دانشکده علوم ریاضی
گروه آمار

پایان‌نامه

برای دریافت درجه کارشناسی ارشد در رشته
آمار، گرایش آمار ریاضی

عنوان

مدل‌بندی داده‌های نرخ و نسبت با رگرسیون بتا

استاد راهنما

دکتر حسین باغیشنی

استاد مشاور

دکتر محمد آرشی

پژوهشگر

میعاد ولیپور پاشاکلایی

دی ۱۳۹۲

نام خانوادگی دانشجو: ولیپور پاشاکلایی نام: میعاد

عنوان: مدل‌بندی داده‌های نرخ و نسبت با رگرسیون بتا

استاد راهنما: دکتر حسین باغیشنی
استاد مشاور: دکتر محمد آرشی

مقطع تحصیلی: کارشناسی ارشد رشته: آمار گرایش: آمار ریاضی

دانشگاه: دانشگاه صنعتی شاهرود دانشکده علوم ریاضی
تاریخ فارغ‌التحصیلی: دی ۱۳۹۲ تعداد صفحات: ۱۰۱

واژگان کلیدی: تابع امتیاز، تابع پیوند، توزیع مجانبی، رگرسیون بتا، ماتریس اطلاع فیشر، مدل‌های خطی تعمیم‌یافته.

چکیده

در بسیاری از کاربردها، محققین علاقه‌مند به کشف رابطه بین یک متغیر با متغیرهای دیگر می‌باشند. برای این منظور، از مدل رگرسیونی که یکی از پرطرفدارترین مدل‌های آماری است، استفاده می‌شود. در بسیاری از کاربردها، ماهیت متغیر پاسخ به صورت نرخ و نسبت می‌باشد. به عنوان چند مثال: در علوم اقتصادی، اقتصاددانان به دنبال درک رابطه بین نرخ رشد، نرخ بیکاری، درصد تولید ملی و غیره با چند متغیر اقتصادی دیگر هستند؛ معمولا برای مدل‌بندی داده‌هایی با دامنه تغییرات $(0,1)$، از مدل‌های لجستیک و پرابیت استفاده می‌شود. اما از آن‌جا که داده‌های نسبت یا نرخ، معمولا در یک زیرفاصله مشخصی از دامنه تغییرات خود متمرکز هستند (به عبارتی توزیع این نوع داده‌ها به شدت چوله است)، مدل‌های لجستیک و پرابیت برای مدل‌بندی آن‌ها مناسب نیستند. به عنوان مثال، به طور معمول نرخ بیکاری بیشتر از ۳۰ درصد یا نسبت مشتریان خوش‌حساب کمتر از ۷۰ درصد نیست. با توجه به این محدودیت‌ها، مدل مناسب و کارا برای این داده‌ها، رگرسیون بتا است. در این پایان‌نامه، پس از معرفی مدل، به برآورد پارامترهای مدل و استنباط آماری می‌پردازیم. سپس با استفاده از شبیه‌سازی، کارایی مدل پیشنهادی را مورد بررسی قرار می‌دهیم. در نهایت کاربرد مدل را در دو مثال واقعی، یکی داده‌های مربوط به درصد بنزین تبدیل‌شده از نفت خام و دیگری داده‌های مربوط به مقاوت بتن، نشان خواهیم داد.

تقدیم به پدر و مادر عزیزم

حمد و سپاس خدایی را که لطف و کرمش شامل من شد و این توانایی را به من عطا نمود تا گامی کوچک در اقیانوس بی‌کران علم و معرفت بردارم.

اکنون که با لطف خداوند متعال، این دوره تحصیلی را به پایان رسانده‌ام، ابتدا، از زحمات بی‌دریغ استاد راهنمای محترم، جناب آقای دکتر حسین باغیشنی که از ابتدای راه و در طی انجام تحقیق، با راهنمایی‌های خود مرا در نگارش این اثر یاری نمودند، سپاسگزاری می‌کنم. از آقای دکتر آرشی نیز که زحمت مشاوره این پایان‌نامه را بر عهده داشتند، تشکر می‌کنم. بر خود لازم می‌دانم که از اساتید محترم داور، جناب آقایان دکتر شاهسونی و دکتر کریمی، و همچنین از تمام اساتید فرهیخته گروه آمار دانشگاه شاهرود تشکر و قدردانی نمایم.

وظیفه خود می‌دانم که از زحمات پدر و مادر بسیار عزیز، دلسوز و فداکارم که پیوسته راهنمای من در زندگی بوده‌اند، قدردانی به عمل آورم. همچنین از برادرم و خواهرانم که همواره پشتیبان من بوده‌اند، صمیمانه سپاسگزاری می‌کنم.

با سپاس بی‌دریغ خدمت دوستان و همکلاسی‌های گران‌مایه‌ام، آقایان احسان اسحقی، علی یعقوبی، حمید کرمی‌کبیر، روح‌الله اکبری، دانیال ثقفی، مسعود شهابی و خانم‌ها رضایی و مقدم که مرا صمیمانه و مشفقانه یاری داده‌اند.

در پایان از همه دوستان گرامی و کسانی که در گردآوری این مجموعه حامی و پشتیبان من بوده‌اند، کمال تشکر را دارم.

میعاد ولیپور پاشاکلایی
دی ۱۳۹۲

پیش‌گفتار

یکی از اهداف بیشتر بررسی‌های آماری، یافتن رابطه‌هایی است که به کمک آن بتوان اثر تغییرات یک یا چند متغیر را بر روی متغیرهای دیگر ارزیابی کرد. معمول‌ترین ابزار برای نیل به این هدف، رگرسیون است. در بسیاری از کاربردها، ماهیت متغیر پاسخ به صورت نرخ و نسبت می‌باشد. به عنوان چند مثال: در علوم اقتصادی، اقتصاددانان به دنبال درک رابطه بین نرخ رشد، نرخ بیکاری، درصد تولید ملی و غیره با چند متغیر اقتصادی دیگر هستند؛ در صنعت، یک مدیر به دنبال یافتن عواملی است که باعث افزایش نرخ ایجاد نقص یک دستگاه مانند توربین می‌باشد؛ در علوم بانکداری، مدیران به دنبال شناخت نرخ بازپس‌دهی وام‌های واگذارشده به مشتریان (شناخت مشتریان خوش‌قول)، برای برنامه‌ریزی‌های آینده هستند؛ و در پزشکی، محققین به دنبال کشف این موضوع هستند که چه نسبتی از کل بیماران سرطانی، در یک منطقه خاص، دارای سرطان ریه می‌باشند. در واقع متغیر پاسخ در این گونه داده‌ها دارای تکیه‌گاهی در فاصله $(۰,۱)$ است. معمولا برای مدل‌بندی داده‌هایی با دامنه تغییرات $(۰,۱)$، از مدل‌های لجستیک و پرابیت استفاده می‌شود. اما از آن‌جا که داده‌های نسبت یا نرخ، معمولا در یک زیرفاصله مشخصی از دامنه تغییرات خود متمرکز هستند (به عبارتی توزیع این نوع داده‌ها به شدت چوله است)، مدل‌های لجستیک و پرابیت برای مدل‌بندی آن‌ها مناسب نیستند. به عنوان مثال، به طور معمول نرخ بیکاری بیشتر از ۳۰ درصد یا نسبت مشتریان خوش‌حساب کمتر از ۷۰ درصد نیست. با توجه به این محدودیت‌ها، مدل مناسب و کارا برای این داده‌ها، رگرسیون بتا است. با توجه به این مقدمه، ساختار این پایان‌نامه به صورت زیر است:

- در فصل اول، تعاریف و مفاهیم اولیه مورد نیاز در این پایان‌نامه را مطرح می‌کنیم.
- در فصل دوم، مدل رگرسیونی بتا، در حالتی که پارامتر دقت آن ثابت است، را معرفی می‌کنیم. سپس نحوه برازش و استنباط در آن را تشریح می‌کنیم و با مطالعه شبیه‌سازی، عملکرد آن را ارزیابی می‌کنیم.

- در فصل سوم، مدل رگرسیونی بتا را برای حالتی که پارامتر دقت نیز تابعی از متغیرهای تبیینی باشد، تعمیم می‌دهیم و نحوه استنباط در آن را تشریح می‌کنیم. سپس با مطالعه شبیه‌سازی، عملکرد این مدل را نیز مورد ارزیابی قرار می‌دهیم.

- در فصل چهارم، نحوه برازش و استنباط در مدل‌های رگرسیونی بتا را با استفاده از بسته $betareg$، موجود در نرم‌افزار R، تشریح می‌کنیم. سپس کاربرد مدل را با دو مثال واقعی شامل داده‌های نسبت بنزین تبدیل‌شده از نفت خام و میزان مقاومت بتن نمایش می‌دهیم.

- پیوست این پایان‌نامه نیز شامل کدهای نوشته‌شده در محیط R برای بازتولید مثال‌های موجود در پایان‌نامه است.

در این پایان‌نامه، مواردی را که برهان توسط خود نویسنده ارایه شده است، با علامت * نمایش داده‌ایم.

فهرست مطالب

فهرست جداول

فهرست تصاویر

فصل ۱
تعاریف و مفاهیم اولیه

۱.۱ مقدمه

در علم آمار، انتخاب تحلیل درست آماری یکی از موضوع‌های بسیار مهم و با اهمیت برای مدل‌بندی یک پدیده می‌باشد. این انتخاب درست به طبیعت داده‌ها و هدفی که از تحلیل آن‌ها داریم، بستگی دارد. آمار با تعریف مسأله و تعیین اهداف شروع می‌شود، سپس داده‌های مورد نیاز گردآوری و تحلیل می‌شوند و سرانجام با استنتاج نتایجی خاتمه می‌یابد.

مدل‌بندی رابطه‌های موجود بین متغیرها را رگرسیون می‌نامند. به عبارتی می‌توان گفت، رگرسیون تعیین روابط نادقیق بین متغیرهای آماری و تحلیل این روابط است (نیرومند، ۱۳۸۷). حال چرا به چنین روش قدرتمند آماری واژه دور از انتظار رگرسیون که معنی واژگانی آن برگشت می‌باشد، داده شده است؟ از نظر تاریخی، نخستین بار فرانسیس گالتن این واژه را به‌کار برد و با آن ارتباط بین میانگین طول قامت پدران و پسران را مورد مطالعه قرار داد. گالتن از روی مشاهداتش نتیجه گرفت طول قامت پسران با والدین خیلی بلندقد (یا کوتاه)، بلندتر (یا کوتاه‌تر) از طول قامت متوسط است؛ البته نه به بلندی (یا کوتاهی) قد والدینشان. این نتیجه‌گیری در سال ۱۸۸۵ با عنوان برگشت[۱] قد ارثی به‌سوی حد متوسط منتشر شد.

[۱] Regression

با این مقدمه، تحلیل رگرسیونی شامل تحلیل داده‌هایی شامل دو متغیر یا بیشتر، برای کشف ماهیت رابطه میان آن‌ها و سپس تحقیق برای هدف‌های پیش‌گویی است. در این‌جا می‌توان دو نوع متغیر را دسته‌بندی کرد: متغیر تبیینی (متغیر کنترل‌شده) و متغیر پاسخ. منظور از متغیرهای تبیینی، آن دسته از متغیرها هستند که می‌توانند مقادیر معینی اختیار کنند و کنترل‌پذیر هستند؛ مانند قیمت یک فرآورده جدید یا غلظت یک ماده شیمیایی سمی. همچنین مقادیری که اختیار می‌کنند می‌توانند قابل مشاهده ولی کنترل‌ناپذیر باشند؛ مانند نرخ تخلیه آب رودخانه و سطح سواد. در نتیجه تغییراتی که، به صورت عمدی یا غیرعمدی، برای متغیرهای تبیینی به‌وجود می‌آیند، یک اثر به متغیر پاسخ منتقل می‌شود؛ مانند امکان فروش یک فرآورده جدید، درصد مرگ و میر پشه‌ها و درصد افراد واجد شرایط برای رای دادن. تفاوت بین متغیرهای تبیینی و پاسخ، همیشه به‌طور کامل محسوس نیست و گاهی به هدف‌های پژوهش بستگی دارد، به این معنی که یک متغیر تبیینی ممکن است به عنوان متغیر پاسخ و یک متغیر پاسخ به عنوان یک متغیر تبیینی در نظر گرفته شود. برای مثال در بررسی رابطه بین طول قامت و وزن انسان‌ها، هم وزن می‌تواند به عنوان متغیر تبیینی ظاهر شود و طول قامت به عنوان متغیر پاسخ و هم برعکس آن.

