp6 && beta6 <= Uvarp6) pc6[o]<-1
if (beta7 >= Lvarp7 && beta7 <= Uvarp7) pc7[o]<-1
if (beta8 >= Lvarp8 && beta8 <= Uvarp8) pc8[o]<-1

censura[o,] = table(d)[1]/N
}
```

```

cat(o,"      ",ite,"  ",round(sum(pc1)/o,3),"  ",round(sum(pc2)/o,3),"
    ",round(sum(pc3)/o,3),"  ",round(sum(pc4)/o,3),"  ",round(sum(pc5)/o,3),"
    ",round(sum(pc6)/o,3),"  ",round(sum(pc7)/o,3),"  ",round(sum(pc8)/o,3),
    "\n"); o<-(o+1);
}

}}

ite<-ite+1
}

qtite  <-c(ite)
s_cens = mean(censura[,1])

ESTI1c <-c(mean(emle[,1]),mean(emle[,2]),mean(emle[,3]),mean(emle[,4])
,mean(emle[,5]),mean(emle[,6]),mean(emle[,7]),mean(emle[,8]))
BIAS1c <-c(mean(emle[,1])-beta1, mean(emle[,2])-beta2,
mean(emle[,3])-beta3, mean(emle[,4])-beta4,
    mean(emle[,5])-beta5, mean(emle[,6])-beta6,
    mean(emle[,7])-beta7, mean(emle[,8])-beta8)
RMSE1c <-c(sqrt(mean((emle[,1]-beta1)^2)),sqrt(mean((emle[,2]-beta2)^2)),
sqrt(mean((emle[,3]-beta3)^2)),sqrt(mean((emle[,4]-beta4)^2)),
sqrt(mean((emle[,5]-beta5)^2)),sqrt(mean((emle[,6]-beta6)^2)),
sqrt(mean((emle[,7]-beta7)^2)),sqrt(mean((emle[,8]-beta8)^2)))
PCc    <-c(sum(pc1)/B,sum(pc2)/B,sum(pc3)/B,sum(pc4)/B,
sum(pc5)/B,sum(pc6)/B,sum(pc7)/B,sum(pc8)/B)

return(list(qtite  = qtite  ,
s_cens  = s_cens ,
ESTI1c  = ESTI1c ,
BIAS1c  = BIAS1c ,
RMSE1c  = RMSE1c ,
PCc      = PCc    ) )
}

```

```
numCores <- 5
cl <- makeCluster(numCores)
finalResults1 <- parLapply(cl, c(100,250,500,750,1000), simFunc)
stopCluster(cl)
```

```
#Plot
```

```
beta1 = 0.5
beta2 = 0.5
beta3 = 1.5
beta4 = 2
beta5 = -3
beta6 = 1
beta7 = -2
beta8 = 0.75
beta1c=c(beta1,beta2,beta3,beta4,beta5,beta6,beta7,beta8)

size = c(100,250,500,750,1000)
```

```
AV1_c = rbind( finalResults1[[1]]$ESTI1c,
finalResults1[[2]]$ESTI1c,
finalResults1[[3]]$ESTI1c,
finalResults1[[4]]$ESTI1c,
finalResults1[[5]]$ESTI1c)
```

```
RMSE1_c = rbind(finalResults1[[1]]$RMSE1c,
finalResults1[[2]]$RMSE1c,
finalResults1[[3]]$RMSE1c,
finalResults1[[4]]$RMSE1c,
finalResults1[[5]]$RMSE1c)
```

```
B1_c = rbind( finalResults1[[1]]$BIAS1c,
finalResults1[[2]]$BIAS1c,
```

```
finalResults1[[3]]$BIAS1c,
finalResults1[[4]]$BIAS1c,
finalResults1[[5]]$BIAS1c)

cob1_c = rbind(finalResults1[[1]]$PCc,
finalResults1[[2]]$PCc,
finalResults1[[3]]$PCc,
finalResults1[[4]]$PCc,
finalResults1[[5]]$PCc)

par(mfrow=c(4,3))

