مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۴، دوره ۱۵، شماره ۴، ص ۱۰۲–۱۱۹

شاپا چاپی: ۸۰۸۸-۲۲۵۱
DOI: 10.22055/jamm.2026.48195.2324

مدل‌بندی ژنتیکی زمان تا رخداد چاقی موش‌ها با رگرسیون فراوانی‌گرا و بیزی

حدیث راسخ فتوحی و سید مرتضی امینی [۱]

گروه آمار، دانشکده ریاضی آمار و علوم کامپیوتر، دانشکدگان علوم، دانشگاه تهران

دبیر مسئول: علیرضا فخارزاده جهرمی

تاریخ دریافت: ۱۴۰۳/۰۷/۲۹ تاریخ پذیرش: ۱۴۰۴/۱۱/۰۴

چکیده: در این مقاله، ارتباط آماری بین پلی‌مورفیسم‌های تک‌نوکلئوتیدی گونه‌ای از موش‌ها با وزن این موش‌ها که در طی شانزده هفته جمع‌آوری شده بودند، با استفاده از مدل رگرسیون مورد بررسی قرار گرفته است. مسئله اصلی مورد بررسی در این پژوهش مسئله زمان تا رخداد است که در تحلیل بقا مورد توجه پژوهش‌گران است. در واقع در این مقاله به مدل‌بندی زمان تا رخداد (چاقی) بر اساس ژن‌ها به عنوان متغیر کمکی و پیش‌گو پرداخته شده است. با توجه به بعد بالای مسئله و وجود تعداد زیادی از ژن‌های بی‌تأثیر در رخداد چاقی، مسئله انتخاب متغیر و مدل‌بندی تنک نیز مورد بررسی قرار گرفته است. دو روش فراوانی‌گرای رگرسیون کاکس لاسو و روش بیزی رگرسیون وایبل در این مقاله معرفی شده و مقایسه می‌شوند. بررسی‌های شبیه‌سازی و مطالعه دادگان واقعی نشان می‌دهد که روش بیزی بر اساس معیارهای گوناگون به روش فراوانی‌گرا برتری دارد.

واژه‌های کلیدی: مطالعات همبستگی سراسر ژنوم، زمان تا رخداد، پلی‌مورفیسم‌های تک‌نوکلئوتیدی، روش رگرسیون کاکس لاسو، روش رگرسیون وایبل بیزی

رده‌بندی ریاضی: 33C45; 49-XX

۱ مقدمه

مطالعات همبستگی سراسر ژنوم، (GWAS)[۲] درک ما از معماری ژنتیکی صفات پیچیده را بسیار گسترش داده است، اما این مطالعات تا حد زیادی بر روی فنوتیپ‌های[۳] دوتایی و صفات کمّی تمرکز کرده‌اند [۲۳] و سن شروع علائم را با وجود اینکه یکی از ویژگی‌های کلیدی در مطالعات بیوبانک بیماری‌های مرتبط با سن می‌باشد، کمتر مورد توجه قرار داده‌اند. درک اساس محیطی و ژنتیکی در زمانی که اولین علائم رخ

[۱] نویسنده مسئول مقاله
morteza.amini@ut.ac.ir (M. Amini), h.fotoohi@ut.ac.ir (H.R. Fotouhi)

[2] Genome wide association studies
[3] Phenotypes

می‌دهد برای برنامه‌های غربالگری اولیه و برای به دست آوردن بینش در مورد توسعه و پیشرفت بیماری مهم است، به ویژه این که فرآیندهای پاتولوژیک بسیاری از بیماری‌های مرتبط با سن ممکن است چندین دهه قبل از ظهور اولین علائم آغاز شوند. شواهد نشان می‌دهد که تحلیل ژنگان که با فنوتیپ‌های مورد کنترل انجام می‌شوند، در مقایسه با روش‌های دیگر تحلیل در سن شروع، قدرت کمتری دارند [۱۸، ۱۹]. با این حال، درک ما از معماری ژنتیکی زمان‌بندی تولید مثل، و سنی که در آن اولین علائم برای اختلالات پیچیده رایج ایجاد می‌شود، محدود است. یکی از حوزه‌های پژوهشی فعال در مطالعات بالینی و فارماکوژنیک حوزه مدل‌سازی آماری داده‌های زمان تا رویداد است. مشابه رگرسیون نشانگر منفرد در تحلیل‌های GWAS، یک مدل نرخ خطر متناسب کاکس[۴] (PH) [۲] برای هر چندریختی تک‌نوکلئوتیدی[۵](SNP) می‌تواند بدین منظور استفاده شود [۱۶، ۲۰، ۲۱].

اخیراً با استفاده از برخی تقریب‌ها برای رگرسیون نرخ خطر متناسب کاکس تک نشانگر، زمان انجام محاسبات رایانه‌ای تا حدی کاهش یافته است [۱]، با این حال، این رویکرد همچنان تنها به براورد اندازه اثر حاشیه‌ای می‌پردازد. روش‌های مبتنی بر مانده‌ها نیز به طور گسترده مورد استفاده قرار گرفته‌اند، که ابتدا فنوتیپ را بر روی متغیرهای کمکی مانند جنسیت یا سن رگرسیون کرده و سپس از مانده‌ها در مدل رگرسیون دوم روی داده‌های SNP استفاده می‌کند. در این روش مانده‌های مارتینگل در مرحله اول به دست می‌آیند [۹، ۱۰]. در مرحله دوم این مانده‌ها با استفاده از مدل نرخ خطر متناسب کاکس روی SNPها رگرسیون می‌شود. روش مانده‌ها مارتینگل می‌تواند بسیار سریع باشد، با این حال، زمان شکست و شاخص سانسور به‌جای اینکه اطلاعات سانسور را به طور خاص از طریق تابع درست‌نمایی وارد کنند، با یک آمار خلاصه ترکیب می‌شوند. بنابراین، این رویکرد از اطلاعات سانسور به طور مؤثر استفاده نمی‌کند که قدرت این مدل را کاهش می‌دهد. علاوه بر این می‌توان در مدل نرخ خطر متناسب کاکس اثرات تصادفی را نیز در نظر گرفت. از جمله پژوهش‌های اخیر در زمینه مدل‌سازی داده‌های زمان تا رخداد با مدل‌های رگرسیونی می‌توان به [۳، ۸، ۱۴] نیز اشاره کرد.

در این پژوهش، یک رویکرد جایگزین برای به دست آوردن نتایج درست در مجموعه داده‌های توالی فنوتیپ – ژنوتیپ[۶] در مقیاس کامل، با پیشنهاد مدل رگرسیون آمیخته با انتخاب متغیر را در چارچوب بیزی مورد مطالعه قرار می‌دهیم.

۲ مفاهیم مقدماتی

۱.۲ استنباط بیزی

استنباط بیزی[۷]، فرآیند برازش یک مدل احتمال به مجموعه‌ای از داده‌ها و خلاصه کردن نتیجه با توزیع احتمال بر روی پارامترهای مدل و مقادیر مشاهده نشده مانند پیش‌بینی برای مشاهدات جدید است. استنباط بیزی که به نام کشیش توماس بیز (۱۷۰۱–۱۷۶۱) نامگذاری شده است، به قضیه بیز متکی است. از استنباط بیزی به طور گسترده در کاربردهای گوناگون استفاده می‌شود.

با شرطی کردن بر روی مقدار مشاهده شده از داده‌ها y و استفاده از ویژگی اصلی احتمال شرطی که قاعده بیز نامیده می‌شود، چگالی پسین به دست می‌آید

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{۱.۲}$$

که برای θ گسسته $p(y) = \sum p(\theta)p(y|\theta)$ و برای θ پیوسته $p(y) = \int p(\theta)p(y|\theta)d\theta$ می‌باشد.

به تابع وزن $p(\theta)$ در رابطه (۱.۲) پیشین و به $p(\theta|y)$ توزیع پسین گفته می‌شود. لازم به ذکر است که $p(\theta)$ می‌تواند تابع چگالی یا جرم احتمال نباشد که در این حالت به آن پیشین ناسره[۸] گفته می‌شود.

می‌توان نوشت

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{۲.۲}$$

رابطه بالا هسته اصلی محاسبه توزیع پسین را نشان می‌دهد که از ضرب تابع درست‌نمایی در پیشین تشکیل می‌شود.

توزیع پیشین نمایانگر باور در مورد مقدار واقعی یک پارامتر است، قبل از اینکه هر گونه شواهد جدیدی در نظر گرفته شود. این بهترین حدس در مورد مقدار پارامتر است. به طور کلی، اگر پارامتر را به عنوان θ نشان دهیم، توزیع پیشین با $p(\theta)$ نشان داده می‌شود. توزیع پسین از توزیع پیشین و تابع درست‌نمایی به دست می‌آید. در واقع توزیع پسین نمایانگر اطلاعات به روز شده در مورد پارامتر پس از در نظر گرفتن شواهد جدیدی است که از داده‌ها به دست آمده است.

[4] Cox Proportional Hazards Model
[5] Single nucleotide polymorphism
[6] Genotype
[7] Bayesian inference
[8] Improper

توزیع پیشین فاقد اطلاع: توزیع پیشین فاقد اطلاع[۹]، که همچنین به عنوان توزیع پراکنده[۱۰] شناخته می‌شود، اطلاعات مبهم یا عمومی در مورد یک متغیر را بیان می‌کند. این توزیع اغلب زمانی استفاده می‌شود که دانش پیشین در مورد پارامتر وجود ندارد. اگر پارامتر با θ نشان داده شود، یک توزیع پیشین فاقد اطلاع می‌تواند به عنوان یک توزیع یکنواخت بر روی مجموعه مقادیر پارامتر نشان داده شود.

دو رویکرد مختلف برای توزیع پیشین فاقد اطلاع وجود دارد. در رویکرد اول توزیع پیشین تخت و یا تا حد امکان تخت مانند توزیع یکنواخت و یا توزیع دارای واریانس بسیار بزرگ به عنوان توزیع پیشین فاقد اطلاع استفاده می‌شود. در رویکرد دوم که رویکرد جفریز[۱۱] نام دارد، توزیع پیشین فاقد اطلاع توزیعی است که تحت تبدیلات یکنوا خانواده توزیع‌های آن تغییر نکند و بر اساس مجذور اطلاع فیشر به دست می‌آید.