در موارد متعددی مقادیر متغیر پاسخ در بازه $(۰, ۱)$ مشاهده می‌شوند، که برای تحلیل آن‌ها می‌توان از مدل‌های رگرسیونی لوجیت و پرابیت (هاسمر و لمشو، ۲۰۱۳) استفاده کرد. اما در زمانی که متغیر پاسخ دارای چولگی باشد، این دو مدل نیز نمی‌توانند به خوبی داده‌ها را برازش دهند. لذا با توجه به انعطاف بسیار خوب توزیع بتا، مدلی را ارایه می‌دهیم که به مدل رگرسیونی بتا معروف است (فراری و سریباری، ۲۰۰۴). انگیزه اصلی معرفی رگرسیون بتا در انعطاف‌پذیری توزیع بتا برای در نظر گرفتن انواع توزیع‌های چوله[۲] در فاصله $(۰, ۱)$ است. بنابراین، لازم است ابتدا توزیع بتا را معرفی کنیم.

[۲] Skewed Distributions

۲.۱ متغیر تصادفی بتا

اگر متغیر تصادفی Y دارای تابع چگالی احتمال

$$f_{p,q}(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-۱}(۱-y)^{q-۱}, \qquad ۰ < y < ۱, \qquad p > ۰, \qquad q > ۰,$$

باشد، گوییم Y دارای توزیع بتا با پارامترهای p و q است و آن را با نماد $Y \sim Beta(p,q)$ نشان می‌دهیم. توزیع بتا می‌تواند برای مدل‌بندی پدیده‌های تصادفی که مجموعه مقادیر ممکن آن یک فاصله متناهی $[c,d]$ است، به‌کار رود. با در نظر گرفتن c به عنوان مبدا و اختیار $d-c$ به عنوان واحد اندازه‌گیری، می‌توان فاصله فوق را به فاصله $[۰,۱]$ تبدیل کرد. توزیع بتا، توزیع احتمالی پیوسته‌ایست که بر بازه $(۰,۱)$ تعریف شده و دارای دو پارامتر شکل[۳] است. به عنوان مثال، فرض کنید متغیر تصادفی Y نشان‌دهنده درصدی از درآمد فرد بوده که پس‌انداز می‌شود. همچنین فرض کنید Y دارای توزیع بتا با پارامترهای $p = ۲$ و $q = ۱۹$ (یعنی $Y \sim B(۲,۱۹)$) با تابع چگالی به‌صورت زیر باشد:

$$f_Y(y) = \frac{\Gamma(۲۱)}{\Gamma(۲)\Gamma(۱۹)} y^{۲-۱}(۱-y)^{۱۹-۱} \quad ۰ < y < ۱.$$

تابع توزیع به‌صورت

$$F_Y(y) = \begin{cases} ۰ & y \le ۰, \\ \int_۰^y ۳۸۰ y(۱-y)^{۱۸} dy = ۱ - (۱-y)^{۱۹}(۱+۱۹y) & ۰ < y < ۱, \\ ۱ & y \ge ۱, \end{cases}$$

به‌دست می‌آید. اکنون، به‌عنوان مثال، احتمال این‌که فردی کمتر از ۰٫۰۵ درآمد خود را پس‌انداز کند، به‌صورت زیر قابل محاسبه است:

$$P(Y < ۰٫۰۵) = F_Y(۰٫۰۵) = ۰٫۲۶۴.$$

این توزیع را برای تحلیل داده‌هایی با ماهیت نرخ، درصد و نسبت، می‌توان استفاده کرد.

[۳] Shape Parameter

۳.۱ مدل رگرسیون خطی

مدل ارتباطی (ضابطه ریاضی) حاکم بر متغیر پاسخ و مجموعه متغیرهای تبیینی، توسط افراد متخصص در زمینه مورد مطالعه، بر مبنای دانش یا قضاوت‌های عینی و ذهنی آن‌ها تعیین می‌شود. این مدل که می‌تواند پارامترهای نامعلوم زیادی را در بر داشته باشد، یک مدل پارامتری نامیده می‌شود. اگر در معادله رگرسیونی تنها یک متغیر تبیینی وجود داشته باشد، آن‌را مدل رگرسیونی خطی ساده می‌نامند. در یک مدل رگرسیونی، هدف بررسی اثر متغیرهای تبیینی بر روی متغیر پاسخ است. معمولا، متغیرهای تبیینی را با X (که می‌تواند بعدی بزرگ‌تر از ۱ داشته باشد) و متغیر پاسخ را با Y نمایش می‌دهند. همچنین این ضابطه می‌تواند خطی یا غیرخطی باشد. باید توجه داشته باشیم که منظور از رابطه خطی یا غیرخطی، در مورد رابطه بین Y و X_iها نیست بلکه در واقع پارامترهای رگرسیونی به‌طور خطی یا غیرخطی وارد معادله رگرسیون می‌شوند. به‌عنوان مثال، هر یک از روابط زیر خطی‌اند هر چند که رابطه بین X و Y ممکن است خطی نباشد:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon, \\ Y &= \beta_0 + \beta_1 e^X + \varepsilon, \\ Y &= \beta_0 + \beta_1 \log X + \varepsilon. \end{aligned}$$

همچنین ε جمله خطا می‌باشد که متغیری است تصادفی و غیر قابل مشاهده. مثال‌هایی از یک مدل رگرسیونی عبارتند از:

(۱) بررسی اثر تعداد قطعات لازم در ساخت یک دستگاه (متغیر تبیینی) بر مدت زمان لازم بر تولید آن (متغیر پاسخ)

(۲) بررسی تاثیر سرانه خالص ملی (متغیر تبیینی) بر تولید خالص ملی (متغیر پاسخ)

(۳) بررسی تاثیر سن (متغیر تبیینی) بر تعداد ضربان قلب (متغیر پاسخ)

در مدل رگرسیونی خطی، متغیر پاسخ Y یک متغیر تصادفی پیوسته است، در حالی که X تصادفی نبوده و توسط تحلیل‌گر کنترل و با خطای قابل اغماضی اندازه‌گیری می‌شود. بنابراین به ازای هر مقدار ممکن X برای Y یک توزیع احتمال وجود دارد. به‌طور کلی، منظور از مدل رگرسیونی این است که $E(Y|x)$ را بر حسب تابعی از X، یا Y را بر حسب تابعی از X، به همراه یک جمله خطا، ε، که امید ریاضی آن صفر فرض می‌شود، بنویسیم. بنابراین

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad E(Y|X) = \beta_0 + \beta_1 X,$$

هر دو معادل‌اند.

هر مدل آماری مبتنی بر پذیره‌هایی ساخته می‌شود. برای مدل رگرسیونی خطی نیز چهار پذیره در نظر گرفته می‌شود:

(۱) خطی بودن ضابطه تابع رگرسیونی $E(Y|X)$

(۲) ثابت بودن واریانس جمله خطا

(۳) ناهمبسته بودن مولفه‌های خطا

(۴) (بر حسب نیاز) نرمال بودن خطاها

تخطی از هر کدام از این پذیره‌ها (به ویژه پذیره اول)، می‌تواند منجر به نامناسب شدن مدل مورد نظر شود.

۴.۱ مدل‌های خطی تعمیم‌یافته

در سال‌های اخیر، مدل‌های خطی تعمیم‌یافته[۴] (GLM) به‌طور چشم‌گیری مورد استفاده قرار گرفته‌اند و متون آماری متعددی در این زمینه به رشته تحریر درآمده‌اند. قسمتی از روند رشد این دسته از مدل‌ها

[۴] Generalized Linear Models

را می‌توان بدین شرح نوشت: مدل‌های خطی تعمیم‌یافته، به عنوان تعمیم مدل‌های خطی، توسط نلدر و ودربرن (۱۹۷۲) معرفی شد. بررسی نظری دقیق این دسته از مدل‌ها در مک کالا و نلدر (۱۹۸۹) آمده است. میر و همکاران (۲۰۰۲) نیز منبع خوبی برای دیدگاه شهودی این مدل‌ها است. لازم به ذکر است که بیشترین استفاده از این مدل‌ها به عنوان رقیبی قوی در برابر مدل‌های لگ خطی[۵] در تحلیل داده‌های گسسته گسترش یافته است. برای اطلاعات بیشتر در این مورد و ارتباط با مدل‌های لگ خطی به اگرستی (۱۹۹۶) مراجعه کنید.

همان‌طور که از نام مدل‌های خطی تعمیم‌یافته بر می‌آید، تعمیمی از مدل‌های خطی مبتنی بر توزیع نرمال هستند. این تعمیم شامل دو قسمت می‌باشد:

۱. در GLM توزیع‌هایی به‌جز نرمال نیز می‌توانند به‌عنوان توزیع متغیر پاسخ در نظر گرفته شوند. البته توزیع در نظر گرفته شده باید متعلق به خانواده نمایی یا شبیه به این خانواده (مک کالا و نلدر، ۱۹۸۹) باشد.

۲. میانگین پاسخ به‌طور مستقیم مدل‌بندی نمی‌شود، بلکه تبدیلی از آن با استفاده از تابع پیوند[۶] مدل‌بندی می‌گردد.

در واقع، شناخت ویژگی‌های خوب خانواده توزیع‌های نمایی و گسترش روش‌های عددی برای برآورد پارامترها، این امکان را به ما می‌دهد تا در حالت‌هایی که متغیرهای پاسخ دارای توزیع‌هایی بجز نرمال هستند و حالت‌هایی که رابطه بین متغیر پاسخ و متغیرهای تبیینی به شکل خطی ساده نیستند، از روش‌های مشابه با آن‌چه برای مدل‌های خطی داده شده‌اند، استفاده کنیم. قبل از بیان مدل خطی تعمیم‌یافته، معرفی خانواده توزیع‌های نمایی لازم است.

[۵] Log-Linear Models
[۶] Link Function

۱.۴.۱ خانواده توزیع‌های نمایی

خانواده‌ای از توزیع‌ها را عضو خانواده توزیع‌های نمایی گوییم، اگر بتوان تابع چگالی احتمال آن را به‌صورت زیر نوشت:

$$f_{Y_i}(y_i) = \exp\left\{\frac{[y_i\gamma_i - b(\gamma_i)]}{\tau^2} - c(y_i,\tau)\right\}, \qquad (۱.۱)$$

که در آن γ_i و τ پارامترهای خانواده توزیع و $b(\gamma_i)$ و $c(y_i,\tau)$ توابع معلوم هستند. برای بسیاری از اعضای این خانواده τ برابر با ۱ است.