## Param[5]
a5    =    0# min(RMSE1_c[,5]) - 0.25
b5    =          max(RMSE1_c[,5]) + 0.25
aa5   =          min(B1_c[,5]) - 0.25
bb5   =          max(B1_c[,5]) + 0.25

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,5],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa5,bb5),ylab =
      expression(paste("BIAS ( ", hat(beta)[10], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,5],xlim=c(min(size),
max(size)), xlab="",ylim=c(a5,b5),ylab =
      expression(paste("RMSE ( ", hat(beta)[10], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")
```

```

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,5],xlim=c(min(size),max(size)),
      xlab="",ylim=c(0.80,1),ylab=
      expression(paste("CP", " (",hat(beta)[10],")")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

```

```
## Param[6]
```

```

a6   =   0# min(RMSE1_c[,6]) - 0.25
b6   =           max(RMSE1_c[,6]) + 0.25
aa6  =           min(B1_c[,6]) - 0.25
bb6  =           max(B1_c[,6]) + 0.25

```

```

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,6],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa6,bb6),ylab =
      expression(paste("BIAS ( ", hat(beta)[11], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

```

```

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,6],xlim=c(min(size),max(size)),
      xlab="",ylim=c(a6,b6),ylab =
      expression(paste("RMSE ( ", hat(beta)[11], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

```

```

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,6],xlim=c(min(size),max(size)),
      xlab="",ylim=c(0.80,1),ylab=
      expression(paste("CP", " (",hat(beta)[11],")")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

```

```
## Param[7]
a7 = 0# min(RMSE1_c[,7])- 0.25
b7 = max(RMSE1_c[,7])+ 0.25
aa7 = min(B1_c[,7])- 0.25
bb7 = max(B1_c[,7])+ 0.25

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,7],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa7,bb7),ylab =
      expression(paste("BIAS ( ", hat(beta)[20], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,7],xlim=c(min(size),max(size)),
      xlab="",ylim=c(a7,b7),ylab =
      expression(paste("RMSE ( ", hat(beta)[20], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,7],xlim=c(min(size),max(size)),
      xlab="",ylim=c(0.80,1),ylab=
      expression(paste("CP", " (",hat(beta)[20],",")")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

## Param[8]
a8 = 0# min(RMSE1_c[,8])- 0.25
b8 = max(RMSE1_c[,8])+ 0.25
aa8 = min(B1_c[,8])- 0.25
bb8 = max(B1_c[,8])+ 0.25
```

```
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,8],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa8,bb8),ylab =
      expression(paste("BIAS ( ", hat(beta)[21], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")
```

```
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,8],xlim=c(min(size),max(size)),
      xlab="",ylim=c(a8,b8),ylab =
      expression(paste("RMSE ( ", hat(beta)[21], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")
```

```
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,8],xlim=c(min(size),max(size)),
      xlab="",ylim=c(0.80,1),ylab=
      expression(paste("CP", " (",hat(beta)[21],")")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")
dev.off()
```

```
par(mfrow=c(4,3))
## Param[1]
a1  = 0#min(RMSE1_c[,1])- 0.25
b1  = max(RMSE1_c[,1])+ 0.25
aa1 = min(B1_c[,1])- 0.25
bb1 = max(B1_c[,1])+ 0.25
```

```
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,1],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa1,bb1),ylab =
      expression(paste("BIAS ( ", hat(beta)[30], " )")),type="o", pch =
```

```

49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,1],xlim=c(min(size),max(size)),
      xlab="",ylim=c(a1,b1),ylab =
      expression(paste("RMSE ( ", hat(beta)[30], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,1],xlim=c(min(size),max(size)),
      xlab="",ylim=c(0.80,1),ylab=
      expression(paste("CP", " (", hat(beta)[30], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

## Param[2]
a2 = 0# min(RMSE1_c[,2])- 0.25
b2 = max(RMSE1_c[,2])+ 0.25
aa2 = min(B1_c[,2])- 0.25
bb2 = max(B1_c[,2])+ 0.25