توزیع پیشین مزدوج: توزیع پیشین مزدوج[۱۲]، توزیع پیشینی است که هنگام ضرب با تابع درستنمایی، منجر به توزیع پسینی می‌شود که در یک خانواده با پیشین است. اهمیت توزیع پیشین مزدوج در این است که با شناخت خانواده توزیع پسین دیگر نیازی به محاسبه ضریب نرمال‌سازی $p(y)$ نیست. با این حال در بسیاری از مسائل توزیع پیشین مزدوج وجود ندارد و به دلیل بعد بالای پارامتر، امکان محاسبه ضریب نرمال‌سازی نیز وجود ندارد. در چنین شرایطی از روش‌های تقریب توزیع پسین استفاده می‌شود، که متداول‌ترین آن‌ها نمونه‌گیری زنجیره مارکوف مونت‌کارلو است.

۱.۱.۲ الگوریتم متروپولیس هستینگز

الگوریتم متروپولیس – هستینگز متداول‌ترین نوع MCMC است. نمونه‌های جدید را بر اساس یک توزیع پیشنهادی، ارائه می‌دهد و با استفاده از نسبت پذیرش آن‌ها را می‌پذیرد یا رد می‌کند و برای پارامترهای پیوسته و گسسته مناسب است. این الگوریتم در الگوریتم ۱ آمده است.

الگوریتم ۱ الگوریتم متروپولیس–هستینگز

ورودی: توزیع پیشین $p(\theta)$، تابع درستنمایی $p(y|\theta)$، توزیع پیشنهادی $q(\theta'|\theta)$، تعداد نمونه‌ها N
خروجی: دنباله‌ای از نمونه‌ها $\theta_1, \theta_2, \ldots, \theta_N$

- انتخاب مقدار اولیه θ_0
- برای $i = 1$ تا N:
  - نمونه‌گیری θ' از توزیع پیشنهادی $q(\theta'|\theta_i)$
  - محاسبه نسبت پذیرش:

$$\alpha = \min\left(1, \frac{p(\theta')p(y|\theta')q(\theta_i|\theta')}{p(\theta_i)p(y|\theta)q(\theta'|\theta_i)}\right)$$

  - نمونه‌گیری از متغیر یکنواخت $u \sim U(0,1)$
  - اگر $u \leq \alpha$: $\theta_{i+1} = \theta'$
  - در غیر این صورت: $\theta_{i+1} = \theta_i$
- دنباله نمونه‌ها $\{\theta_1, \theta_2, \ldots, \theta_N\}$ را بازگشت دهید.

[9] Non-informative prior distribution
[10] Diffuse prior
[11] Jeffreys
[12] Conjugate prior distribution

۲.۲ تحلیل بقا

تحلیل بقا[۱۳] شاخه‌ای از آمار است که به مطالعه و مدل‌سازی زمان وقوع یک رویداد خاص می‌پردازد. این رویداد می‌تواند وقوع مرگ، بروز بیماری، شکست یک قطعه صنعتی، یا هر رویداد دیگری باشد که برای آن زمان وقوع اهمیت دارد. تحلیل بقا در بسیاری از زمینه‌ها از جمله پزشکی، مهندسی، علوم اجتماعی و اقتصاد کاربرد دارد.

۱.۲.۲ تابع بقا

تابع بقا [۱۴] احتمال آن را که یک فرد یا سیستم تا زمان مشخصی زنده بماند، نشان می‌دهد. این تابع معمولاً به صورت زیر تعریف می‌شود:

$$S(t) = P(T > t)$$

که در آن T زمان بقا و t زمان مورد نظر است.

۲.۲.۲ تابع نرخ خطر

تابع نرخ خطر، که با $h(t)$ نشان داده می‌شود، در زمان t، پتانسیل لحظه‌ای در هر واحد زمان برای وقوع رخداد مورد نظر است، با توجه به اینکه فرد تا زمان t زنده مانده است. به صورت ریاضی، این به صورت زیر تعریف می‌شود:

$$h(t) = \lim_{\Delta t \to ۰} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{۳.۲}$$

که در آن T زمان رویداد است.

۳.۲.۲ نسبت خطر

نسبت خطر[۱۵] (HR) در تجزیه و تحلیل بقا، خطر نسبی یک رویداد (مانند مرگ، پیشرفت بیماری یا هر پیامد دیگری) را بین دو گروه مجزا اندازه‌گیری می‌کند. این گروه‌ها معمولاً با سطوح مختلف یک متغیر درمانی مورد علاقه مشخص می‌شوند (به عنوان مثال، درمان شده در مقابل کنترل).

به طور خاص، (HR) نشان دهنده نسبت نرخ خطر (خطرات آنی) تجربه رخداد در این دو گروه است.

به عنوان مثال، در یک آزمایش بالینی، (HR) بین دو گروه درمان به صورت زیر خواهد بود

$$\mathrm{HR} = \frac{h_۱(t)}{h_۰(t)} \tag{۴.۲}$$

که در آن $h_۰(t)$ و $h_۱(۱)$ به ترتیب نرخ خطر در گروه‌های ۰ و ۱ هستند. اگر $\mathrm{HR} > ۱$ باشد، نرخ خطر در گروه ۱ بیشتر و اگر $\mathrm{HR} < ۱$ نرخ خطر در گروه ۰ بیشتر است.

۳.۲ مفاهیم زیستی

DNA و کروموزوم: مولکول DNA، یا اسید دئوکسی ریبونوکلئیک اسید، ماده ارثی در انسان و تقریباً تمام موجودات دیگر است. تقریباً تمام سلول‌های بدن یک فرد حاوی نسخهٔ DNA مشابهی است. بیشترین DNA در هسته سلول قرار دارد. اما مقدار کمی از DNA نیز در میتوکندری یافت می‌شود.

اطلاعات در DNA با استفاده از کُد ساخته شده از چهار باز آلی: آدنین (A)، گوانین (G)، سیتوزین (C) و تایمین (T) ذخیره می‌شود. DNA انسانی شامل حدود ۳ میلیارد باز است و بیش از ۹۹ درصد از باز در همه افراد یکسان است.

بازهای DNA با یکدیگر، A با T و C با G جفت می‌شوند تا واحدهای به نام جفت باز ایجاد شوند. هر باز نیز به یک مولکول قند و یک مولکول فسفات وصل می‌شود. یک باز، قند و فسفات با هم، نوکلئوتید نامیده می‌شود. نوکلئوتیدها از دو رشته طولانی تشکیل شده‌اند

[13] Survival Analysis
[14] Survival Function
[15] hazard ratio

که یک مارپیچ را به نام مارپیچ دوتایی تشکیل می‌دهند. ساختار مارپیچ دوتایی تا حدودی شبیه یک نردبان است؛ جفت‌های باز نردبان‌ها را تشکیل می‌دهند و مولکول‌های قند و فسفات، قسمت‌های کناری نردبان را تشکیل می‌دهند.

در هسته هر سلول، مولکول DNA به ساختارهای رشته مانند به نام کروموزوم‌ها بسته‌بندی می‌شود. هر کروموزوم از DNAای که چندین بار به طور مارپیچی در اطراف پروتئین‌هایی به نام هیستون‌ها قرار گرفته‌اند، ساخته شده‌اند که این هیستون‌ها از ساختار کروموزوم پشتیبانی می‌کند.

ژن: ژن پایه‌ای‌ترین واحد فیزیکی و عملکردی ارثی است. ژن‌ها از DNA ساخته می‌شوند. بیشتر ژن‌ها به‌عنوان دستورالعملی برای تولید مولکول‌هایی به نام پروتئین عمل می‌کنند. بااین‌حال، برخی از ژن‌ها تولید کننده پروتئین نیستند. در انسان، طول ژن‌ها از چند صد باز DNA به بیش از ۲ میلیون باز تغییر می‌کند. پروژه ژنوم انسان تخمین زده است که انسان بین ۲۰۰۰۰ تا ۲۵۰۰۰ ژن دارد.

هر فرد دارای دو نسخه از هر ژن است که از هریک از پدر و مادر به ارث برده می‌شود. اکثر ژن‌ها در همه افراد یکسان هستند، اما تعداد کمی از ژن‌ها (کمتر از ۱ درصد از کل) بین افراد کمی متفاوت هستند.

پلی مورفیسم تک‌نوکلئوتیدی (SNP): یک پلی‌مورفیسم تک‌نوکلئوتیدی یک دنباله از جهش‌های DNA است و وقتی رخ می‌دهد که یک تک‌نوکلئوتید C،T G، یا A در ژنگان، در بین افراد مختلف از یک‌گونه و یا بین جفت کروموزوم‌های یک فرد تغییر می‌کند. برای مثال دو دنباله از DNA دو فرد مختلف را در نظر بگیریم: AAGCCTA و AAGCTTA . این دو دنباله در یک نوکلئوتید باهم تفاوت دارند. در این صورت گفته می‌شود که دو الل T و C وجود دارند. همه هاSNP فقط دارای دو الل هستند. در یک جمعیت یک مقدار به نام حداقل فراوانی اللی (MAF) برای هر SNP تعریف می‌شود که اشاره به اللی با کمترین فراوانی بین دو الل در یک SNP دارد [۶].

فنوتیپ یا رُخ نمود: خصوصیات و ویژگی‌هایی که در ظاهر ارگانیسم‌ها آشکار می‌شود، فنوتیپ نامیده می‌شود. فنوتیپی که در یک موجود زنده آشکار می‌شود، نتیجه بیان ژنوتیپ و تأثیر محیط می‌باشد. به همین دلیل است که وقتی ژنوتیپ در موجودات زنده یکی باشد، نمی‌تواند فنوتیپ یکسانی نیز داشته باشند چون محیط نیز بر آن تأثیر می‌گذارد [۴].

ژنوتیپ: ژنوتیپ یکی از سه عاملی است که فنوتیپ فرد را مشخص می‌نماید (دو عامل دیگر اپی ژنتیک ارثی و عوامل محیطی غیرارثی می‌باشند). بنابراین شبیه بودن ژنوتیپ دو موجود زنده تضمین‌کننده شباهت فنوتیپی آن دو موجود نیست. ژنوتیپ‌ها بیانگر نحوه چینش تفاوت‌ها در افراد و این تفاوت می‌تواند باعث افتراق موجودات از یکدیگر شود.