ویژگی‌های خانواده توزیع‌های نمایی در سال‌های ۱۹۳۷ و ۱۹۳۸، به‌طور مستقل، توسط سه آماردان به نام‌های کوپمن، پیتمن و دارمویس مطالعه و بررسی شده‌اند. می‌توان نشان داد که اگر Y_i دارای توزیعی از خانواده نمایی باشد، امید ریاضی و واریانس آن به‌صورت زیر خواهند بود:

$$E(Y_i) = \mu = b'(\gamma_i), \qquad (۲.۱)$$

$$Var(Y_i) = \sigma_i^2 = b''(\gamma_i)\tau^2. \qquad (۳.۱)$$

ازجمله توزیع‌های نمایی معروف می‌توان به توزیع نرمال، دوجمله‌ای، پواسن، نمایی، گاما، بتا و توزیع گاوسی معکوس اشاره کرد. به عنوان مثال، توزیع نرمال را در نظر بگیرید:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\big[-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}\big].$$

که می‌توان آن را به شکل خانواده نمایی نوشت:

$$f(y_i) = exp[\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)]$$

میانگین و واریانس آن نیز مشابه روابط (۲.۱) و (۳.۱) به‌دست می‌آیند:

$$E(Y_i) = b'(\gamma_i) = \gamma_i = \mu_i,$$

$$var(Y_i) = b''(\gamma_i)\tau^2 = \sigma^2.$$

۲.۴.۱ مولفه‌های مدل‌های خطی تعمیم‌یافته

قبل از معرفی مولفه‌های یک مدل خطی تعمیم‌یافته، چند سوال مطرح می‌کنیم:

۱. داده‌ها دارای چه توزیعی هستند؟

۲. پیشگوی خطی چه باشد؟

۳. چه تابعی از میانگین برای پیشگوی خطی در نظر گرفته شود؟

با طرح این سوال‌ها، مولفه‌های مدل‌های خطی تعمیم‌یافته را به‌صورت زیر معرفی می‌کنیم:

۱. **مولفه تصادفی:** مشاهدات y_i، $i = ۱, \cdots, n$، از هم مستقل هستند و توزیع متغیر پاسخ، عضوی از خانواده نمایی به‌صورت (۱.۱) است.

۲. **مولفه سیستماتیک:** این مولفه شامل متغیرهای تبیینی $x_۱, ..., x_p$ و بردار پارامترهای $\boldsymbol{\beta}$ است که پیشگوی خطی را به‌صورت $\eta = \sum_{j=۱}^{p} x_j \boldsymbol{\beta}_j$، تشکیل می‌دهند.

۳. **تابع پیوند:** تابع پیوند به منظور ارتباط دادن مولفه سیستماتیک به مولفه تصادفی استفاده می‌شود. تابع پیوند $g(\cdot)$ یک تابع یکنوا و مشتق‌پذیر است، به‌طوری‌که

$$g(\mu_i) = \eta_i = x_i' \beta.$$

در مدل‌های خطی، پیشگوی خطی و میانگین یکی هستند؛ یعنی $\eta = \mu$. به عبارتی، تابع پیوند یک تابع همانی است. دومین تعمیم GLM مربوط به انتخاب تابع پیوند است که هر تابع مشتق‌پذیر یکنوا را شامل می‌شود. به‌عنوان مثال، در توزیع پواسن که معمولا برای مدل‌بندی داده‌های شمارشی به‌کار می‌رود، میانگین توزیع، μ، نامنفی است، در حالی که ممکن است پیشگوی خطی، η، مثبت نباشد. پس تابع پیوند همانی نمی‌تواند انتخاب مناسبی باشد. در این حالت، می‌توان تابع پیوند را به‌صورت لگاریتمی در نظر گرفت:

$$\eta = g(\mu) = \ln(\mu).$$

در این صورت خواهیم داشت

$$\mu = e^{\eta} > ۰.$$

به این ترتیب اثرات جمعی موثر در η به اثرات ضربی موثر در μ تبدیل می‌شوند. به‌عنوان مثالی دیگر، در توزیع برنولی، $ber(p)$، میانگین توزیع برابر p است که همان احتمال موفقیت تعبیر می‌شود. در این حالت $p \in (۰, ۱)$. بنابراین تابع پیوند باید طوری انتخاب شود که پیشگوی خطی را به فاصله $(۰, ۱)$ بنگارد. به‌عنوان مثال، می‌توان تابع پیوند لوجیت را، $g(p) = logit(p) = \log\left(\frac{p}{۱-p}\right)$، تعریف کرد.

هر توزیع خانواده نمایی یک تابع پیوند ویژه دارد که برای آن آماره بسنده وجود دارد. چنین تابع پیوندی، زمانی ایجاد می‌شود که $\gamma = \eta$ باشد و آن را تابع پیوند متعارف[۷] می‌نامند. برای توزیع پواسن، تابع پیوند لگاریتمی[۸] تابع پیوند متعارف است. پیوند متعارف برای توزیع نرمال، همانی و برای توزیع برنولی، لجیت است.

از آن‌جایی‌که تابع پیوند، یکنوا و در نتیجه معکوس‌پذیر است، می‌توان رابطه زیر را نوشت:

$$\mu_i = g^{-۱}(\eta_i) = g^{-۱}(x_i'\beta).$$

۳.۴.۱ برازش مدل‌های خطی تعمیم‌یافته

اگر y دارای تابع چگالی به‌صورت (۱.۱) باشد، آن‌گاه تابع لگاریتم درست‌نمایی آن به‌صورت زیر خواهد بود:

$$\ell = \sum_{i=۱}^{n} \frac{[y_i\gamma_i - b(\gamma_i)]}{\tau^{۲}} - \sum_{i=۱}^{n} c(y_i, \tau). \qquad (۴.۱)$$

قبل از این‌که معادله درست‌نمایی را به‌دست آوریم، نیاز است که برخی مشخصه‌های مفید آن را بیان کنیم. این مشخصه‌ها، شامل امید ریاضی و واریانس می‌باشند که طبق شرایط نظم (کسلا و برگر، ۲۰۰۲)، خواهیم داشت

$$E\left[\frac{\partial \ell}{\partial \gamma_i}\right] = ۰, \qquad (۵.۱)$$

[۷] Canonical Link Function
[۸] Log Link Function

و

$$var\left(\frac{\partial \ell}{\partial \gamma_i}\right) = -E\left[\frac{\partial^2 \ell}{\partial \gamma_i^2}\right]. \tag{۶.۱}$$

با توجه با روابط (۱.۱) و (۵.۱)، داریم

$$E\left[\frac{\left\{Y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\}}{\tau^2}\right] = 0,$$

یا به عبارت دیگر

$$E[Y_i] = \mu_i = \frac{\partial b(\gamma_i)}{\partial \gamma_i}. \tag{۷.۱}$$

همچنین با استفاده از دو رابطه (۱.۱) و (۶.۱)، می‌توان نوشت

$$var\left(\frac{\left\{Y_i - \frac{\partial b(\gamma_i)}{\partial \gamma_i}\right\}}{\tau^2}\right) = -E\left[-\frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right],$$

که با توجه به رابطه (۷.۱)، داریم

$$var\left(\frac{Y_i - \mu_i}{\tau^2}\right) = \frac{1}{\tau^2}\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2},$$

یا

$$var(Y_i) = \tau^2 \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}.$$

در نهایت می‌توان نوشت

$$var(Y_i) \equiv \tau^2 v(\mu_i),$$

که $v(\mu_i) = \frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}$ به‌عنوان تابع واریانس برای متغیر پاسخ y_i شناخته می‌شود. در همین راستا، می‌توان به دو ویژگی دیگر اشاره کرد:

$$\frac{\partial \gamma_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \gamma_i}\right)^{-1} = \left(\frac{\partial^2 b(\gamma_i)}{\partial \gamma_i^2}\right)^{-1} = \frac{1}{v(\mu_i)}, \tag{۸.۱}$$

و با توجه به رابطه $g(\mu_i) = x_i'\beta$ و قاعده زنجیره‌ای، می‌توان نوشت

$$\begin{aligned}\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial \mu_i}{\partial g(\mu_i)}\frac{\partial g(\mu_i)}{\partial \beta} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1}\frac{\partial x_i'\beta}{\partial \beta} \\ &= \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} x_i' = \frac{1}{g_\mu(\mu_i)}x_i,\end{aligned} \tag{۹.۱}$$

که در آن $g_\mu(\mu_i) = \frac{\partial g(\mu_i)}{\partial \mu_i}$. حال می‌توان معادله درست‌نمایی را برای β به‌دست آورد. با توجه به رابطه (۴.۱)، داریم

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\tau^2} \sum_i \left[y_i \frac{\partial \gamma_i}{\partial \beta} - \frac{\partial b(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \beta} \right].$$

لذا بنابر رابطه (۷.۱) می‌توان نوشت

$$\frac{1}{\tau^2} \sum_i \left[y_i \frac{\partial \gamma_i}{\partial \beta} - \frac{\partial b(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \beta} \right] = \frac{1}{\tau^2} \sum_i (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \beta},$$

که با استفاده از قاعده زنجیره‌ای، نتیجه می‌شود

$$\frac{1}{\tau^2} \sum_i (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \beta} = \frac{1}{\tau^2} \sum_i (y_i - \mu_i) \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta}.$$

در نهایت با توجه به روابط (۸.۱) و (۹.۱)، خواهیم داشت

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \frac{1}{\tau^2} \sum_i \frac{(y_i - \mu_i)}{v(\mu_i) g_\mu(\mu_i)} x_i' \\ &= \frac{1}{\tau^2} \sum_i (y_i - \mu_i) w_i g_\mu(\mu_i) x_i', \end{aligned} \tag{۱۰.۱}$$

که در آن $w_i = [v(\mu_i) g_\mu^2(\mu_i)]^{-1}$. شکل ماتریسی رابطه (۱۰.۱) را می‌توان به‌صورت زیر نوشت:

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\tau^2} X' W \Delta (y - \mu), \tag{۱۱.۱}$$

که در آن $W = diag(w_i)$ و $\Delta = diag(g_\mu(\mu_i))$ و منظور از $diag(w_i)$، یک ماتریس قطری n بعدی با مولفه‌های w_i است. بنابراین معادله درست‌نمایی را می‌توان به‌صورت

$$X' W \Delta y = X' W \Delta \mu, \tag{۱۲.۱}$$

نوشت، به‌طوری‌که W، Δ و μ، شامل پارامترهای نامعلوم β هستند. معمولا این‌ها توابعی غیرخطی از β هستند و نمی‌توان با استفاده از روابط جبری سرراست، رابطه (۱۲.۱) را حل کرد. به عنوان مثال، برای مدل با تابع پیوند پرابیت، با توجه به رابطه $p = \Phi(X\beta)$، که در آن y دارای توزیع برنولی با پارامتر p است و $\Phi(\cdot)$ تابع توزیع نرمال استاندارد، لگاریتم تابع درست‌نمایی و مشتق آن نسبت به β، به‌صورت

زیر محاسبه می‌شوند:

$$\ell = \sum_{i=۱}^{n} (y_i \{\log \Phi(x_i'\beta) - \log[۱ - \Phi(x_i'\beta)]\} + \log[۱ - \Phi(x_i'\beta)]),$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta} &= \sum_{i=۱}^{n} \left[y_i \left(\frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} x_i' + \frac{\phi(x_i'\beta)}{۱ - \Phi(x_i'\beta)} x_i' \right) - \frac{\phi(x_i'\beta)}{۱ - \Phi(x_i'\beta)} x_i' \right] \\
&= \sum_{i=۱}^{n} \frac{[y_i - \Phi(x_i'\beta)]\phi(x_i'\beta)}{\Phi(x_i'\beta)[۱ - \Phi(x_i'\beta)]} x_i' \\
&= \sum_{i=۱}^{n} \frac{(y_i - \mu_i)\phi(x_i'\beta)}{\mu_i(۱ - \mu_i)} x_i'.
\end{aligned}$$