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,2],xlim=c(min(size),max(size)),
      xlab="",ylim=c(aa2,bb2),ylab =
      expression(paste("BIAS ( ", hat(beta)[31], " )")),type="o", pch =
      49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,2],xlim=c(min(size),max(size)),
      xlab="",ylim=c(a2,b2),ylab =
      expression(paste("RMSE ( ", hat(beta)[31], " )")),type="o", pch =

```

```

    49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,2],xlim=c(min(size),max(size)),
     xlab="",ylim=c(0.80,1),ylab=
     expression(paste("CP", " (",hat(beta)[31],")")),type="o", pch =
     49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

## Param[3]
a3   =   0# min(RMSE1_c[,3])- 0.25
b3   =       max(RMSE1_c[,3])+ 0.25
aa3  =       min(B1_c[,3])- 0.25
bb3  =       max(B1_c[,3])+ 0.25

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,3],xlim=c(min(size),max(size)),
     xlab="",ylim=c(aa3,bb3),ylab =
     expression(paste("BIAS ( ", hat(beta)[40], " )")),type="o", pch =
     49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,3],xlim=c(min(size),max(size)),
     xlab="",ylim=c(a3,b3),ylab =
     expression(paste("RMSE ( ", hat(beta)[40], " )")),type="o", pch =
     49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,3],xlim=c(min(size),max(size)),
     xlab="",ylim=c(0.80,1),ylab=
     expression(paste("CP", " (",hat(beta)[40],")")),type="o", pch =

```

```

49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")

## Param[4]
a4    =    0# min(RMSE1_c[,4])- 0.25
b4    =    max(RMSE1_c[,4])+ 0.25
aa4   =    min(B1_c[,4])- 0.25
bb4   =    max(B1_c[,4])+ 0.25

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,B1_c[,4],xlim=c(min(size),max(size)),
xlab="",ylim=c(aa4,bb4),ylab =
expression(paste("BIAS ( ", hat(beta)[41], " )")),type="o", pch =
49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,RMSE1_c[,4],xlim=c(min(size),max(size)),
xlab="",ylim=c(a4,b4),ylab =
expression(paste("RMSE ( ", hat(beta)[41], " )")),type="o", pch =
49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0, v = 1, col = "gray60")

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size,cob1_c[,4],xlim=c(min(size),max(size)),
xlab="",ylim=c(0.80,1),ylab=
expression(paste("CP"," (",hat(beta)[41],")")),type="o", pch =
49,col="1",lwd=1,lty=2,main=NULL)
abline(h = 0.95, v = 0.95, col = "gray60")
dev.off()

par(mfrow=c(4,2))
inf5   =    min(AV1_c[,5]) - 0.25
sup5   =    max(AV1_c[,5]) + 0.25

```

```

par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,5],xlim=c(min(size),max(size)),
      ylim=c(inf5,sup5), xlab="",ylab =
      expression(paste(hat(beta)[10])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[5], v = 1, col = "gray60")

inf6 = min(AV1_c[,6]) - 0.25
sup6 = max(AV1_c[,6]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,6],xlim=c(min(size),max(size)),
      ylim=c(inf6,sup6), xlab="",ylab =
      expression(paste(hat(beta)[11])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[6], v = 1, col = "gray60")

inf7 = min(AV1_c[,7]) - 0.25
sup7 = max(AV1_c[,7]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,7],xlim=c(min(size),max(size)),
      ylim=c(inf7,sup7), xlab="",ylab =
      expression(paste(hat(beta)[20])),type="o", pch = 49,
      col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[7], v = 1, col = "gray60")

inf8 = min(AV1_c[,8]) - 0.25
sup8 = max(AV1_c[,8]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,8],xlim=c(min(size),max(size)),
      ylim=c(inf8,sup8), xlab="",ylab =
      expression(paste(hat(beta)[21])),type="o", pch = 49,
      col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[8], v = 1, col = "gray60")

```

```
inf1 = min(AV1_c[,1]) - 0.25
sup1 = max(AV1_c[,1]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,1],xlim=c(min(size),max(size)),
      ylim=c(inf1,sup1), xlab="",ylab =
      expression(paste(hat(beta)[30])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[1], v = 1, col = "gray60")
```