۳ مدل رگرسیون کاکس تنک

به طور کلی تحلیل بقا مجموعه‌ای از روش‌های آماری برای تحلیل داده‌هایی است که متغیر پاسخ در آن‌ها زمان تا رخداد یک پیشامد خاص است. منظور از زمان در تحلیل بقا می‌تواند تعداد سال‌ها، ماه‌ها، هفته‌ها یا روزها از شروع پیگیری یک فرد تا رخداد پیشامد مورد نظر برای وی باشد. همچنین می‌توان زمان را سن یک فرد در هنگام رخداد پیشامد در نظر گرفت.

پیشامد در تحلیل بقا ممکن است مرگ، وقوع بیماری، عود بیماری پس از بهبودی، باز یافتن توانایی و بهبودی (مثلاً بازگشت به کار)، یا هر تجربه مورد نظر دیگری که ممکن است برای فرد رخ دهد، باشد. در تحلیل‌های بقا معمولاً متغیر زمان را زمان بقا می‌نامیم زیرا این متغیر تعیین کننده مدت زمانی است که یک فرد در طول دوره پیگیری بقا یافته است. همچنین چون معمولاً در این نوع تحلیل‌ها، پیشامدهای مورد نظر مرگ، وقوع بیماری یا سایر تجربه‌های فردی است، پیشامد مورد نظر را شکست می‌نامیم. در حالی که زمان بقا ممکن است زمان تا برگشت به کار پس از یک عمل جراحی باشد، که در این صورت شکست یک پیشامد مثبت خواهد بود.

تحلیل داده‌های زمان تا رخداد برای درک رخدادها و زمان‌بندی آن‌ها بسیار مهم است که در این تحلیل با نمادها و اصطلاحاتی مواجه هستیم که به شرح زیر می‌باشد:

- T: نشان دهنده‌ی متغیر تصادفی زمان بقا، برای یک فرد است و بنابراین تکیه‌گاه این متغیر تصادفی اعداد نامنفی است.
- d: یک متغیر دودویی است و از آن برای نشان دادن وضعیت رخداد پیشامد (شکست) و یا سانسور شدن فرد استفاده می‌کنیم. $d = ۱$ برای زمانی است که در طول مطالعه، پیشامد مورد نظر برای فرد رخ داده و $d = ۰$ برای حالتی است که به یکی از دلایل گفته شده، زمان بقا برای فرد سانسور شده باشد.

کاکس[16] در سال ۱۹۷۲ اولین کسی بود که مدل زیر را ارائه داد:

$$h(t, x, \beta) = h_{۰}(t)e^{x\beta} \tag{۱.۳}$$

و نسبت خطر به صورت زیر است:

$$\mathrm{HR}(t, x_{۱}, x_{۰}) = e^{\beta(x_{۱}-x_{۰})} \tag{۲.۳}$$

[16] Cox

این مدل، در ادبیات با عناوین مختلفی همچون مدل کاکس[۱۷]، مدل خطرات نسبی کاکس یا به سادگی مدل خطرات نسبی[۱۸] شناخته می‌شود. بخشی از جذابیت مدل کاکس در تفسیر (۲.۳) به عنوان نوعی نسبت خطر نسبی است. در مدل (۱.۳) پارامتر β میزان تغییرات نمایی نرخ خطر را به ازای یک واحد تغییر در مقدار متغیر پیشگوی x نشان می‌دهد. در واقع به ازای یک واحد تغییر در متغیر x لگاریتم نرخ خطر به میزان β تغییر می‌کند. این مسئله در رابطه (۲.۳) به صورتی دیگر نیز بیان می‌شود. در این رابطه بیان می‌شود که با تغییر متغیر پیشگوی x از $x_۰$ به $x_۱$ نسبت نرخ خطر به صورت نمایی تعیین می‌شود. اگر $HR > ۱$ باشد، نرخ خطر $x_۱$ بیشتر از $x_۰$ است، و بالعکس.

می‌خواهیم روش بیشینه درست‌نمایی را به مدل خطرات متناسب در معادله (۱.۳) اعمال کنیم. فرض کنید n مشاهده مستقل داریم که هر کدام شامل اطلاعاتی در مورد طول زمانی که یک موضوع مشاهده شده است، این‌که آیا مشاهده یک زمان بقا بوده یا سانسور از راست شده و یک متغیر کمکی که مقدار آن در زمان شروع مشاهده تعیین می‌شود و در طول پیگیری موضوع در همان مقدار باقی می‌ماند. متغیرهای کمکی که مقادیر آن‌ها در مقدار اندازه‌گیری شده در ابتدای پیگیری ثابت می‌مانند، در عمل بیشتر مشاهده می‌شوند. برای سادگی و همچنین برای کاربرد پایه‌ای، بررسی‌ها را به متغیرهای کمکی که مقادیر آن‌ها ثابت است محدود می‌کنیم. بردار سه‌تایی از زمان مشاهده شده، متغیر کمکی و متغیر سانسور را به صورت (t_i, x_i, d_i)، $i = ۱, ۲, ..., n$ نشان می‌دهیم.

از آنجا که مشاهدات مستقل فرض شده‌اند، تابع درست‌نمایی کامل به صورت زیر است:

$$l(\beta) = \prod_{i=۱}^{n} [f(t_i, \beta, x_i)]^{d_i} \times [S(t_i, \beta, x_i)]^{۱-d_i} . \qquad (۳.۳)$$

برای به دست آوردن بیشینه درست‌نمایی نسبت به پارامتر مورد نظر، β، تابع لگاریتم درست‌نمایی را بیشینه می‌کنیم

$$\begin{aligned} L(\beta) &= \sum_{i=۱}^{n} [d_i \ln[f(t_i, \beta, x_i)] + (۱ - d_i) \ln[S(t_i, \beta, x_i)]] \\ &= \sum_{i=۱}^{n} \left[d_i \ln[h_۰(t_i)] + d_i x_i^T \beta + e^{x_i^T \beta} \ln[S_۰(t_i)] \right] . \end{aligned}$$

بیشینه درست‌نمایی کامل نیاز دارد که تابع درست‌نمایی را نسبت به پارامتر ناشناخته مورد نظر، β، و مخاطرات پایه و توابع بقا نامشخص $h_۰$ و $S_۰$ بیشینه کنیم. استفاده از تابع لگاریتم درست‌نمایی کامل ممکن نیست. کاکس در سال ۱۹۷۲ پیشنهاد استفاده از تابع درست‌نمایی جزئی را مطرح کرد. این تابع تنها به پارامتر مورد علاقه وابسته است. براوردگرهای پارامتر از تابع درست‌نمایی جزئی، همان خواص توزیعی براوردگرهای بیشینه درست‌نمایی کامل را خواهند داشت. تابع درست‌نمایی جزئی به صورت زیر داده شده است:

$$l_p(\beta) = \prod_{i=۱}^{n} \left[\frac{e^{x_i^T \beta}}{\sum_{j \in R(t_i)} e^{x_j^T \beta}} \right] \qquad (۴.۳)$$

که در آن مجموع در مخرج روی همه موضوعات در مجموعه ریسک در زمان t_i، که با $R(t_i)$ نشان داده می‌شود، انجام می‌گیرد. لازم به ذکر است که مجموعه ریسک شامل تمام موضوعاتی با زمان بقای بیشتر یا مساوی با زمان مشخص شده است.

عبارت در (۴.۳) فرض می‌کند که زمان‌ها برابر نیستند و اغلب برای حذف عباراتی که $d_i = ۰$ هستند، به صورت زیر اصلاح می‌شود:

$$l_p(\beta) = \prod_{i=۱}^{m} \left[\frac{e^{x_{(i)}^T \beta}}{\sum_{j \in R(t_{(i)})} e^{x_j^T \beta}} \right] \qquad (۵.۳)$$

که در آن حاصل‌ضرب بر روی m زمان‌های بقای مرتب شده متفاوت است و $x_{(i)}$ مقدار متغیر کمکی برای موضوعی با زمان بقای مرتب شده $t_{(i)}$ را نشان می‌دهد. تابع درست‌نمایی جزئی لگاریتمی به صورت زیر است:

[17] Cox model
[18] Proportional hazards model

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ x_{(i)}^T \beta - \log \left[\sum_{j \in R(t_{(i)})} e^{x_j^T \beta} \right] \right\} \tag{۶.۳}$$

مدل رگرسیون کاکس لاسو[19] مدل خطر متناسب کاکس را با لاسو ترکیب می‌کند که هدف آن شناسایی مجموعه‌ای پراکنده از ژن‌ها (یا ویژگی‌ها) است که عملکرد پیش‌آگاهی[20] بالایی در پیش‌بینی بقا دارند.

با اعمال روش لاسو در چارچوب رگرسیون کاکس، Cox–Lasso ژن‌های مرتبط را انتخاب می‌کند و در عین حال تأثیر آن‌ها بر زمان بقا را در نظر می‌گیرد.

لاسو و رگرسیون کاکس لاسو نقش مهمی در پیش‌بینی بقا با متعادل کردن انتخاب ویژگی[21] و منظم‌سازی[22] دارند. در حالی که لاسو بر انتخاب ژن فردی متمرکز است، Cox–Lasso این مفهوم را به تجزیه و تحلیل بقا گسترش می‌دهد، و آن را به ویژه برای تحقیقات سرطان و شناسایی نشانگرهای زیستی قوی بقا مفید می‌کند.

لاسو تعمیم یافته در واقع یک تکنیک منظم سازی برای براورد مدل‌های خطی تعمیم یافته (GLM) است. لاسو تعمیم یافته رگرسیون لاسو را به این مدل‌ها تعمیم می‌دهد.

مدل‌های خطی تعمیم‌یافته مدل‌های رگرسیون مختلفی از جمله رگرسیون خطی، رگرسیون لجستیک، رگرسیون پواسون و مانند آن را در بر می‌گیرند، که توزیع متغیر پاسخ در آن‌ها می‌تواند از خانواده نمایی بوده و لزوماً گاوسی نباشد. همانند لاسوی خطی، لاسو تعمیم یافته شامل یک عبارت تاوان است که اندازه ضرایب براورد شده را محدود می‌کند.