برای حل معادله‌های تابع درست‌نمایی، باید از مشتق دوم تابع لگاریتم درست‌نمایی، امید ریاضی بگیریم. به همین منظور، می‌توان نوشت

$$\frac{\partial^۲ \ell}{\partial \beta \partial \beta'} = -\frac{۱}{\tau^۲} X' W \Delta \frac{\partial \mu}{\partial \beta'} + \frac{۱}{\tau^۲} X' \frac{\partial W \Delta}{\partial \beta'} (y - \mu).$$

بنابراین

$$-E\left[\frac{\partial^۲ \ell}{\partial \beta \partial \beta'}\right] = \frac{۱}{\tau^۲} X' W \Delta \frac{\partial \mu}{\partial \beta'} + ۰.$$

لذا با استفاده از رابطه (۹.۱)، می‌توان نوشت

$$\frac{۱}{\tau^۲} X' W \Delta \frac{\partial \mu}{\partial \beta'} + ۰ = \frac{۱}{\tau^۲} X' W \Delta \Delta^{-۱} X.$$

در نهایت خواهیم داشت

$$-E\left[\frac{\partial^۲ \ell}{\partial \beta \partial \beta'}\right] = \frac{۱}{\tau^۲} X' W X, \tag{۱۳.۱}$$

که در آن $W = diag(w_i) = \left([v(\mu_i) g_\mu^۲(\mu_i)]^{-۱}\right)$. همچنین بنابر مک‌کالاک و سیرل (۲۰۰۱)، داریم

$$\begin{aligned}
-E\left[\frac{\partial^۲ \ell}{\partial \beta \partial \tau^۲}\right] &= -E\left[\frac{\partial}{\partial \tau^۲} \frac{۱}{\tau^۲} X' W \Delta (y - \mu)\right] \\
&= \frac{۱}{\tau^۴} X' W \Delta E[y - \mu] \qquad (۱۴.۱) \\
&= ۰.
\end{aligned}$$

برای حل معادله ماکسیمم درست‌نمایی در رابطه (۱۲.۱) ، معمولا از روش تکراری کمترین توان‌های دوم موزون[۹] استفاده می‌شود. استفاده از این روش به فرمول تکراری زیر منتهی می‌شود:

$$\theta^{(m+1)} = \theta^{(m)} + I(\theta^{(m)})^{-1}\frac{\partial l}{\partial \theta}|_{\theta=\theta^{(m)}},$$

که در آن m تعداد تکرارها، $I(\theta)$ ماتریس اطلاع فیشر[۱۰] و θ بردار پارامترها می‌باشند.

حال با توجه به روابط (۱۱.۱)، (۱۳.۱) و (۱۴.۱) می‌توان نوشت

$$\beta^{(m+1)} = \beta^{(m)} + (X'WX)^{-1}X'W\Delta(y-\mu).$$

۴.۴.۱ ارزیابی مدل

یکی از مراحل اصلی مدل‌بندی آماری، ارزیابی یا همان سنجش نیکویی برازش مدل است. برای مدل‌های خطی تعمیم‌یافته، این سنجش توسط معیارهایی انجام می‌شود که در این قسمت به آن‌ها اشاره می‌کنیم.

- **آزمون نسبت درست‌نمایی و انحراف**

کیبش[۱۱] یا نیکویی برازش مدل می‌تواند به روش‌های مختلف بررسی شود. یکی از روش‌های مهم در این رده از مدل‌ها، لگاریتم نسبت درست‌نمایی‌هاست که انحراف نامیده می‌شود. در واقع این معیار، درست‌نمایی مربوط به مدل مورد نظر را با درست‌نمایی مدل کامل مقایسه می‌کند. مدل کامل، مدلی است که به اندازه مشاهدات پارامتر دارد (مک کالاک و سیرل، ۲۰۰۱). به‌جای آن‌که تابع درست‌نمایی را تابعی از پارامتر متعارف، γ، در نظر بگیریم، مناسب‌تر است که آن را بر حسب پارامتر میانگین، μ، بیان کنیم. به این ترتیب، اگر $L(\widehat{\mu},\tau,y)$ ماکسیمم تابع درست‌نمایی برحسب β باشد، به‌طوری‌که ϕ مقداری ثابت است و $L(\mu,\tau,y)$ ماکسیمم تابع درست‌نمایی تحت مدل کامل با n پارامتر در نظر گرفته شود، آن‌گاه آماره نسبت درست‌نمایی تعمیم‌یافته

$$\lambda = \frac{L(\mu,\tau,y)}{L(\widehat{\mu},\tau,y)},$$

[۹] Minimum Weighted Least Square
[۱۰] Fisher's Information Matrix
[۱۱] Deviance

اندازه‌ای از نیکویی برازش مدل را نشان می‌دهد.

اگر مدل مورد نظر، داده‌ها را به خوبی توصیف کند، در این صورت $L(\widehat{\mu}, \tau, y)$، باید به‌طور تقریبی برابر با $L(\mu, \tau, y)$ شود. آماره منهای لگاریتم نسبت درست‌نمایی را به صورت زیر تعریف می‌کنیم و آن را انحراف برازش می‌نامیم:

$$-۲\log\lambda = ۲\sum_{i=۱}^{n}\frac{y_i(\tilde{\gamma}_i - \widehat{\gamma}_i) - b(\tilde{\gamma}_i) + b(\widehat{\gamma}_i)}{a_i(\tau)},$$

که در آن $\widehat{\gamma}_i = \gamma_i(\hat{\beta})$ و $\tilde{\gamma}_i = \gamma_i(\tilde{\beta})$ به ترتیب برآوردهای پارامترهای مدل مورد نظر و مدل کامل می‌باشند.

بافرض این‌که $a_i(\tau) = \frac{\tau}{p_i}$، آن‌گاه کیبش را می‌توان به‌صورت

$$D_{\widehat{\mu}}(y) = ۲\sum_{i}^{n} p_i[y_i(\tilde{\gamma}_i - \widehat{\gamma}_i) - b(\tilde{\gamma}_i) + b(\widehat{\gamma}_i)],$$

تعریف کرد که $D_{\widehat{\mu}}(y)$ را کیبش مقیاس‌بندی‌شده[۱۲] می‌نامند.

مثال ۱.۴.۱. در توزیع نرمال $N(\mu, \sigma^۲)$، داریم $\gamma_i = \mu_i$، $b(\gamma_i) = \frac{۱}{۲}\gamma_i^۲$ و $a_i(\tau) = \sigma^۲$. در این توزیع p_iها برابر یک هستند. بنابراین

$$\begin{aligned} D_{\widehat{\mu}}(y) &= ۲\sum_{i}\{y_i(y_i - \widehat{\mu}_i) - \frac{۱}{۲}y_i^۲ + \frac{۱}{۲}\widehat{\mu}_i^۲\} \\ &= ۲\sum_{i}\{\frac{۱}{۲}y_i^۲ - y_i\widehat{\mu}_i^۲ + \frac{۱}{۲}\widehat{\mu}_i^۲ \\ &= \sum_{i}(y_i - \widehat{\mu}_i)^۲. \end{aligned}$$

ملاحظه می‌کنید که برای توزیع نرمال، کیبش مقیاس‌بندی‌شده، برابر با مجموع توان‌های دوم باقی‌مانده‌هاست.

برای مقایسه دو مدل $M_۱$ با $p_۱$ پارامتر و $M_۲$ با $p_۲$ پارامتر، به‌طوری‌که $M_۱ \in M_۲$ و $p_۲ > p_۱$ (که در واقع مدل $M_۱$ در مدل $M_۲$ لانه‌ای[۱۳] است)، لگاریتم نسبت درست‌نمایی ماکسیمم برای

[۱۲] Scaled Deviance
[۱۳] Nested

دو مدل را می‌توان به‌صورت تفاضل انحراف‌های آن دو نوشت:

$$-۲\log\lambda = \frac{D(M_۱) - D(M_۲)}{\phi}.$$

انحراف برای مدل‌های خطی دارای توزیع کی‌دو می‌باشد و برای مدل‌های خطی تعمیم‌یافته به طور مجانبی دارای توزیع کی‌دو است (بیکل و داکسوم، ۲۰۰۱). مزیت کلی انحراف، به‌عنوان یک اندازه انحراف، این است که اگر از برآوردهای درست‌نمایی ماکسیمم استفاده شود، انحراف برای مدل‌های لانه‌ای جمع‌پذیر است، در صورتی که آماره کی‌دو چنین خاصیتی ندارد (مک‌کالاک و سیرل، ۲۰۰۱). نقش تحلیل انحراف در GLM مشابه تحلیل واریانس در مدل‌های خطی کلاسیک می‌باشد و برای بررسی معنی‌داری اثر متغیرهای تبیینی بر متغیر پاسخ نیز مورد استفاده قرار می‌گیرد. به عنوان مثال اگر انحراف کوچک باشد، می‌توانیم بگوییم مدل برازش‌شده خوب است و اگر بزرگ باشد، می‌گوییم مدل برازش‌شده خوب نیست (ضعیف است).

انحراف را می‌توان به صورت $D_{\hat{\mu},\phi}(y) = \sum_{i=۱}^{n}(r_i^d)^۲$ نیز نوشت، که در آن

$$r_i^d = sign(y_i - \hat{\mu}_i)\left\{۲(\ell_i(\mu_i,\phi) - \ell_i(\hat{\mu}_i,\phi))\right\}^{\frac{۱}{۲}}.$$

همچنین r_i^d را باقی‌مانده‌های انحراف نیز می‌نامند. معیار مهم دیگری که برای انحراف برازش وجود دارد، آماره کی‌دو پیرسون[۱۴] است که به‌صورت

$$X^۲ = \sum \frac{(y-\hat{\mu})^۲}{V(\hat{\mu})},$$

تعریف می‌شود.

- **آزمون والد[۱۵]**

یکی دیگر از روش‌های بسیار عمومی برای آزمون فرضیه، آزمون والد است (والد، ۱۹۴۳). اگر $\hat{\theta}$ یک برآورد از θ با بعد p و $I(\theta)$ ماتریس اطلاع فیشر برای $\hat{\theta}$ باشد، آنگاه

$$(\hat{\theta} - \theta_*)'[I(\theta_*)]^{-۱}(\hat{\theta} - \theta_*),$$

[۱۴] Chi-Square Pearson Statistic
[۱۵] Wald's Test

یک آماره آزمون (آماره والد) برای فرضیه $H_۰ : \theta = \theta_*$ است. توزیع مجانبی این آماره کی‌دو با p درجه آزادی، $\chi^۲_p$، می‌باشد. البته توزیع آن برای مدل‌های خطی، دقیقا $\chi^۲_p$ است (یانکینگ و ونگ، ۲۰۱۱).

- **تحلیل باقی‌مانده‌ها**

آماره‌های نیکویی برازش، معیار کلی برای کفایت یک الگو را فراهم می‌کنند، اما بررسی جنبه‌های مختلف GLM، شامل انتخاب تابع واریانس، تابع پیوند و جمله‌های پیشگوی خطی، به منظور بررسی کفایت یک مدل برازش‌شده، با استفاده از باقی‌مانده‌ها انجام می‌شود. همچنین باقی‌مانده‌ها مقادیر غیرعادی را شناسایی می‌کنند. برای مدل‌های خطی تعمیم‌یافته، به تعمیم تعریف باقی‌مانده‌ها نیاز داریم، به‌طوری‌که برای همه توزیع‌هایی که می‌توانند به‌جای نرمال جایگزین شوند، قابل کابرد باشد.

سه شکل تعمیم‌یافته از باقی‌مانده‌ها در GLM تعریف می‌شوند که عبارتند از باقی‌مانده‌های پیرسونی[۱۶]، انس‌کمب[۱۷] و انحراف (مک کالا و نلدر، ۱۹۸۹). به عنوان مثال، باقی‌مانده‌های پیرسونی به صورت $r_p = \frac{y-\mu}{\sqrt{v(\mu)}}$ تعریف می‌شوند.