```
inf2 = min(AV1_c[,2]) - 0.25
sup2 = max(AV1_c[,2]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,2],xlim=c(min(size),max(size)),
      ylim=c(inf2,sup2), xlab="",ylab =
      expression(paste(hat(beta)[31])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[2], v = 1, col = "gray60")
```

```
inf3 = min(AV1_c[,3]) - 0.25
sup3 = max(AV1_c[,3]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,3],xlim=c(min(size),max(size)),
      ylim=c(inf3,sup3), xlab="",ylab =
      expression(paste(hat(beta)[40])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
abline(h = beta1c[3], v = 1, col = "gray60")
```

```
inf4 = min(AV1_c[,4]) - 0.25
sup4 = max(AV1_c[,4]) + 0.25
par(mai=c(0.45,0.9,0.1, 0.1))
plot(size, AV1_c[,4],xlim=c(min(size),max(size)),
      ylim=c(inf4,sup4), xlab="",ylab =
      expression(paste(hat(beta)[41])),type="o",
      pch = 49,col="1",lwd=1,lty=2,main=NULL)
```

```
abline(h = beta1c[4], v = 1, col = "gray60")
dev.off()

save.image("model1.RData")
```


واژه‌نامه فارسی به انگلیسی

Logistics	لجستیک
Survival Analysis	تحلیل بقا
Censore	سانسور
Weibull	وایبل
Cut	برش
Likelihood	درست‌نمایی
Newton-Raphson	نیوتون-رافسون
Information Matrix	ماتریس اطلاع
Hessian	هسیان
Inflated Zero	آماسیده در صفر
Cure Rate	نرخ شفایافتگی
Mixture	آمیخته
Simulation	شبیه‌سازی

واژه‌نامه انگلیسی به فارسی

Logistics	لجستیک
Survival Analysis	تحلیل بقا
Censore	سانسور
Weibull	وایبل
Cut	برش
Likelihood	درست‌نمایی
Newton-Raphson	نیوتون-رافسون
Information Matrix	ماتریس اطلاع
Hessian	هسیان
Inflated Zero	آماسیده در صفر
Cure Rate	نرخ شفایافتگی
Mixture	آمیخته
Simulation	شبیه‌سازی

Aabstract

Banks and financial institutions give their customers loans under some conditions, always. customers can be divided in to three groups based on their performance. The first group of customers cut off their relationship with the bank, completely. This group of customers are called dishonest which their survival time to repay the loan is equal to zero. The second group is those customers who do not care about their loan repayment. Contrary to the first group, they repay their loan somedeal and then cut off. This group has positive survival time of repayment of loan, but finally they are turned to dishonest one. We call these customers defaulter. At last, the third group is the customers who repay their loan in a timely manner.

In order to maximize the profit of a bank, it is necessary to minimize the rates of dishonest and defaulter customers and maximize good customers (the third group). Therefore, analyzing the survival time of loan repayment can be vital in any bank. Analyzing of the timelength of the data till the occurance of a phenomenon (here, repayment of loan) places in the survival analysis domain. In the analysis of the survival data, the main problem is to find a suitable model for measuring the affection of the various factors on the length of the survival time.

Accordingly to the nature of the data attained from the bank customers, the survival of non-repayment of dishonest and defaulter customers do not tend to zero. A suitable category for these type of the survival data is Healing rate models. For these models, two subclasses are introduced: mixture and non-mixture models. There is a problem with using of the Healing rate models for analyzing the borrowers behaviour, and that is the first group of customers which their survival repayment tend to zero. This case causes an unwanted fitness in the Healing rate model. In such cases, it is recommended to use zero-inflated models. In this thesis, first we introduce the zero-inflated rate model and then it for data analysis we will refund some of our clients to four types of registered banks in the central bank of Iran.

Keywords: cure rate, survival analysis, zero-inflated, bank loans



Shahrood University of Technology

Faculty Of Mathematical Sciences

MSc Thesis in: Statistics

**A zero-inflated cure rate regression model
and its application in bank loans**

By: Mona Abedi Mehr

Supervisor

Negar Eghbal

Advisor

Hossein Baghishani

January 2018