مسئله بهینه‌سازی رگرسیون لاسو تعمیم یافته می‌تواند به صورت زیر نمایش داده شود:

$$\max \left(\sum_{i=1}^{n} \ell(y_i; \sum_{j=0}^{p} \beta_j x_{ij}) - \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{۷.۳}$$

که در آن $\ell(\cdot;\cdot)$ تابع لگاریتم درست‌نمایی در مدل GLM است.

مدل رگرسیون لاسو دو محدودیت بزرگ دارد. اولین محدودیت این است که مدل لاسو در صورتی که $p > n$ باشد، حداکثر می‌تواند n ضریب را به صورت غیرصفر براورد کند. این مسئله به طور خاص در مدل‌های ژنتیکی یک چالش است، زیرا در این مدل‌ها تعداد ژن‌ها بسیار بزرگ‌تر از تعداد نمونه‌ها است و تعداد ژن‌های مؤثر ممکن است بسیار بیشتر از تعداد نمونه باشد. دومین محدودیت مدل لاسو این است که اگر دو متغیر کمکی با یکدیگر رابطه خطی قوی داشته باشند و یا به عبارت دیگر در مدل هم‌خطی چندگانه وجود داشته باشد، روش لاسو از بین این دو متغیر، تنها یکی را در مدل نگاه داشته و دیگری را حذف می‌کند. این مسئله نیز در مدل‌های ژنتیکی یک معضل بزرگ است، زیرا پژوهش‌گران ژنتیک به دنبال کشف تمامی ژن‌های مؤثر در یک فنوتیپ هستند و به طور خاص نیز می‌خواهند بدانند کدام ژن‌ها با هم به صورت گروهی کار می‌کنند. مدل رگرسیون تور کشسان، علاوه بر برطرف کردن دو محدودیت بیان شده، خاصیت گروه‌بندی را نیز دارد، بدین معنی که ضریب رگرسیونی متغیرهای کمکی دارای همبستگی بالا با یکدیگر را یکسان براورد می‌کند. نام‌گذاری این روش از سوی افرون به نام تور کشسان[23] نیز به این دلیل است که این روش مانند یک تور ماهی‌گیری کشسان، تمام ماهی‌های بزرگ را صید می‌کند، به این معنی که تمام متغیرهای کمکی تأثیرگذار را چه هم‌خطی چندگانه داشته باشند و یا نداشته باشند در مدل نگاه می‌دارد.

تور کشسان در واقع دارای یک عبارت منظم‌سازی که ترکیبی از تاوان‌های L_1 و L_2 است می‌باشد و مسئله بهینه‌سازی را به صورت زیر مدل‌بندی می‌کند:

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \left[-\frac{1}{n} \sum_{i=1}^{n} \ell(y_i; \beta_0 - x_i^T \beta) + \lambda\big((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1\big) \right] \tag{۸.۳}$$

با جایگذاری تابع لگاریتم درست‌نمایی مدل رگرسیون کاکس (۶.۳) در مدل رگرسیون لاسوی تعمیم‌یافته (۷.۳) به مدل کاکس لاسو دست می‌یابیم که می‌تواند برای تنک‌سازی مدل رگرسیون کاکس استفاده شود. همچنین می‌توان از مدل رگرسیون تورکشسان تعمیم‌یافته (۸.۳) نیز با تابع لگاریتم درست‌نمایی کاکس بهره جست.

[19] COX-Lasso
[20] prognostic
[21] feature selection
[22] regularization
[23] Elastic-Net

۴ مدل رگرسیون وایبل سانسورشده بیزی

در این بخش از یک حالت خاص از مدل ارائه شده توسط [۱۵] استفاده کرده و آن را با مدل کاکس تنک مقایسه خواهیم کرد. این مدل با یک روش غربالگری اولیه که در ادامه معرفی خواهد شد به کار گرفته می‌شود. در این حالت خاص تنها یک گروه برای خصوصیات ژنتیکی در نظر گرفته شده و توزیع آمیخته ارائه شده نیز تنها دو جزء دارد که همان توزیع پیشین تنک‌ساز میخ و تخته است.

یک نمای کلی از مدل مورد بررسی به شرح زیر است. برای فرد i فرض می‌کنیم که سن شروع یک رخداد مانند بیماری، Y_i توزیع وایبل دارد. به بیان دقیق‌تر $Y_i \sim \text{Weibull}(\alpha, b_i)$، که در آن α و b_i به ترتیب پارامترهای شکل و مقیاس هستند. بنابراین تابع بقا به صورت زیر است

$$S_i(y) = \exp\left(-\left(\frac{y}{b_i}\right)^a\right) \tag{۱.۴}$$

میانگین و واریانس توزیع وایبل وابسته هستند، زیرا هر دو پارامتر را در عبارات خود به اشتراک می‌گذارند. علاوه بر این، از آن‌جا که عبارات میانگین و واریانس حاوی توابع گاما هستند، تشریح میانگین و واریانس که فقط به یک پارامتر وابسته باشند، نسبتاً دشوار است. یک راه حل استفاده از $\log(Y_i)$ و گشتاورهای آن به جای Y_i است. اگر $Y_i \sim W(\alpha, b_i)$ آن‌گاه $\log Y_i$ دارای توزیع گامبل با میانگین و واریانس به صورت زیر است:

$$\mathrm{E}(\log Y_i | b_i, \alpha) = \log b_i - \frac{K}{\alpha}, \tag{۲.۴}$$

$$\mathrm{Var}(\log Y_i | b_i, \alpha) = \frac{\pi^۲}{۶\alpha^۲} \tag{۳.۴}$$

که در آن K ثابت اویلر ماشرونی $K \approx ۰/۵۷۷۲۱$ است.

واریانس فقط به پارامتر α بستگی دارد. از آن‌جا که هدف، مدل‌سازی اثرات SNP (β)، متغیرهای کمکی δ و میانگین مقیاس برای زمان تا رویداد μ (عرض از مبدأ) است، می‌توان آن‌ها را به صورت زیر مدل‌بندی کرد:

$$b_i = \exp\left(\mu + {x_i}^T\beta + {z_i}^T\delta + \frac{K}{\alpha}\right) \tag{۴.۴}$$

در نتیجه

$$\mathrm{E}(\log Y_i | \mu, \beta, \delta, \alpha) = \mu + {x_i}^T\beta + z_i^T\delta, \tag{۵.۴}$$

$$\mathrm{Var}(\log Y_i | \mu, \beta, \delta, \alpha) = \frac{\pi^۲}{۶\alpha^۲}, \tag{۶.۴}$$

که در آن μ عرض از مبدأ، x_i مقادیر نشانگر استاندارد شده برای همه SNP ها، β اثر نشانگرها، z_i متغیر کمکی اضافی (مانند جنسیت یا اجزای اصلی ژنتیکی)، δ اثر متغیر کمکی اضافی و α پارامتر شکل توزیع وایبل است. فرض می‌کنیم که β از توزیع آمیخته گاوسی با تعداد اجزای L توزیع شده است. برای $j \in \{۱, \ldots, M\}$ هر نشانگر β_j با یک متغیر شاخص متناظر $\gamma_j \in \{۰, \ldots, L\}$ مرتبط است. فرض می‌کنیم جزء آمیخته kام ($\gamma_j = k$) به صورت نرمال $N(۰, C_k\sigma^۲)$ توزیع شده است، $\sigma^۲$ نشان دهنده واریانس فنوتیپی قابل انتساب به نشانگرها و C_k ضریب خاص واریانس و جزء آمیخته kام است.

در این مقاله ما تنها حالت توزیع پیشین میخ پیوسته را در نظر می‌گیریم که در آن $C_۱ = ۱$ و $C_۲ = ۱ \times ۱۰^{-۱۲}$ که اولین ضریب متناظر با ضرایب غیرصفر (تأثیرگذار) و دومین ضریب متناظر با ضرایبی است که می‌خواهیم صفر براورد شوند (ژن‌های بی‌تأثیر).

تابع بقا و چگالی برای Y_i به ترتیب برابرند با

$$S_i(y) = \exp\{-y^\alpha \exp(-\alpha(\mu + x_i^T\beta + z_i^T\delta) - K)\}, \tag{۷.۴}$$

$$\begin{aligned} f_i(y) = \exp\{&-K - \alpha(\mu + x_i^T\beta + z_i^T\delta) \\ &- y^\alpha \exp(-\alpha(\mu + x_i^T\beta + z_i^T\delta) - K)\} y^{\alpha-۱}\alpha \end{aligned} \tag{۸.۴}$$

در نتیجه تابع درست‌نمایی برای داده‌های سانسور شده از راست و بریده شده از چپ (در سن a_i) برای n مشاهده به صورت زیر است (متغیرهای تصادفی با حروف کوچک به معنای مقدارهای مشاهده شده آن‌ها هستند)

$$\begin{aligned} p(D|\alpha,\beta,\delta,\mu) &= \prod_{i=1}^{n} \frac{1}{S(a_i)} \prod_{i=1}^{n} f(y_i)^{d_i} S(y_i)^{1-d_i} \\ &= \alpha^d \exp\Big\{ -Kd + (\alpha-1)\sum_{i=1}^{n} \log y_i) - \alpha \sum_{i=1}^{n} d_i(\mu + x_i^T\beta + z_i^T\delta) \\ &- \sum_{i=1}^{n} y_i^\alpha \exp(-\alpha(\mu + x_i^T\beta + z_i^T\delta) - K) \\ &+ \sum_{i=1}^{n} a_i^\alpha \exp(-\alpha(\mu + x_i^T\beta + z_i^T\delta) - K\Big\} \end{aligned} \tag{۹.۴}$$

که در آن d_i نشانگر شکست و d تعداد رویدادها در پایان دوره‌ها، a_i زمان برش چپ است. ممکن است از مدل بدون برش چپ استفاده شود. برای انجام این کار، برای هر i، $a_i = 0$ در نظر گرفته می‌شود. هرگاه $a_i = 0$ باشد، همچنین فرض می‌شود

$$\exp\big(\alpha(\log(a_i) - \mu - x_i^T\beta - z_i^T\delta)\big) = 0 \tag{۱۰.۴}$$