۵.۱ معیارهای اندازه‌گیری دقت مدل

به‌منظور ارزیابی توانایی مدل‌ها در برآورد متغیر پاسخ، از برخی از معیارها استفاده می‌شود تا بتوان نتایج مدل‌های مختلف را با یکدیگر مقایسه کرد. در واقع، این معیارها با اندازه‌گیری خطای برآورد یا دقت برآورد یک مدل، نتایج حاصل را قابل ارزیابی و قیاس‌پذیر با سایر نتایج می‌کنند. در آمار، معیارهای متفاوتی برای این هدف وجود دارند. در این زیربخش، سه معیار پرطرفدار را معرفی می‌کنیم.

[۱۶] Pearson Residuals
[۱۷] Anscombe Residuals

۱.۵.۱ میانگین توان دوم خطا

یکی از راه‌های اندازه‌گیری خطای برآورد مدل، استفاده از معیار میانگین توان دوم خطا[۱۸] (MSE) است. برای مقایسه برآوردگرهای اریب با برآوردگرهای نااریب یا مقایسه دو برآوردگر یک پارامتر، با اریبی‌های مختلف، MSE ملاکی مفید است و مقدار آن به‌صورت

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^۲],$$

به‌دست می‌آید که در آن، θ مقدار واقعی پارامتر و $\hat{\theta}$ مقدار برآوردشده پارامتر می‌باشند. چنان‌چه مقدار $MSE(\hat{\theta})$ به سمت صفر میل کند، بیان‌گر سازگار[۱۹] بودن $\hat{\theta}$ است.

۲.۵.۱ اریبی

در آمار، اریبی[۲۰] برآوردگر یک پارامتر برابر است با اختلاف امید ریاضی برآوردگر و مقدار واقعی آن پارامتر. به‌عبارتی اگر $\hat{\theta}$ برآوردگر پارامتر باشد، اریبی به صورت زیر تعریف می‌شود:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta = E(\hat{\theta} - \theta).$$

۳.۵.۱ معیار اطلاع آکاییک

معیار اطلاع آکاییک[۲۱] (AIC)، معیاری برای انتخاب بهترین مدل آماری از بین چند نامزد است. این معیار، نیکویی برازش مدل را همراه با پیچیدگی آن، به‌طور توام، در نظر می‌گیرد که توسط آکاییک در سال ۱۹۷۴ پیشنهاد شد. با داده‌های مشاهده‌شده، چند مدل رقیب بر حسب مقدار AIC رتبه‌بندی می‌شوند و مدل با کمترین AIC بهترین است. در حالت کلی، AIC به صورت

$$AIC = ۲k - ۲ln(L),$$

[۱۸] Mean Squared Error
[۱۹] Consistency
[۲۰] Bias
[۲۱] Akaike Information Criterion

تعریف می‌شود که در آن k تعداد پارامترهای مدل و L مقدار ماکسیمم تابع درست‌نمایی برای مدل برآوردشده است.

۶.۱ سایر تعاریف لازم

در این بخش، چند تعریف که در این پایان‌نامه از آن‌ها استفاده می‌شود را بیان می‌کنیم. جزییات بیشتر این تعاریف را می‌توانید در (استرانگ ۲۰۰۳) بیابید.

تعریف ۱.۶.۱. ماتریس ترانهاده[۲۲]

اگر جای سطرها یا ستون‌های ماتریس A را با ستون یا سطرهای آن تعویض نماییم، ماتریس حاصل ترانهاده ماتریس A نامیده می‌شود و با نماد A' نمایش می‌دهیم:

$$A' = (a_{ij})' = (a_{ji}),$$

که در آن a_{ij} عنصر موجود در سطر iام و ستون jام می‌باشد.

تعریف ۲.۶.۱. دترمینان[۲۳]

دترمینان ماتریس $A_{p\times p}$ را با نماد $|A|$ نشان داده که تابعی اسکالر است و به‌صورت زیر تعریف می‌شود:

$$\begin{cases} |A| = a_{۱۱} & p = ۱, \\ |A| = \sum_{j=۱}^{p} |A_{۱j}|(-۱)^{۱+j} & p > ۱, \end{cases}$$

که در آن $A_{۱j}$ ماتریسی $(n-۱)\times(n-۱)$ است و با حذف jامین سطر و ستون ماتریس A به‌دست می‌آید.

تعریف ۳.۶.۱. ماتریس همسازه[۲۴]

[۲۲] Transpose Matrix
[۲۳] Determinant
[۲۴] Cofactor Matrix

ماتریس همسازه، ماتریسی است که درایه‌های آن به وسیله رابطه زیر به‌دست می‌آیند:

$$A_{ij} = (-۱)^{i+j}|M_{ij}|,$$

که در آن M_{ij} به معنی i و jامین کهاد[۲۵] ماتریس مورد نظر می‌باشد که از حذف سطر iام و ستون jام ماتریس اولیه به‌دست می‌آید. پس از محاسبه تک تک درایه‌ها با استفاده از فرمول فوق، ماتریس همسازه به‌دست می‌آید.

تعریف ۴.۶.۱. ماتریس الحاقی[۲۶]

ماتریس الحاقی از معکوس کردن ماتریس همسازه نسبت به قطر اصلی به‌دست می‌آید.

تعریف ۵.۶.۱. معکوس ماتریس[۲۷]

اگر A یک ماتریس $n \times n$ باشد، آنگاه

$$A'A^* = |A|I_n,$$

که در آن A^* ماتریس الحاقی، A' ترانهاده ماتریس A و $|A|$ دترمینان ماتریس A است. اگر $|A| \neq ۰$، یعنی ماتریس وارون‌پذیر باشد، آنگاه معکوس ماتریس A به صورت

$$A^{-۱} = \frac{۱}{|A|}(A^*)',$$

محاسبه می‌شود. پس برای یافتن وارون ماتریس A، ابتدا ماتریس الحاقی را می‌یابیم و سپس ترانهاده آن را بر دترمینان A تقسیم می‌کنیم.

تعریف ۶.۶.۱. ماتریس معین مثبت[۲۸]

اگر A یک ماتریس متقارن با ویژگی $X'AX > ۰$ ($X'AX \geq ۰$) به ازای هر $X \neq ۰$ باشد، آنگاه شکل درجه دوم $X'AX > ۰$ را معین مثبت (نیمه معین مثبت[۲۹]) و ماتریس A را ماتریس معین مثبت (نیمه معین مثبت) می‌نامند.

---

[۲۵] Minor
[۲۶] Adjoint Matrix
[۲۷] Inverse of a Matrix
[۲۸] Positive Definite Matrix
[۲۹] Positive Semi Matrix

تعریف ۷.۶.۱. ماتریس اطلاع فیشر

برای متغیر تصادفی X با تابع (چگالی) احتمال $f_\theta(x)$، ماتریس اطلاع فیشر را با $I_X(\theta)$ نشان می‌دهیم و ماتریسی نیمه معین مثبت است که $\theta \in \Theta \subseteq R^K$، برابر است با

$$I_X(\theta) = E[(\frac{\partial}{\partial\theta} ln f_\theta(x))^2].$$

شکل ماتریسی آن نیز به‌صورت

$$I_X(\theta) = -E\begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2}\ln f_\theta(x) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ln f_\theta(x) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_K}\ln f_\theta(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial\theta_K\partial\theta_1}\ln f_\theta(x) & \frac{\partial^2}{\partial\theta_K\partial\theta_2}\ln f_\theta(x) & \cdots & \frac{\partial^2}{\partial\theta_K^2}\ln f_\theta(x) \end{bmatrix},$$

می‌باشد.

تعریف ۸.۶.۱. تابع چندگاما [۳۰]

به‌طورکلی تابع چندگاما را به‌صورت $\psi^m(\cdot), m = 0, 1, \cdots$، نشان می‌دهیم، و به‌صورت زیر تعریف می‌شود:

$$\psi^m(x) = \frac{d^{m+1}}{dx^{m+1}} \log\Gamma(x), \quad x > 0.$$

تعریف ۹.۶.۱. شرایط نظم

فرض کنید X یک متغیر تصادفی از خانواده توزیع‌های احتمال $F = \{f_\theta, \theta \in \Theta\}$ است، در این صورت با برقراری شرایط زیر می‌گوییم شرایط نظم برقرار است.

۱. $\Theta \subseteq R$ یک فاصله باز از اعداد حقیقی باشد و تکیه‌گاه X، $S_x = \{x; f_\theta(x) > 0\}$ به θ بستگی نداشته باشد.

۲. $\frac{\partial}{\partial\theta}\log f_\theta(x)$ به ازای تمام مقادیر $\theta \in \Theta$ موجود و متناهی باشد.

[۳۰] Polygamma Function

۳. هرگاه $h(X)$ یک تابع حقیقی از x باشد، به‌طوری‌که $E_\theta|h(X)| < \infty$، آن‌گاه رابطه زیر برقرار می‌باشد:

$$\begin{aligned}\frac{\partial}{\partial\theta}E(h(X)) &= \frac{\partial}{\partial\theta}\int h(X)f_\theta(x)dx \\ &= \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)dx.\end{aligned}$$

۴.

$$۰ < E[(\frac{\partial}{\partial\theta}lnf_\theta(x))^۲] < \infty.$$

تعریف ۱.۱۰.۶. (واشنگتن و همکاران، ۲۰۱۱) **فاصله کوک**[۳۱]

فاصله کوک برای شناسایی نقاط دورافتاده و موثر مورد استفاده قرار می‌گیرد. با توجه به رگرسیون Y روی $(x_۱, \cdots, x_k)$ و با استفاده از مجموعه داده $(y_j, x_{۱j}, \cdots, x_{kj})$، $j = ۱, \cdots, n$ داریم

$$D_i = \frac{\sum_{j=۱}^{n}(\hat{y}_j - \hat{y}_j(i))^۲}{(k+۱)^۲ s^۲}, \quad i = ۱, \cdots, n,$$

که در آن s، ریشه میانگین توان‌های دوم خطا، $\hat{y}_j$، میانگین شرطی، $E(Y_j|x_{۱j}, \cdots, x_{jk})$، و $\hat{y}_j(i))$، میانگین شرطی پس از حذف مشاهده iام می‌باشد.

[۳۱] Cook's Distance

فصل ۲
مدل رگرسیونی بتا

در این فصل، یک مدل رگرسیونی مناسب را برای مدل‌بندی متغیر وابسته‌ای که به‌طور پیوسته در فاصله واحد استاندارد، یعنی $۰ < y < ۱$، اندازه‌گیری می‌شود، معرفی می‌کنیم. مدل بر اساس پذیره پیروی متغیر پاسخ از توزیع بتا، بنا نهاده شده است.

در ادامه ابتدا با ارایه شکل پارامتری مناسبی از توزیع چگالی بتا برای مدل رگرسیونی، به معرفی مدل می‌پردازیم. از آن‌جایی که توزیع بتا عضوی از خانواده نمایی است، روند استنباط برای مدل رگرسیونی بتا، مشابه مدل‌های خطی تعمیم‌یافته است. بنابراین، برازش مدل را با روش درست‌نمایی ماکسیمم انجام می‌دهیم. سپس توزیع مجانبی برآوردگرها را معرفی کرده و فاصله اطمینان و آزمون فرضیه‌ها را برای این مدل مورد بررسی قرار می‌دهیم. در پایان، با مطالعه شبیه‌سازی، ویژگی‌های مجانبی برآوردگرها را در کوچک نمونه ارزیابی می‌کنیم.