۱.۴ توزیع‌های پیشین

توزیع پیشین α یک توزیع گاما با پارامترهای α_0 و κ_0 به صورت زیر در نظر گرفته شده است:

$$p(\alpha) \propto \alpha^{\alpha_0 - 1} \exp(-\kappa_0 \alpha) \tag{۱۱.۴}$$

توزیع پیشین برای β_j نرمال به صورت زیر فرض شده است:

$$p(\beta_j|\sigma_\beta^2; y_j = k) \propto \left(\frac{1}{C_k\sigma_G^2}\right)^{0/5} \exp\left[-\frac{1}{2C_k\sigma_G^2}\beta_j{}^2\right] \tag{۱۲.۴}$$

توزیع پیشین برای σ_G^2 گامای معکوس با پارامترهای α_σ و β_σ به صورت زیر است:

$$p(\sigma_G^2) \propto \left(\frac{1}{\sigma_G^2}\right)^{\alpha_\sigma+1} \exp\left(-\frac{\beta_\sigma}{\sigma_G^2}\right) \tag{۱۳.۴}$$

توزیع پیشین برای δ_q (qامین متغیر کمکی) نرمال به صورت زیر

$$p(\delta_q) \propto \exp\left(-\frac{1}{2\sigma_\delta^2}\delta_q^2\right) \tag{۱۴.۴}$$

توزیع پیشین برای μ نرمال به صورت زیر:

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_\mu^2}\mu^2\right) \tag{۱۵.۴}$$

و توزیع پیشین برای γ_j چند نقطه‌ای به صورت زیر است:

$$p(\gamma_j|\pi) = \pi_0^{I(\gamma_j=0)} \cdot \ldots \cdot \pi_L^{I(\gamma_j=L)} \tag{۱۶.۴}$$

توزیع ابرپیشین متعلق به هر یک از توزیع‌های آمیخته دیریکله به صورت زیر هستند:

$$p(\pi|\gamma) = \text{Dirichlet}(P_L) \tag{۱۷.۴}$$

که در آن P_L بردار $L+1$ بعدی با مقادیر معین است.

۲.۴ توزیع‌های پسین شرطی

به آسانی مشاهده می‌شود که توزیع پسین شرطی برای σ_G^2 توزیع گامای معکوس به ترتیب به صورت زیر است:

$$\alpha_\sigma + ۰/۵ \sum_{k=۱}^{L} |\gamma^k| \tag{۱۸.۴}$$

و

$$۰/۵ \sum_{k=۱}^{L} |\gamma^k| \beta_{\gamma^k}^T \beta_{\gamma^k} + \beta_\sigma, \tag{۱۹.۴}$$

که γ^k مجموعه‌ای از شاخص‌های j را نشان می‌دهد که $\gamma_j = k$ است. توزیع پسین شرطی برای π توزیع دیریکله به صورت زیر است:

$$p(\pi|\gamma) = \text{Dirichlet}(P_L + (|\gamma^۰|, \ldots, |\gamma^L|)) \tag{۲۰.۴}$$

هیچ شکل ساده‌ای برای توزیع‌های پسین شرطی برای α، μ، β_j و δ_q وجود ندارد و توزیع‌های پسین شرطی آن‌ها لگاریتم مقعر[۲۴] هستند. بنابراین نمونه‌برداری برای α، μ، β_j و δ_q می‌تواند با استفاده از نمونه‌گیری متروپولیس هستینگز که فقط به لگاریتم پسین نیاز دارد، انجام شود.

۵ مطالعه شبیه‌سازی

در این بخش یک مطالعه شبیه‌سازی برای مقایسه عملکرد دو مدل رگرسیون کاکس لاسو و مدل بیزی کاکس اجرا شده است. حلقه شبیه‌سازی در اینجا ۱۰۰ بار اجرا می‌شود و به تکرارهای متعدد اجازه داده می‌شود تا اطمینان حاصل شود که نتایج تا حد امکان دقیق هستند.

میانگین توان دوم خطای پیش‌بینی (MSPE) معیاری است که میزان خطاهای مربعی بین مقادیر پیش‌بینی‌شده و مقادیر واقعی را برای داده‌های آزمون اندازه‌گیری می‌کند. این معیار به صورت زیر تعریف می‌شود:

$$\text{MSPE} = \frac{۱}{M} \sum_{j=۱}^{M} \frac{۱}{n} \sum_{i=۱}^{n} (y_i - \hat{y}_{ij})^۲$$

که در آن

- y_i مقدار واقعی متغیر پاسخ برای نمونه iام است.
- $\hat{y}_{ij}$ مقدار پیش‌بینی‌شده متغیر پاسخ برای نمونه iام در تکرار jام است.
- n تعداد نمونه‌ها است.

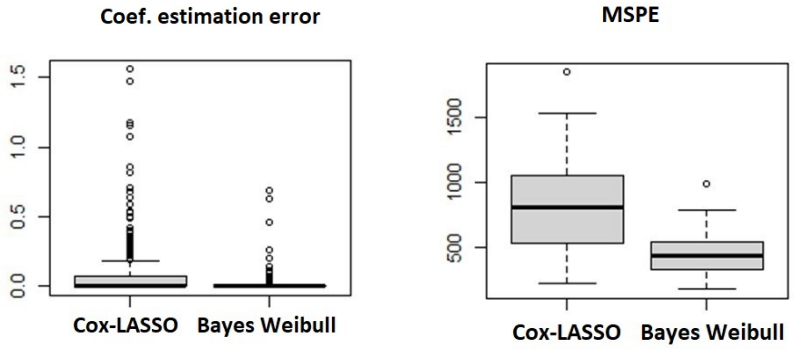
تفاوت MSPE با MSE (میانگین توان دوم خطا) این است که معیار MSPE برای پیش‌بینی نمونه‌های آینده و یا خارج از نمونه آموزشی و معیار MSE برای براورد پارامترها استفاده می‌شوند.

خطای شبیه‌سازی براورد ضریب رگرسیونی نشان‌دهنده تفاوت بین ضریب رگرسیون واقعی و ضریب رگرسیون براورد شده توسط مدل است. این خطا به صورت زیر تعریف می‌شود:

$$\text{خطای براورد ضریب رگرسیونی} = \frac{۱}{M} \sum_{j=۱}^{M} (\hat{\beta}_j - \beta_j)^۲$$

که در آن

[24] log-concave



شکل ۱: نمودارهای جعبه‌ای نتایج شبیه‌سازی

- β_j ضریب رگرسیون واقعی در تکرار j است.
- $\hat{\beta}_j$ ضریب رگرسیون براورد شده در تکرار j است.
- M تعداد تکرارهای شبیه‌سازی است.

این بردارها توان دوم خطاهای ضرایب براورد شده را ذخیره می‌کنند و این به ارزیابی نزدیک بودن ضرایب براورد شده به ضرایب واقعی مورد استفاده برای تولید داده‌ها کمک می‌کند.

با مقداردهی اولیه این بردارها ساختاری برای جمع‌آوری و ذخیره نتایج از هر تکرار شبیه‌سازی تنظیم می‌شود که امکان مقایسه جامع عملکرد دو مدل را در چندین معیار فراهم می‌کند.

۱۰ عدد تصادفی از یک توزیع نرمال با میانگین ۰ و انحراف معیار ۱/۰ تولید می‌شود که این اعداد نشان دهنده ضرایب واقعی برای داده‌های شبیه‌سازی شده اند. همچنین ۱۰ مقدار باینری از یک توزیع دوجمله‌ای با احتمال ۳/۰ هم تولید می‌شود که این مقادیر نیز به عنوان شاخص برای تعیین ضرایب فعال (غیرصفر) عمل می‌کنند. این مقدارها باعث می‌شوند نرخ سیگنال نسبت به نرخ خطا مقدار مناسبی باشد به طوری که مدل آماری قادر به تشخیص سیگنال باشد. ضرایب بر اساس شاخص‌های باینری تنظیم می‌شوند و برخی مقادیر را صفر قرار می‌دهند. این فرآیند تضمین می‌کند که هر تکرار حلقه مجموعه جدیدی از ضرایب واقعی را با برخی از ضرایب صفر تولید می‌کند و سناریوهای مختلفی را برای ارزیابی مدل‌ها شبیه‌سازی می‌کند. ۸۰ درصد از داده‌ها به داده‌های آموزشی و ۲۰ درصد به داده‌های آزمایشی اختصاص داده می‌شود .

در اینجا ۱۰متغیر کمکی، زمان بقا تا ۵۰ و نرخ سانسور ۱۰ درصد در نظر گرفته شده است و با استفاده از تابع sim.survdata مجموعه داده‌ها، با ایجاد زمان بقا بر اساس ضرایب و متغیرهای کمکی مشخص شده و سپس اعمال نرخ سانسور برای برخی از مشاهدات، تولید می‌شوند.

در مرحله بعدی هر دو مدل بر روی مجموعه داده آموزشی، آموزش داده می‌شوند تا رابطه بین متغیرهای کمکی و زمان بقا را بیاموزند. سپس از مدل‌های آموزش دیده برای پیش‌بینی روی مجموعه داده آزمایشی برای ارزیابی عملکرد آن‌ها استفاده می‌شود و بدین ترتیب مقایسه عملکرد این دو مدل از نظر دقت پیش‌بینی و براورد ضریب انجام می‌شود که در بخش بعدی به نتایج به دست آمده که با استفاده از نمودارهای جعبه‌ای تصویرسازی شده است، پرداخته می‌شود.

نمودار شکل ۱ سمت چپ نمودار جعبه‌ای میانگین توان دوم خطای پیش‌بینی MSPE را برای دو روش و تمامی تکرارهای شبیه‌سازی نشان می‌دهد. همان‌طور که از نمودار می‌توان دید، روش وایبل بیزی به طرز معنی‌داری دارای خطاهای کمتری در پیش‌بینی متغیر پاسخ که همان زمان تا رخداد است، برای دادگان آزمون است.

نمودار شکل ۱ سمت راست نمودار جعبه‌ای خطای براورد ضرایب رگرسیونی را برای دو روش و تمامی تکرارهای شبیه‌سازی نشان می‌دهد. همانطور که از نمودار می‌توان دید، روش وایبل بیزی به طرز معنی‌داری دارای خطاهای کمتری در براورد دقیق ضریب رگرسیونی است.

۶ تحلیل دادگان وزن موش‌ها

در این فصل علاوه بر معیار خطای پیش‌بینی که در شبیه‌سازی استفاده شد، معیار دیگری به نام C–index یا شاخص هماهنگی هارل[۲۵] نیز برای ارزیابی مدل استفاده می‌شود.

شاخص هماهنگی هارل (C-index) یک معیار ارزیابی برای مدل‌های رگرسیون در داده‌های بقا است. این شاخص، میزان تطابق پیش‌بینی‌های مدل با داده‌های مشاهده شده را اندازه‌گیری می‌کند. به عبارت دیگر، C-index به ما می‌گوید که مدل چقدر می‌تواند به درستی ترتیب زمان‌های رویداد (مانند زمان مرگ یا شکست) را پیش‌بینی کند.

C-index به صورت زیر تعریف می‌شود:

$$C = \frac{\sum_{i,j} \mathbb{I}(u_i > u_j) \cdot \mathbb{I}(\hat{u}_i > \hat{u}_j) + ۰٫۵ \times \sum_{i,j} \mathbb{I}(\hat{u}_i = \hat{u}_j)}{\sum_{i,j} \mathbb{I}(u_i > u_j)}, \tag{۱.۶}$$

که در آن

- u_i و u_j زمان‌های واقعی وقوع رویداد برای افراد i و j هستند.
- $\hat{u}_i$ و $\hat{u}_j$ زمان‌های پیش‌بینی شده برای همان افراد هستند.
- $\mathbb{I}(.)$ تابع نشان‌گر است

به زبان ساده، C-index درصد جفت‌هایی است که مدل به درستی ترتیب وقوع رویداد آن‌ها را پیش‌بینی کرده است. مقدار این شاخص بین ۰ و ۱ است؛ مقدار ۱ نشان‌دهنده یک مدل کاملاً دقیق و مقدار ۰.۵ نشان‌دهنده یک مدل تصادفی است.

۱.۶ معرفی دادگان

در این مطالعه از بزرگترین موش وحشی در جهان که در جزیره گاف سکونت دارد استفاده شده است. موش‌ها در جزیره گاف جمع‌آوری شده و به آزمایشگاه برده می‌شوند. پنجاه موش در سپتامبر ۲۰۰۹ در جزیره گاف نزدیک ایستگاه تحقیقاتی به دام افتادند. چهار موش ازبین رفتند و از دو موش باردار (پنج موش) در حین انتقال از جزیره گوف به مرکز آموزشی چارمانی در دانشگاه ویسکانسین مدیسون به دنیا آمدند. ۲۵ موش ماده بالغ و ۲۱ موش نر بالغ برای شروع یک کلونی پرورشی استفاده شدند. همه موش‌ها در قفس‌های میکرو ایزوله قرار گرفتند و بر اساس جنسیت از هم جدا شدند. موش‌های نر فقط در صورتی گروه‌بندی می‌شدند که از یک زایمان باشند و قبلاً جفت نگرفته باشند. موش‌های انتخاب شده پس از ۸ هفتگی شروع به جفت‌گیری می‌کردند. وزن موش‌ها از ۱ هفتگی تا هفته ۱۶، با نزدیک‌ترین دقت میلی‌گرم ثبت شد. برای شناسایی موش‌ها با شروع هفته اول، خالکوبی انگشت پا با استفاده از لانست‌های استریل و خمیر تاتو انجام شد. برای تشخیص موش‌های بالغ، زمانی که آن‌ها ۳–۴ هفته بودند و از شیر گرفته می‌شدند، علامتی روی گوش گذاشته می‌شد.

با شروع جفت‌گیری از موش‌های وحشی در آزمایشگاه، چهار نسل از این موش‌ها متولد شدند. سپس با جفت‌اندازی نسل چهارم با موش WSB که موش آزمایشگاه جکسون بوده، در انتها چهار نسل ۲F تولید شد. موش WSB به این دلیل انتخاب شده است که یک سویه از حیوانات وحشی است، دارای اندازه بدن معمول برای موش‌های وحشی است، متعلق به همان زیرگونه موش‌های جزیره گاف است، کاملاً دارای نژاد است و دارای ژنوم توالی‌یابی شده است. در نهایت ۱۳۷۴ موش نسل ۲F شامل ۴۹۷ موش از خط نسلی A و ۸۷۷ موش از خط نسلی B تولید گردید. اطلاعات فنوتیپی موش‌ها که شامل وزن آن‌ها بود با دقت میلی‌گرم و با شروع از هفته‌ی اول بعد ار تولد، در طی ۱۶ هفته جمع‌آوری گردید. همچنین تمامی موش‌ها با استفاده از آرایه جهانی ژنوتیپ Mouse Mega ژنوتیپ شدند. با ارزیابی ژنوتیپ‌ها خطاهای بیولوژیک، فنی و خطاهایی که در هنگام ثبت داده به وجود آمده‌اند، شناسایی شدند. همچنین با پاکسازی نشان‌گرهای ژنتیکی‌ای که مقادیر گم شده بالایی داشتند و یا اطلاعات خاصی را در اختیار نمی‌گذاشتند، مجموعه‌ای از ۱۱۸۳۳ نشانگر ژنتیکی حاصل گشت. همچنین با حذف نمونه موش‌هایی که در ۴ نقطه زمانی و یا بیشتر، وزن آن‌ها مشخص نبود، مجموعه‌ای متشکل از ۱۲۱۲ موش نسل ۲F ساخته شد.

۲.۶ پیش پردازش داده‌ها

در پایگاه داده مورد استفاده از موش‌ها ژنوتیپ‌ها به صورت GW GG، و WW آورده شده بودند و برای استفاده از آن‌ها در ورودی الگوریتم باید مورد پیش‌پردازش قرار می‌گرفتند و به شکل عددی و استاندارد آورده می‌شدند، که این کار انجام شد و ژنوتیپ‌ها به صورت استاندارد شده

[25] Harrell's concordance index

درآمدند. برای استاندارد کردن ژنوتیپ‌ها =۲GG، =۱GW و =۰WW در نظر گرفته شد و با استفاده از فرمول z-score مقدار ژنوتیپ استاندارد شده به روش زیر محاسبه گردید

$$z_i = \frac{x_i - \mu}{\sigma}$$

که در آن پارامترها به صورت زیر است:

Z_i: مقدار ژنوتیپ استاندارد شده

x_i: مقدار عددی مربوط به هر ژنوتیپ (برای مثال =۲GG)

μ: میانگین کل ژنوتیپ‌ها

σ: انحراف معیار کل ژنوتیپ‌ها

قدم بعدی در مرحله‌ی پیش‌پردازش رسیدگی به گمشدگی‌ها[۲۶] در وزن موش‌ها بود، در برخی از هفته‌ها وزن تعدادی از موش‌ها گمشدگی داشتند که با استفاده از روش جانهی میانگین، مقادیر گمشده جانهی شده‌اند.

مسئله دیگری که به آن پرداخته شد، بحث تعریف چاقی در موش‌ها بود. از آن‌جایی که این پژوهش در حوزه زمان تا رخداد انجام شده است و در اینجا رخداد چاقی موش‌ها است، باید تعریف دقیقی از چاقی در موش‌ها مشخص می‌شد.

از این رو چارک سوم وزن کل موش‌ها که برابر با ۳/۲۰ به دست آمده بود به عنوان معیار چاقی موش‌ها در نظرگرفته شد. بدین ترتیب موش‌هایی که وزن ۳/۲۰ و بالاتر را تجربه کرده بودند درواقع رخداد چاقی برایشان روی داده بود و زمان رسیدن به چاقی (رخداد) برایشان به دست آمد. برای مثال اگر وزن موشی در هفته ششم برای اولین بار عددی برابر یا بزرگتر از ۳/۲۰ بود، رخداد آن برابر با ۱ و زمان آن برابر با ۶ قرار گرفت و بدین صورت وضعیت زمان‌ها و رخدادها برای ۱۲۱۲ موش مشخص گردید.

۳.۶ غربالگری اولیه

با توجه به بعد بالای داده‌ها (تعداد زیاد ژن‌ها) و با توجه به حجم بالای محاسبات در روش بیزی که از روش نمونه‌گیری زنجیره مارکوف مونت‌کارلو استفاده می‌کند، برای مقایسه‌پذیری دو روش ارائه شده، ابتدا تعداد زیادی از ژن‌ها را در یک غربالگری اولیه حذف کرده‌ایم. برای این کار ابتدا ژن‌هایی را که دارای همبستگی بسیار نزدیک به صفر با متغیر پاسخ (زمان تا رخداد چاقی) نسبت به سایر ژن‌ها بودند حذف کردیم. همچنین از میان سایر ژن‌ها آن‌هایی که دارای واریانس بسیار نزدیک به صفر بودند را نیز حذف کردیم. این کار به این دلیل است که اگر ژنی دارای واریانس بسیار پایین باشد نمی‌توان آن را به عنوان یک ویژگی بارز و یک عامل انتخاب نمود. این کار در روش‌هایی مانند روش مؤلفه‌های اصلی[۲۷] در انتخاب ویژگی متداول است و همچنین در ژنتیک نیز به عنوان یک روش غربالگری اولیه به وفور استفاده می‌شود. این غربالگری را در چهار فاز انجام داده‌ایم و نتایج مدل رگرسیون کاکس لاسو را بررسی کرده‌ایم تا ببینیم غربالگری چه تأثیری در دقت مدل دارد.

به بیان دقیق‌تر در هر فاز چارک اول واریانس و چارک اول همبستگی را به دست آورده و تنها ژن‌های با واریانس و همبستگی بالاتر از این دو چارک را انتخاب کرده‌ایم. نمودارهای شکل ۲ نتایج این چهار فاز را نشان می‌دهند. همچنین نمودار ۳ نمودار اعتبارسنجی متقابل برای به دست آوردن پارامتر تنظیم بهینه در روش رگرسیون کاکس لاسو را در فاز نهایی نشان می‌دهند.