۱.۲ معرفی مدل

فرض کنید متغیر پاسخ y دارای توزیع بتا با پارامترهای p و q باشد. به‌طور معمول، مدل رگرسیونی بر حسب میانگین پاسخ فرمول‌بندی می‌شود. بنابراین، برای یک مدل رگرسیونی مناسب بر مبنای توزیع بتا، نیازمند صورت متفاوتی از تابع چگالی توزیع بتا هستیم، که بر حسب پارامترهای میانگین و

دقت[۱] تعریف شده باشد. برای این منظور، اگر تعریف کنیم $\mu = \frac{p}{p+q}$ و $\phi = p + q$، خواهیم داشت $q = (۱ - \mu)\phi$ و $p = \mu\phi$. پس می‌توان نوشت

$$E(Y) = \mu,$$

$$Var(Y) = \frac{\phi^۲\mu(۱-\mu)}{\phi^۲(\phi+۱)} = \frac{\mu(۱-\mu)}{\phi+۱} = \frac{V(\mu)}{۱+\phi}.$$

با توجه به آن‌که $V(\mu) = \mu(۱ - \mu)$، که در آن μ میانگین متغیر پاسخ است، برای μ ثابت هرگاه ϕ افزایش یابد، واریانس متغیر پاسخ کاهش می‌یابد. بنابراین می‌توان ϕ را به‌عنوان یک پارامتر ارزیابی دقت مدل تعبیر کرد. با تعریف پارامترهای جدید، تابع چگالی متغیر پاسخ به‌صورت زیر حاصل می‌شود:

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱}, \quad ۰ < y < ۱, \; ۰ < \mu < ۱, \; \phi > ۰. \tag{۱.۲}$$

با استفاده از این دو پارامتر، تابع چگالی می‌توانند به شکل‌های مختلفی نمایش داده شود. شکل ۱.۲ توابع چگالی متفاوتی از توزیع بتا با پارامترهای مختلف (μ, ϕ) را نمایش می‌دهد. این مقادیر در دامنه‌های $\phi \in (۵, ۱۰۰)$ و $\mu \in (۰/۰۰۲, ۰/۹۹)$ قرار دارند. به خصوص می‌توانند متقارن، $(\mu = \frac{۱}{۲})$ یا نامتقارن $(\mu \neq \frac{۱}{۲})$ باشند.

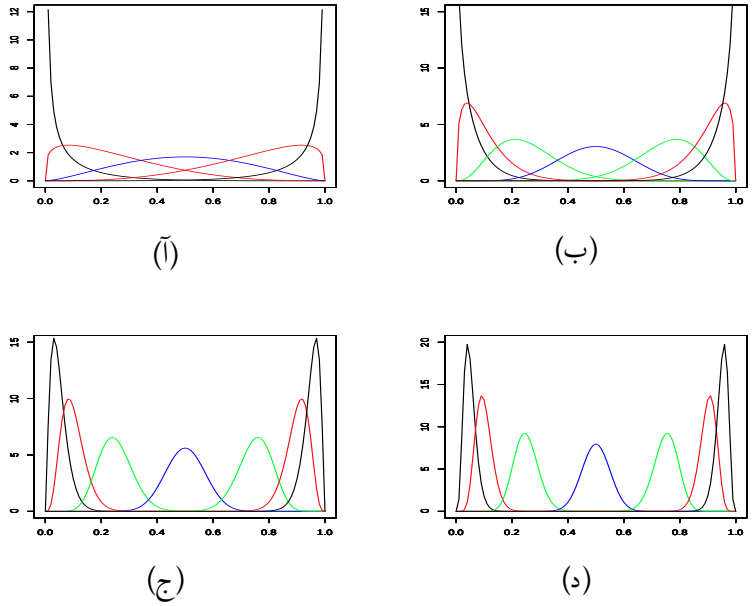
فرض کنید $y_۱, \cdots, y_n$ متغیرهای مستقل باشند، به‌طوری‌که تابع چگالی احتمال y_i، $i = ۱, \cdots, n$، در (۱.۲) با میانگین μ_i و پارامتر دقت نامعلوم ϕ نشان داده شده است. با در نظر گرفتن μ_i به‌عنوان پارامتر میانگین، پیشگوی خطی مدل به‌صورت زیر به میانگین پیوند داده می‌شود:

$$g(\mu_i) = \sum_{j=۱}^{k} x_{ij}\beta_j = \eta_i,$$

که در آن $\boldsymbol{\beta} = (\beta_۱, \cdots, \beta_k)' \; (\boldsymbol{\beta} \in \mathbb{R}^k)$ بردار پارامترهای مجهول رگرسیون است؛ $x_{i۱}, \cdots, x_{ik}$، مشاهدات روی k متغیر $(k < n)$ می‌باشند و $g(\cdot)$ تابع پیوند است.

برای تابع پیوند $g(\cdot)$ چندین انتخاب وجود دارد. به‌عنوان مثال، می‌توان از تابع پیوند لوجیت،

[۱] Precision



شکل ۱.۲: نمودارهای توابع چگالی بتا به ازای مقادیر مختلفی از ϕ و μ

$g(\mu) = \log\{\frac{\mu}{(۱-\mu)}\}$، پرابیت، $g(\mu) = \Phi^{-۱}(\mu)$، و لگاریتم مکمل، $g(\mu) = -\log\{-\log(\mu)\}$، استفاده کرد (مک کالا و نلدر، ۱۹۸۹).

۲.۲ برازش مدل

تابع درست‌نمایی برای یک نمونه به حجم n، به‌صورت

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=۱}^{n} \ell_i(\mu_i, \phi),$$

بیان می‌شود، که در آن

$$\ell_i(\mu_i, \phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((۱-\mu_i)\phi)$$
$$+ (\mu_i\phi - ۱)\log y_i + \{(۱-\mu_i)\phi - ۱\}\log(۱-y_i). \qquad (۲.۲)$$

برای برآورد پارامترها، باید توابع امتیاز [۲] متناظر با $\boldsymbol{\beta}$ و ϕ را به‌دست آوریم.

قضیه ۱.۲.۲. * توابع امتیاز متناظر با β و ϕ عبارتند از

$$\begin{aligned} U_{\beta}(\boldsymbol{\beta},\phi) &= \phi X' T(y^* - \mu^*), \\ U_{\phi}(\boldsymbol{\beta},\phi) &= \sum_{i=1}^{n}\{\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) + \psi((1-\mu_i)\phi) + \psi(\phi)\}. \end{aligned}$$

که در آن $y_i^* = \log\frac{y_i}{1-y_i}$ و $\mu_i^* = \psi(\mu_i\phi) - \psi((1-\mu_i)\phi)$، $T = diag(\frac{1}{g'(\mu_1)},\ldots,\frac{1}{g'(\mu_n)})$، $y^* = (y_1^*,\ldots,y_n^*)'$ و $\mu^* = (\mu_1^*,\ldots,\mu_n^*)'$.

برهان. با استفاده از مشتق زنجیره‌ای، برای محاسبه تابع امتیاز β_i، داریم

$$U_{\beta_r}(\boldsymbol{\beta},\phi) = \frac{\partial \ell(\boldsymbol{\beta},\phi)}{\partial \beta_r} = \sum_{i=1}^{n}\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r}. \tag{۳.۲}$$

بنابراین برای محاسبه (۳.۲)، باید مشتق‌های جزیی مجموع سمت راست آن را به ترتیب، محاسبه و جایگذاری کنیم. ابتدا $\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i}$ را به‌دست می‌آوریم. با توجه به رابطه (۲.۲)، داریم

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} = -\phi\frac{d}{d\mu_i}\log\Gamma(\mu_i\phi) + \phi\frac{d}{d\mu_i}\log\Gamma((1-\mu_i)\phi) + \phi\log y_i - \phi\log(1-y_i).$$

با تعریف $\frac{d}{dz}\log\Gamma(z) = \psi(z)$، نتیجه می‌شود

$$\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i} = \phi\left[\log\frac{y_i}{1-y_i} - \{\psi(\mu_i\phi) + \psi((1-\mu_i)\phi)\}\right]. \tag{۴.۲}$$

با توجه به تعریف μ_i^* و y_i^* در بالا، رابطه (۴.۲) به‌صورت $\phi(y^* - \mu^*)$ قابل بازنویسی است. بنابراین (۳.۲) را می‌توان به‌صورت

$$\frac{\partial \ell(\boldsymbol{\beta},\phi)}{\partial \beta_r} = \phi\sum_{i=1}^{n}(y_i^* - \mu_i^*)\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r}, \tag{۵.۲}$$

نوشت. با توجه به رابطه $\mu_i = g^{-1}(\eta_i)$، داریم

$$\frac{d\mu_i}{d\eta_i} = \frac{1}{\frac{d}{d\mu_i}g(\mu_i)}. \tag{۶.۲}$$

[۲] Score Function

از طرفی

$$\frac{\partial \eta_i}{\partial \beta_r} = x_{ir}. \tag{۷.۲}$$

در نهایت با جایگذاری روابط (۵.۲) و (۷.۲) در (۳.۲)، نتیجه می‌شود

$$U_{\beta_r}(\boldsymbol{\beta}, \phi) = \phi \sum_{i=1}^{n} \frac{(y_i^* - \mu_i^*)}{g'(\mu_i)} x_{ir}.$$

بنابراین

$$U_{\boldsymbol{\beta}} = \begin{pmatrix} U_{\beta_1} \\ \vdots \\ U_{\beta_k} \end{pmatrix} = \phi \begin{pmatrix} \sum_{i=1}^{n} \frac{(y_i^* - \mu_i^*)}{g'(\mu_i)} x_{i1} \\ \vdots \\ \sum_{i=1}^{n} \frac{(y_i^* - \mu_i^*)}{g'(\mu_i)} x_{ik} \end{pmatrix} = \phi X' T (y^* - \mu^*).$$

به‌طور مشابه برای محاسبه تابع امتیاز ϕ، داریم

$$U_\phi(\boldsymbol{\beta}, \phi) = \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \phi},$$

که با مشتق‌گیری از رابطه (۲.۲) خواهیم داشت

$$\begin{aligned}
\frac{\partial \ell_i(\mu_i, \phi)}{\partial \phi} &= \frac{d}{d\phi}\log\Gamma(\phi) - \mu_i \frac{d}{d\phi}\log\Gamma(\mu_i\phi) - (1-\mu_i)\frac{d}{d\phi}\log\Gamma((1-\mu_i)\phi) \\
&\quad + \mu_i \log y_i + (1-\mu_i)\log(1-y_i) \\
&= \frac{d}{d\phi}\log\Gamma(\phi) - \mu_i \frac{d}{d\phi}\log\Gamma(\mu_i\phi) - \frac{d}{d\phi}\log\Gamma((1-\mu_i)\phi) \\
&\quad + \mu_i \frac{d}{d\phi}\log\Gamma((1-\mu_i)\phi) + \mu_i \log y_i + \log(1-y_i) - \mu_i \log(1-y_i) \\
&= \mu_i \log \frac{y_i}{1-y_i} - \mu_i \psi(\mu_i\phi) - \psi((1-\mu_i)\phi) + \mu_i \psi((1-\mu_i)\phi) \\
&\quad + \log(1-y_i) + \log\Gamma(\phi) \\
&= \mu_i y_i^* - \mu_i \psi(\mu_i\phi) - \psi((1-\mu_i)\phi) + \mu_i \psi((1-\mu_i)\phi) \\
&\quad + \log(1-y_i) + \log\Gamma(\phi) \\
&= \mu_i y_i^* - \mu_i \mu_i^* - \psi((1-\mu_i)\phi) + \log(1-y_i) + \log\Gamma(\phi) \\
&= \mu_i(y_i^* - \mu_i^*) + \log(1-y_i) + \psi((1-\mu_i)\phi) + \psi(\phi),
\end{aligned}$$

بنابراین

$$U_\phi(\boldsymbol{\beta},\phi)=\sum_{i=1}^{n}\frac{\partial \ell_i(\mu_i,\phi)}{\partial\phi}=\sum_{i=1}^{n}\mu_i(y_i^*-\mu_i^*)+\log(1-y_i)+\psi((1-\mu_i)\phi)+\psi(\phi). \quad (۸.۲)$$

□

برای تکمیل فرآیند برازش مدل، نیازمند اطلاع از دقت برآوردگرهای به‌دست آمده هستیم. بنابراین باید ماتریس اطلاع فیشر برآوردگرها را به‌دست آوریم. معکوس این ماتریس، ماتریس کوواریانس برآوردگرهای مدل را نتیجه می‌دهد. قضیه زیر نحوه محاسبه ماتریس اطلاع فیشر برای مدل را بیان می‌کند.