همانطور که از نمودار ۲ می‌توان دید، در فاز اول شاخص c-index و میانگین توان دوم خطای پیش‌بینی[۲۸] برای داده‌های آزمون مقداری کاهش می‌یابند. توجه کنید که این کاهش برای c-index خبر بد و برای خطا خبر خوبی است. بنابراین در فاز اول نمی‌توان به طور قاطع گفت که غربالگری تأثیر مثبت و یا منفی گذاشته است. با این حال با توجه به این که معمولاً خطای پیش‌بینی در عمل اهمیت بالاتری دارد می‌توان نتیجه را متمایل به مثبت ارزیابی کرد. همچنین در فازهای بعدی تغییر چندان محسوسی در این دو معیار در نمودارها دیده نمی‌شود. در واقع شاخص c-index افزایش بسیار کم و MSPE هم افزایش و سپس کاهش بسیار نرمی دارد که در نهایت هر دو خبر خوبی هستند. با این حال نمودار ۲ ارزش افزوده بسیار بالای بدست آمده از روش غربالگری را نشان می‌دهد که همان کاهش بعد و تعداد کل ژن‌ها از حدود ۱۲۰۰۰ ژن به کمتر از ۲۰۰۰ ژن است که بسیار خبر خوبی به ویژه برای روش بیزی است.

۴.۶ پیاده‌سازی روش رگرسیون کاکس لاسو

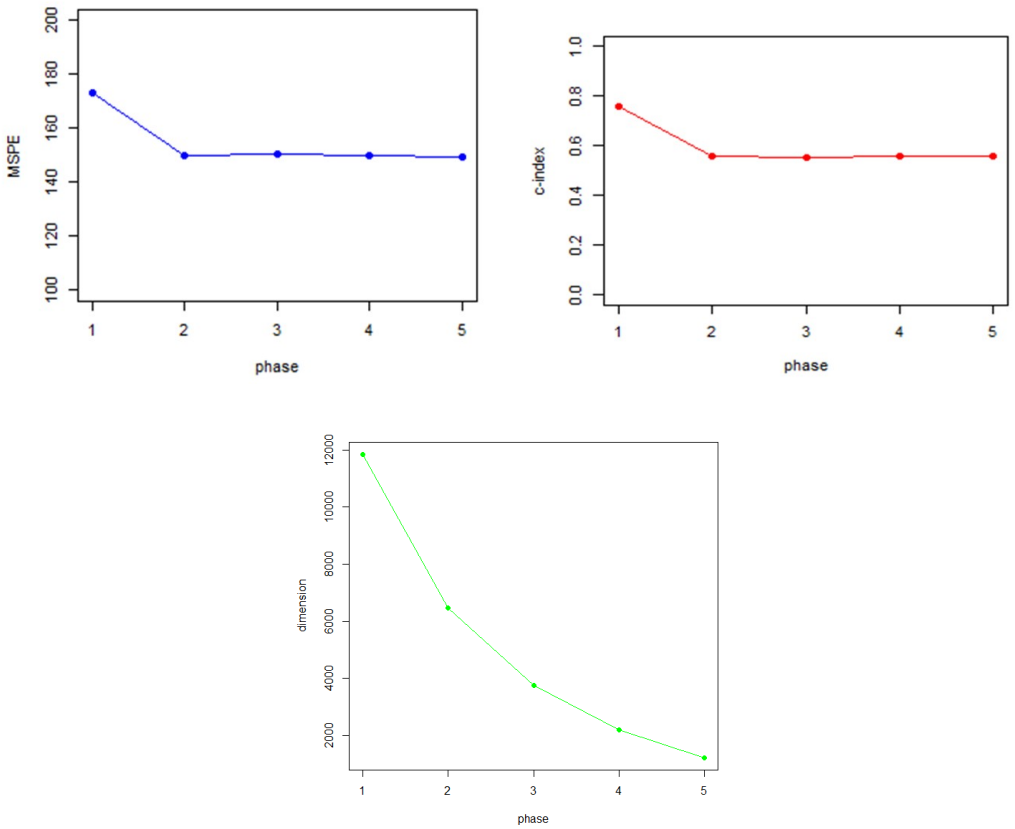
برای اجرای این الگوریتم، دو ورودی مورد نیاز است. ورودی اول یا X در واقع متغیرهای کمکی می‌باشد که در این‌جا همان ژنوتیپ‌های استاندارد شده‌اند و ورودی دوم یا Y نیز شامل ستون زمان رخداد و وضعیت رخداد می‌باشد که پیش‌تر در رابطه با نحوه به دست آوردن آن‌ها توضیح داده شده است.

داده‌ها را با نسبت به ترتیب ۸۰ درصد و ۲۰ درصد به دو مجموعه داده آموزش و آزمون تقسیم کرده‌ایم. این مجموعه‌ها در هر دو روش رگرسیون کاکس لاسو و رگرسیون وایبل بیزی یکسان هستند تا بتوان دو روش را بر اساس دقت پیش‌بینی با هم مقایسه کرد.

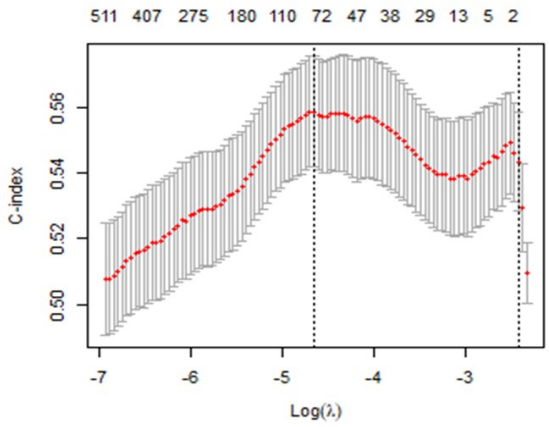
[26] missing

[27] Principal Component Analysis (PCA)

[28] Mean Squared Prediction Error (MSPE)



شکل ۲: نمودار c–index (بالا چپ) MSPE (بالا راست) و تعداد ژن‌های باقی‌مانده (پایین) برای کل ژن‌ها و فازهای یک تا چهار



شکل ۳: نمودار اعتبارسنجی متقابل برای انتخاب پارامتر تنظیم بهینه λ مربوط به فاز ۴

برای تعیین مقدار بهینه پارامتر تنظیم روش کاکس لاسو از اعتبارسنجی متقابل استفاده شده است. شاخص پیش‌فرض در اعتبارسنجی متقابل در بسته نرم‌افزاری glmnet شاخص هماهنگی هارل است. بر اساس مقدار بهینه پارامتر تنظیم، مدل نهایی کاکس لاسو برازش داده شده است.

ضرایب به دست آمده نهایی که به شیوه تنک براورد شده‌اند، علاوه بر استفاده در پیش‌بینی متغیر پاسخ که همان زمان تا رخداد چاقی موش‌ها است، در تعیین ژن‌های تأثیرگذار در این رخداد استفاده می‌شوند. ضرایبی که غیرصفر براورد شده‌اند متناظر با ژن‌های تأثیرگذار و ضرایب صفر متناظر با ژن‌های بی‌تأثیر هستند.

در اجرای روش رگرسیون کاکس لاسو $\mathrm{MSPE} = ۱۴۹$ است که این مقدار زیاد است و نشان می‌دهد که مدل نسبت به روش رقیب که در بخش بعد بیان می‌شود، عملکرد خوبی نداشته است.

۵.۶ پیاده‌سازی روش رگرسیون وایبل بیزی

برای انجام نمونه‌گیری زنجیره مارکوف مونت‌کارلو از مدل بیزی وایبل معرفی شده از کتابخانه rjags در نرم‌افزار R استفاده شده است.

براساس تابع درست‌نمایی رابطه (۹.۴)، و توزیع‌های پیشین که در فصل ۳ از فرمول (۱۱.۴) تا فرمول (۱۷.۴) بیان شده‌اند، الگوریتم متروپولیس هستینگر و نمونه‌گیری رد تطبیقی انجام می‌شود.

ابرپارامترها به صورت زیر مقداردهی شده‌اند:

$$\alpha_\circ = \circ/\circ\circ\circ\,۱ ،\ \kappa_\circ = \circ/\circ\circ\circ\,۱ ،\ \alpha_\sigma = \circ/\circ\circ\circ\,۱ ،\ \beta_\sigma = \circ/\circ\circ\circ\,۱ ،\ \sigma_\delta^۲ = ۱\circ\circ\circ\circ\circ ،\ \sigma_\mu^۲ = ۱\circ\circ\circ\circ\circ ،$$
$$\alpha_{\pi,۱} = ۱/\circ\circ ۱ ،\ \alpha_{\pi,۲} = ۱/\circ\circ ۱$$

نمونه‌گیری زنجیره مارکوف مونت کارلو با تعداد تکرار[29] برابر ۴۰۰۰۰ و دوره داغیدن[30] برابر ۱۰۰۰۰ برای نمونه‌گیری از توزیع‌های پسین انجام شده است. همچنین برای جلوگیری از خودهمبستگی بالای زنجیره مارکوف، نمونه‌ها با تاخیرهای ۳۰ تایی برداشته می‌شوند.

براورد پارامترها بر اساس میانگین نمونه‌های نهایی به دست می‌آیند. همچنین برازش متغیر پاسخ بر اساس نمونه‌گیری از توزیع پیش‌بین پسین انجام شده است. بر اساس این مقدارهای برازش یافته پیش‌بینی شده، میانگین توان دوم خطای پیش‌بینی داده‌های آزمون در روش رگرسیون وایبل بیزی مقدار ۷۴/۲۳ به دست آمده است که مقدار بسیار کمتری نسبت به معیار متناظر در الگوریتم کاکس لاسو است. بنابراین به وضوح روش وایبل بیزی در پیش‌بینی زمان تا رخداد چاقی بر اساس ژن‌ها بسیار موفق‌تر عمل کرده است.