قضیه ۲.۲.۲. * ماتریس اطلاع فیشر برای پارامترهای مدل (۱.۲) به‌صورت

$$K=K(\boldsymbol{\beta},\phi)=\begin{pmatrix}K_{\beta\beta} & K_{\beta\phi}\\ K_{\phi\beta} & K_{\phi\phi}\end{pmatrix}, \quad (۹.۲)$$

می‌باشد که در آن $K_{\beta\beta}=\phi X'WX$، $K_{\beta\phi}=K'_{\phi\beta}=X'Tc$ و $K_{\phi\phi}=tr(D)$ به‌طوری که

$$\begin{aligned}
c &= (c_1,\cdots,c_n)',\\
c &= -\phi\{\frac{d}{d\phi}\psi(\mu_i\phi)\mu_i-\frac{d}{d\phi}\psi((1-\mu_i)\phi)(1-\mu_i)\},\\
W &= diag\Big(w_1,\cdots,w_n\Big),\\
D &= diag\Big(d_1,\cdots,d_n\Big),\\
d_i &= \frac{d}{d\phi}\psi(\mu_i\phi)\mu_i^2+\frac{d}{d\phi}\psi((1-\mu_i)\phi)(1-\mu_i)^2-\frac{d}{d\phi}\psi(\phi),
\end{aligned}$$

و T در قضیه ۱.۲.۲ تعریف شده است.

برهان. با مشتق‌گیری از (۳.۲) نسبت به β، داریم

$$\begin{aligned}
\frac{\partial^2\ell(\boldsymbol{\beta},\phi)}{\partial\beta_r\partial\beta_s} &= \sum_{i=1}^{n}\frac{\partial}{\partial\mu_i}\left(\frac{\partial\ell_i(\mu_i,\phi)}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\right)\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_s}x_{ir}\\
&= \sum_{i=1}^{n}\left(\frac{\partial^2\ell(\beta,\phi)}{\partial\mu_i^2}\frac{d\mu_i}{d\eta_i}+\frac{\partial\ell(\beta,\phi)}{\partial\mu_i}\frac{\partial}{\partial\mu_i}\frac{d\mu_i}{d\eta_i}\right)\frac{d\mu_i}{d\eta_i}x_{ir}x_{is}.
\end{aligned}$$

از آن‌جایی که $E(\frac{\partial \ell_i(\mu_i,\phi)}{\partial \mu_i}) = ۰$،

$$E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial\beta_r \partial\beta_s}\right) = \sum_{i=۱}^{n} E\left(\frac{\partial^۲ \ell_i(\mu_i,\phi)}{\partial \mu_i^۲}\right)\left(\frac{d\mu_i}{d\eta_i}\right)^۲ x_{ir}x_{is}.$$

با مشتق‌گیری از رابطه (۴.۲) نسبت به μ_i، می‌توان نتیجه گرفت

$$\begin{aligned}\frac{\partial^۲ \ell_i(\mu_i,\phi)}{\partial \mu_i^۲} &= \frac{d}{d\mu_i}\left(\phi\left[\log\frac{y_i}{۱-y_i} - \psi(\mu\phi) + \psi((۱-\mu_i)\phi)\right]\right)\\ &= \phi\left(-\phi\frac{d}{d\mu_i}\psi(\mu_i\phi) - \phi\frac{d}{d\mu_i}\psi((۱-\mu_i)\phi)\right)\\ &= -\phi^۲\left(\frac{d}{d\mu_i}\psi(\mu_i\phi) + \frac{d}{d\mu_i}\psi((۱-\mu_i)\phi)\right).\end{aligned}$$

همچنین بنابر (۶.۲)، خواهیم داشت

$$(\frac{d\mu_i}{d\eta_i})^۲ = \frac{۱}{(g'(\mu_i))^۲}.$$

فرض کنید $W = diag\Big(w_۱,...,w_n\Big)$، که در آن $w_i = \phi\left\{\frac{d}{d\mu_i}\psi(\mu_i\phi) + \frac{d}{d\mu_i}\psi((۱-\mu_i)\phi)\right\}\frac{۱}{\{g'(\mu_i)\}^۲}$ و $\frac{d}{d\mu_i}\psi(.)$ تابع تری‌گاما[۳] (بای‌نی و فنگ، ۲۰۱۰) می‌باشد. بنابراین

$$E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial\beta_r \partial\beta_s}\right) = -\phi\sum_{i=۱}^{n} w_i x_{ir}x_{is}.$$

در نتیجه می‌توان نوشت

$$K_{\beta\beta} = -E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial\beta\partial\beta'}\right) = \phi X'WX.$$

به‌طور مشابه، برای $K_{\beta\phi}$ و $K_{\phi\beta}$ با مشتق‌گیری از رابطه (۴.۲)، نسبت به ϕ، داریم

$$\begin{aligned}\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial\beta_r\partial\phi} &= \frac{d}{d\phi}\left[\phi\sum_{i=۱}^{n}(y_i^* - \mu_i^*)\frac{۱}{g'(\mu_i)}x_{ir}\right]\\ &= \frac{d}{d\phi}\left[\phi\sum_{i=۱}^{n} y_i^*\frac{۱}{g'(\mu_i)}x_{ir} - \phi\sum_{i=۱}^{n}\mu_i^*\frac{۱}{g'(\mu_i)}x_{ir}\right]\\ &= \sum_{i=۱}^{n} y_i^*\frac{۱}{g'(\mu_i)}x_{ir} - \left[\sum_{i=۱}^{n}\mu_i^*\frac{۱}{g'(\mu_i)}x_{ir} + \phi\frac{\partial\mu_i^*}{\partial\phi}\right]\\ &= \sum_{i=۱}^{n}\left[(y_i^* - \mu_i^*) - \phi\frac{\partial\mu_i^*}{\partial\phi}\right]\frac{۱}{g'(\mu_r)}x_{ir}. \qquad (۱۰.۲)\end{aligned}$$

[۳] Trigamma Function

از طرفی می‌توان نوشت

$$\begin{aligned}\frac{\partial \mu_i^*}{\partial \phi} &= \frac{d}{d\phi}\left(\psi(\mu_i\phi) + \psi((۱-\mu_i)\phi)\right) \\ &= \mu_i \frac{d}{d\phi}\psi(\mu_i\phi) - (۱-\mu_i)\frac{d}{d\phi}\psi((۱-\mu_i)\phi). \end{aligned} \tag{۱۱.۲}$$

با جایگذاری (۱۱.۲) در (۱۰.۲)، نتیجه می‌شود

$$\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial \beta_r \partial \phi} = \sum_{i=۱}^{n}\left[(y_i^* - \mu_i^*) - \phi(\mu_i \frac{d}{d\phi}\psi(\mu_i\phi) - (۱-\mu_i)\frac{d}{d\phi}\psi((۱-\mu_i)\phi))\right]\frac{۱}{g'(\mu_i)}x_{ir}.$$

از آن‌جایی که $E(y_i^*) = \mu_i^*$، پس

$$E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial \beta_r \partial \phi}\right) = -\sum_{i=۱}^{n} c_i \frac{۱}{g'(\mu_i)}x_{ir}. \tag{۱۲.۲}$$

شکل ماتریسی رابطه (۱۲.۲) به‌صورت

$$E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial \beta \partial \phi}\right) = -X'Tc,$$

نوشته می‌شود. بنابراین

$$K_{\beta\phi} = -E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial \beta \partial \phi}\right) = X'Tc.$$

برای $K_{\phi\phi}$، با استفاده از رابطه (۸.۲)، می‌توان نوشت

$$\begin{aligned}\frac{\partial^۲ \ell(\boldsymbol{\beta},\phi)}{\partial \phi^۲} &= \frac{d}{d\phi}\left(\sum_{i=۱}^{n}\{\mu_i(y_i^* - \mu_i^*) + \log(۱-y_i) - \psi((۱-\mu_i)\phi) + \psi(\phi)\}\right) \\ &= \frac{d}{d\phi}\left(\sum_{i=۱}^{n}\Big\{\mu_i y_i^* - \mu_i\psi(\mu_i\phi) + \mu_i\psi((۱-\mu_i)\phi) + \log(۱-y_i)\right. \\ &\quad \left.-\psi((۱-\mu_i)\phi) + \psi(\phi)\Big\}\right) \\ &= \sum_{i=۱}^{n}\left\{-\mu_i^۲\frac{d}{d\phi}\psi(\mu_i\phi) + \mu_i(۱-\mu_i)\frac{d}{d\phi}\psi((۱-\mu_i)\phi)\right. \\ &\quad \left.-(۱-\mu_i)\frac{d}{d\phi}\psi((۱-\mu_i)\phi) + \frac{d}{d\phi}\psi(\phi)\right\} \\ &= \sum_{i=۱}^{n} -\left(\mu_i^۲\frac{d}{d\phi}\psi(\mu_i\phi) + (۱-\mu_i)^۲\frac{d}{d\phi}\psi((۱-\mu_i)\phi) - \frac{d}{d\phi}\psi(\phi)\right).\end{aligned}$$

لذا

$$\frac{\partial^۲ \ell(\boldsymbol{\beta}, \phi)}{\partial \phi^۲} = -\sum_{i=۱}^{n} d_i = -tr(D).$$

در نتیجه

$$K_{\phi\phi} = -E(\frac{\partial^۲ \ell(\boldsymbol{\beta}, \phi)}{\partial \phi^۲}) = tr(D).$$

□

۳.۲ توزیع مجانبی برآوردگرها

قبل از بیان توزیع مجانبی برآوردگرهای پارامترهای مدل، ابتدا باید لم زیر را مطرح کنیم.

لم ۱.۳.۲. (استرانگ، ۲۰۰۳) فرض کنید K به‌صورت زیر تجزیه شده باشد:

$$K = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

آن‌گاه معکوس ماتریس K به صورت

$$K^{-۱} = \begin{pmatrix} A^{-۱} + A^{-۱}B(D - CA^{-۱}B)^{-۱}CA^{-۱} & -A^{-۱}B(D - CA^{-۱}B)^{-۱} \\ -(D - CA^{-۱}B)^{-۱}CA^{-۱} & (D - CA^{-۱}B)^{-۱} \end{pmatrix},$$

قابل محاسبه می‌باشد.