۷ جمع‌بندی نتیجه‌گیری و پیشنهادات

در این مقاله، رابطه بین پلی‌مورفیسم‌های تک‌نوکلئوتیدی و زمان تا رخداد چاقی گونه‌ای از موش‌های وحشی، به دو شیوه فراوانی‌گرا و بیزی مورد بررسی قرار گرفته است. دو روش کاکس لاسو و رگرسیون وایبل بیزی که علاوه بر مدل‌بندی رگرسیونی انتخاب متغیر را نیز به روش مدل‌بندی تنک انجام می‌دهند بدین منظور به کار گرفته شده‌اند. با توجه به تعداد بالای ژن‌های انتخاب شده، یعنی ژن‌هایی که ضریب رگرسیونی آن‌ها غیرصفر براورد شده است، امکان لیست کردن آن‌ها در این مقاله وجود نداشت، با این حال می‌توان به راحتی در خروجی مدل‌ها به این ژن‌ها دست یافت. در ادامه طبق نتایج و نمودارهای به دست آمده از دو الگوریتم رگرسیون کاکس لاسو و رگرسیون وایبل بیزی، مشخص شد که الگوریتم رگرسیون وایبل بیزی با میانگین توان دوم خطای ۷۴.۲۳ عملکرد خیلی بهتری نسبت به الگوریتم رگرسیون کاکس لاسو با میانگین توان دوم خطای ۱۴۹ دارد. با توجه به بعد بسیار بالای مسئله برای ممکن ساختن اجرای الگوریتم‌ها به‌خصوص الگوریتم بیزی که با توجه به نمونه‌گیری زنجیره مارکوف مونت‌کارلو زمان و حجم محاسبات بسیار بالایی دارد، ابتدا با یک روش غربالگری اولیه تعداد زیادی از متغیرهای پیش‌گو را که دارای همبستگی بسیار نزدیک به صفر با متغیر پاسخ (نسبت به سایر متغیرهای پیش‌گو) و یا واریانس (به دلیل این‌که اگر ژنی دارای واریانس بسیار کم باشد بدین معنی است که این ژن در افراد مختلف تقریباً یکسان بوده و نمی‌تواند به عنوان یک عامل خاص شناخته شود) بسیار نزدیک به صفر (نسبت به بقیه) بودند حذف کردیم و سپس دو الگوریتم را بر روی باقی‌مانده متغیرهای پیش‌گو که باز هم تعداد قابل توجهی بودند اجرا کردیم.

اگرچه الگوریتم بیزی نسبت به الگوریتم فراوانی‌گرا زمان بسیار زیادتری برای محاسبات صرف می‌کند، با این حال با توجه به خطای بسیار پایین‌تر به نظر می‌رسد که صرف این زمان ارزش داشته است. علاوه بر این با توجه به ظهور روش‌های بسیار سریع‌تر در استنباط بیزی، مانند روش‌های بیز گوناگونی[31] که در این مقاله به آن‌ها پرداخته نشده است، می‌توان امیدوار بود که با توسعه نتایج ارائه شده در این مقاله بتوان به روش‌هایی با هم دقت و هم سرعت بالاتر در استنباط بیزی این مدل‌ها دست یافت. بنابراین اولین پیشنهاد در خصوص این پژوهش می‌تواند توسعه مدل رگرسیون وایبل بیزی با استفاده از روش‌های بیز گوناگونی باشد.

در انتها می‌توان از تکنیک‌های پیشرفته یادگیری ماشین نیز برای براورد زمان تا رخداد در این مسائل استفاده کرد.

کدهای مورد استفاده در این مقاله در آدرس زیر موجود است

https://github.com/mortamini/BayesWeibullReg/

مراجع

[1] Bi, W., Fritsche, L.G., Mukherjee, B., Kim, S. and Lee, S., 2020. A fast and accurate method for genome-wide time-to-event data analysis and its application to UK Biobank. *The American Journal of Human Genetics*, 107(2), pp.222–233. doi:10.1016/j.ajhg.2020.06.003

[29] iterations
[30] burn-in period
[31] Variational Bayes

[2] Cox, D.R., 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), pp.187–220. doi:10.1111/j.2517-6161.1972.tb00899.x

[3] Dey, T., Lipsitz, S.R., Cooper, Z., Trinh, Q.D., Krzywinski, M. and Altman, N., 2022. Regression modeling of time-to-event data with censoring. *Nature Methods*, 19(8), pp.888–890. doi:10.1038/s41592-022-01689-8

[4] Elston, R.C., Satagopan, J.M. and Sun, S., 2012. Genetic terminology. *Methods in Molecular Biology*, 850, pp.1–9. doi:10.1007/978-1-61779-555-8_1

[5] Flint, J. and Eskin, E., 2012. Genome-wide association studies in mice. *Nature Reviews Genetics*, 13(11), pp.807–817. doi:10.1038/nrg3335

[6] Ganesalingam, J. and Bowser, R., 2010. The application of biomarkers in clinical trials for motor neuron disease. *Biomarkers in Medicine*, 4(2), pp.281–297. doi:10.2217/bmm.09.71

[7] Gentleman, R., Hornik, K. and Parmigiani, G., 2006. Use R! Springer, New York, NY.

[8] Hua, K., Wojdyla, D., Carnicelli, A., Granger, C., Wang, X. and Hong, H., 2025. Network meta-analysis with individual participant-level data of time-to-event outcomes using Cox regression. *Statistics in Medicine*, 44(5), p.e70027. doi:10.1002/sim.70027

[9] Joshi, P.K., Fischer, K., Schraut, K.E., Campbell, H., Esko, T. and Wilson, J.F., 2016. Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nature Communications*, 7, p.11174. doi:10.1038/ncomms11174

[10] Joshi, P.K., Pirastu, N., Pietzner, M., Jansen, R., Robinson, M.R., Winkler, T.W., Sala, C.F. and Wilson, J.F., 2017. Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity. *Nature Communications*, 8, p.910. doi:10.1038/s41467-017-00934-5

[11] Laber, S. and Cox, R.D., 2017. Mouse models of human GWAS hits for obesity and diabetes in the post-genomic era: time for reevaluation. *Frontiers in Endocrinology*, 8, p.11. doi:10.3389/fendo.2017.00011

[12] Maeshima, K., Ide, S. and Babokhov, M., 2019. Dynamic chromatin organization without the 30-nm fiber. *Current Opinion in Cell Biology*, 58, pp.95–104. doi:10.1016/j.ceb.2019.02.003

[13] Mondal, T., Rasmussen, S.H., Pandey, G.K., Isaksson, A. and Kanduri, C., 2010. Characterization of the RNA content of chromatin. *Genome Research*, 20(7), pp.899–907. doi:10.1101/gr.103473.109

[14] Nagpal, C., Jeanselme, V. and Dubrawski, A., 2021. Deep parametric time-to-event regression with time-varying covariates. *Survival Prediction – Algorithms, Challenges and Applications*, pp.184–193.

[15] Ojavee, S.E., Kousathanas, A., Trejo Banos, D., Orliac, E.J., Patxot, M., Läll, K., Mägi, R., Fischer, K., Kutalik, Z. and Robinson, M.R., 2021. Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis. *Nature Communications*, 12(1), p.2337. doi:10.1038/s41467-021-22538-w

[16] Rizvi, A.A., Karaesmen, E., Morgan, M., Preus, L., Wang, J., Sovic, M., Hahn, T. and Sucheston-Campbell, L., 2018. gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics*, 35(11), pp.1968–1970. doi:10.1093/bioinformatics/bty920

[17] Sillanpää, E. and Shahzad, K., 2015. Supplier development and buyer-supplier relationship strategies – a literature review. *International Journal of Procurement Management*, 8(1-2), pp.227–250. doi:10.1504/IJPM.2015.066283

[18] Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J. et al., 2017. A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics*, 25(7), pp.854–862. doi:10.1038/ejhg.2017.78

[19] Syed, H., Jørgensen, A.L. and Morris, A.P., 2016. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics*, 17(8), pp.907–915. doi:10.2217/pgs.16.19

[20] Syed, H., Jørgensen, A.L. and Morris, A.P., 2017. Survival GWAS-SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinformatics*, 18(1), p.265. doi:10.1186/s12859-017-1683-z

[21] Syed, H., Jørgensen, A.L. and Morris, A.P., 2016. SurvivalGWAS_Power: a user friendly tool for power calculations in pharmacogenetic studies with "time to event" outcomes. *BMC Bioinformatics*, 17(1), p.523. doi:10.1186/s12859-016-1407-9

[22] Uffelmann, E., Huang, Q.Q., Munung, N.S., de Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D., 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1, p.59. doi:10.1038/s43586-021-00056-9

[23] Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. and Yang, J., 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1), pp.5–22. doi:10.1016/j.ajhg.2017.06.005

[24] https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies

[25] www.rasekhoon.net

[26] www.colifelabs.com/blog/what-is-dna

[27] www.genome.gov/about-genomics

[28] www.researchgate.net

[29] www.jalalisah.iut.ac.ir

[30] www.sciencedirect.com

[31] https://gentis.com.vn/public/media/dich-vu/adn/snp.jpg

Genetic Modeling of Time to Obesity Event of Mice Using Frequentist and Bayesian Regression Models

Hadiss Rasekh Fotouhi and Morteza Amini [32]

Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran.

Communicated by: Alireza Fakharzadeh Jahromi

Received: 20 October 2024 Accepted: 24 January 2026

Abstract: This paper examines the statistical relationship between the single-nucleotide polymorphisms of various mice and their weight, collected over sixteen weeks, using regression models. The primary concern of this paper is the time-to-event analysis, which is of interest to researchers in survival analysis. Indeed, the time-to-event (fatness) is modeled as a function of the genes as the covariates in this paper. Considering the high dimension of the problem and the high number of ineffective genes on the fatness event, variable selection and sparse modeling are also studied. The frequentist Cox-LASSO regression and the Bayesian Weibull regression models are introduced and compared in this paper. The simulation study and real data analysis show that the Bayesian model performs better than the frequentist method.

Keywords: Genome-wide association studies, time to event, single nucleotide polymorphisms, Bayes Weibull regression method, cox-LASSO regression method.

©2024 Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-NonComertial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

32

Corresponding Author *E-mail addresses:* morteza.amini@ut.ac.ir (M. Amini), h.fotoohi@ut.ac.ir (H.R. Fotouhi)