قضیه ۱.۳.۲. * تحت شرایط نظم برای برآوردگرهای ماکسیمم درست‌نمایی (بیکل و داکسوم، ۲۰۰۱)، توزیع مجانبی، وقتی اندازه نمونه بزرگ باشد، عبارت است از

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+۱} \left(\begin{pmatrix} \boldsymbol{\beta}_{\circ} \\ \phi_{\circ} \end{pmatrix}, \quad K^{-۱} \right),$$

که در آن $\beta_{\circ}$ و $\phi_{\circ}$ مقادیر واقعی پارامترها هستند. همچنین معکوس ماتریس اطلاع فیشر (ماتریس کوواریانس مجانبی برآوردگرها) به‌صورت

$$K^{-۱} = K^{-۱}(\boldsymbol{\beta}, \phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix},$$

است، که در آن

$$K^{\beta\beta} = \frac{1}{\phi}(X'WX)^{-1}\left\{I_k + \frac{X'Tcc'T'X\,(X'WX)^{-1}}{\gamma\phi}\right\},$$
$$K^{\beta\phi} = (K^{\phi\beta})' = -\frac{1}{\gamma\phi}(X'WX)^{-1}X'Tc,$$
$$K^{\phi\phi} = \left[tr(D) - \phi^{-1}c'T'X(X'WX)^{-1}X'Tc\right]^{-1},$$
$$\gamma = tr(D) - \phi^{-1}c'T'X(X'WX)^{-1}X'Tc,$$

و I_k یک ماتریس همانی با بعد k است.

برهان. توزیع مجانبی نرمال برآوردگرهای $(\hat{\beta},\hat{\phi})$ بنابر لم (۲.۳.۱) در بیکل و داکسوم (۲۰۰۱) قابل استنتاج است. پس کافیست معکوس ماتریس اطلاع فیشر را محاسبه کنیم. برای به‌دست آوردن معکوس ماتریس اطلاع فیشر و بنابر لم ۱.۳.۲ می‌توان نوشت

$$K^{-1} = K^{-1}(\beta,\phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix} \qquad (۱۳.۲)$$

که برابر است با

$$\begin{pmatrix} K_{\beta\beta}^{-1} + K_{\beta\beta}^{-1}K_{\beta\phi}(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1}K_{\phi\beta}K_{\beta\beta}^{-1} & -K_{\beta\beta}^{-1}K_{\beta\phi}(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1} \\ -(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1}K_{\phi\beta}K_{\beta\beta}^{-1} & (K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1} \end{pmatrix}.$$

برای محاسبه $K^{\beta\beta}$، با توجه به روابط (۹.۲) و (۱۳.۲) داریم

$$\begin{aligned} K^{\beta\beta} &= K_{\beta\beta}^{-1} + K_{\beta\beta}^{-1}K_{\beta\phi}(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1}K_{\phi\beta}K_{\beta\beta}^{-1} \\ &= \frac{1}{\phi}(X'WX)^{-1} + \frac{1}{\phi}(X'WX)^{-1}X'Tc\underbrace{(tr(D) - c'T'X\frac{1}{\phi}(X'WX)^{-1}X'Tc)^{-1}}_{\gamma} \\ &\quad \times c'T'X\frac{1}{\phi}(X'WX)^{-1} \\ &= \frac{1}{\phi}(X'WX)^{-1}\left[I + \frac{X'Tcc'T'X(X'WX)^{-1}}{\gamma\phi}\right] \end{aligned}$$

به طور مشابه برای $K^{\phi\phi}$، خواهیم داشت

$$K^{\phi\phi} = (K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1}$$

$$= \left(tr(D) - \phi^{-1} c'T'X(X'WX)^{-1}X'Tc\right)^{-1}$$
$$= \gamma^{-1}$$

همچنین برای $K^{\phi\beta}$ و $K^{\beta\phi}$ داریم

$$\begin{aligned} K^{\beta\phi} &= -K_{\beta\beta}^{-1}K_{\beta\phi}(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1} \\ &= -\frac{1}{\phi}(X'WX)^{-1}X'Tc\underbrace{(tr(D) - c'T'X\frac{1}{\phi}(X'WX)^{-1}X'Tc)}_{\gamma}{}^{-1} \\ &= -\frac{1}{\phi\gamma}(X'WX)^{-1}X'Tc, \end{aligned}$$

$$\begin{aligned} K^{\phi\beta} &= -(K_{\phi\phi} - K_{\phi\beta}K_{\beta\beta}^{-1}K_{\beta\phi})^{-1}K_{\phi\beta}K_{\beta\beta}^{-1} \\ &= -\underbrace{(tr(D) - c'T'X\frac{1}{\phi}(X'WX)^{-1}X'Tc)}_{\gamma}{}^{-1}c'T'X\frac{1}{\phi}(X'WX)^{-1} \\ &= -\frac{1}{\phi\gamma}c'T'X(X'WX)^{-1}, \end{aligned}$$

که $K^{\beta\phi} = (K^{\phi\beta})'$ می‌باشد. □

توزیع‌های مجانبی، معمولا برای سایر استنباط‌های آماری، شامل فاصله اطمینان و آزمون فرضیه‌ها، استفاده می‌شوند.

۴.۲ فاصله اطمینان و آزمون فرضیه

مفاهیم برآورد فاصله‌ای و آزمون فرضیه‌ها ارتباط نزدیکی با هم دارند و هر یک از آن‌ها می‌تواند منجر به معرفی مفهوم دیگری شود. نظیر برآورد نقطه‌ای، برآورد فاصله‌ای نیز از دو بخش تشکیل می‌شود. ابتدا تعیین برآورد فاصله‌ای و به دنبال آن به‌دست آوردن فاصله‌ای خوب یا به عبارت دیگر تعیین برآورد فاصله‌ای بهینه. از جمله انواع برآوردهای فاصله‌ای، می‌توان به فاصله‌های اطمینان با دم‌های برابر، کوتاهترین، ناریب و مجانبی اشاره کرد.

علی‌رغم متفاوت بودن فاصله‌های اطمینان یادشده که از کمیت‌های محوری با توزیع‌های نامتقارن به‌دست می‌آیند، با افزایش اندازه نمونه، همه فاصله‌های اطمینان ذکرشده به یکدیگر نزدیک می‌شوند. از طرف دیگر، با توجه به قضیه حد مرکزی (کسلا و برگر، ۲۰۰۲)، می‌توان نشان داد که بسیاری از کمیت‌های محوری دارای توزیع تقریبی نرمال هستند. فواصل اطمینان به‌دست آمده با اندازه‌های بزرگ را فواصل اطمینان مجانبی می‌نامند. در این بخش، نرمال مجانبی بودن آماره‌های آزمون نسبت درست‌نمایی و والد را نشان می‌دهیم. همچنین بیان می‌کنیم که آزمون نسبت درست‌نمایی، نسبت به آزمون والد برتری دارد.

فرض کنید یک نمونه بزرگ از مشاهدات در اختیار داریم. می‌خواهیم فرضیه $H_0 : \beta_1 = \beta_1^{(0)}$ در مقابل $H_1 : \beta_1 \neq \beta_1^{(0)}$ را آزمون کنیم، که در آن $\beta_1 = (\beta_1, \ldots, \beta_m)'$ و $\beta_1^{(0)} = (\beta_1^{(0)}, \ldots, \beta_m^{(0)})$. همچنین $m < k$ که در آن k تعداد کل پارامترهای رگرسیونی است و $\beta_1^{(0)}$، مقداری معلوم است.

از فصل ۱، می‌دانیم لگاریتم آماره نسبت درست‌نمایی به صورت $w_1 = 2\{\ell(\hat{\beta}, \hat{\phi}) - \ell(\tilde{\beta}, \tilde{\phi})\}$ می‌باشد و $(\tilde{\beta}', \tilde{\phi})'$ نیز، بردار برآورد ماکسیمم درست‌نمایی محدودشده بردار پارامترهای $(\beta', \phi)'$ است، که تحت فرضیه صفر محاسبه شده است. از طرفی، تحت شرایط نظم، وقتی حجم نمونه زیاد باشد، آماره آزمون مورد نظر به توزیع کی‌دو با m درجه آزادی نزدیک می‌شود (مک کالاک و سیرل، ۲۰۰۱). بنابراین می‌توان برای معنی‌داری آماره آزمون نسبت درست‌نمایی تحت فرضیه صفر، آن را با توزیع χ^2_m مقایسه کرد.

وقتی پارامترهای مزاحم وجود داشته باشند، تابع درست‌نمایی، خواص مطلوب خود را نخواهد داشت. به عبارت دیگر، میانگین تابع امتیاز صفر نیست و برابری اطلاعات برقرار نمی‌باشد (سریباری و مارسلو، ۲۰۱۲). آزمون فرضیه‌ها، بر مبنای ویژگی‌های تابع درست‌نمایی، می‌توانند از ریخت‌افتادگی‌های حجمی[۴] مورد نظر در نمونه‌های کوچک را نشان دهند، به ویژه وقتی چندین پارامتر مزاحم وجود داشته باشد. هنگامی‌که پارامتر مورد علاقه، بعد برابر ۱ داشته باشد، آزمون فرضیه صفر را می‌توان با استفاده از آماره

[۴] Display Considerable Size Distortions

درست‌نمایی علامت‌دار، آزمون کرد (سریباری مارسلو، ۲۰۱۲):

$$R = sign(\beta - \hat{\beta})\{۲(\ell(\tilde{\beta}, \tilde{\phi}) - \ell(\hat{\beta}, \hat{\phi}))\}^{\frac{۱}{۲}}.$$

توزیع آماره آزمون فوق را می‌توان تحت فرضیه صفر با توزیع نرمال استاندارد تقریب زد (سریباری و مارسلو، ۲۰۱۲). با این حال، این تقریب برای نمونه‌های کوچک ممکن است ضعیف باشد. بارندرف نیلسن (۱۹۸۳) آماره نسبت درست‌نمایی علامت‌دار اصلاح‌شده زیر را معرفی کرد:

$$R^* = R + R^{-۱} \log(\frac{u}{R}), \qquad (۱۴.۲)$$

که در آن u شامل مشتق‌های فضای نمونه می‌باشد. این آماره آزمون نیز دارای توزیع تقریبی نرمال با دقت بیشتری نسبت به آماره نسبت درست‌نمایی علامت‌دار می‌باشد. باید اشاره کرد که در اغلب اوقات به‌دست آوردن این آماره کار بسیار دشواری است، زیرا باید مشتق‌های فضای نمونه را محاسبه کرد. در همین راستا اسکوگارد (۲۰۰۱) یک تقریب برای آماره (۱۴.۲) به‌دست آورد که نتایج او بعدها برای مواقعی که پارامتر مورد نظر، برداری باشد نیز، تعمیم پیدا کرد (سریباری و مارسلو، ۲۰۱۲).

با استفاده از آزمون والد نیز می‌توان استنباط مجانبی را انجام داد. آماره آزمون والد برای آزمون $H_۰ : \beta_۱ = \beta_۱^{(۰)}$ به صورت

$$w_۲ = \left(\hat{\beta}_۱ - \beta_۱^{(۰)}\right)' \left(\hat{K}_{۱۱}^{\beta\beta}\right)^{-۱} \left(\hat{\beta}_۱ - \beta_۱^{(۰)}\right),$$

است، که در آن $\hat{K}_{۱۱}^{\beta\beta}$ ماتریس $m \times m$ اطلاع فیشر برای برآوردگر $\hat{\beta}_۱$ است که در $\hat{\beta}_۱$ مقداردهی شده‌است و $\hat{\beta}_۱$ برآوردگر درست‌نمایی $\beta_۱$ است. برای این آزمون نیز، تحت شرایط نظم، وقتی حجم نمونه زیاد باشد، آماره آزمون مورد نظر به توزیع کی‌دو با m درجه آزادی نزدیک می‌شود.

آزمون‌های والد و نسبت درست‌نمایی، اغلب منجر به نتایج مشابه می‌شوند. ولی، با این حال، در بعضی موارد اختلاف محسوسی دارند که در نتیجه‌گیری تاثیرگذار است (فراری و سریباری، ۲۰۰۴).

به چند دلیل، آزمون نسبت درست‌نمایی را به آزمون والد ترجیح می‌دهند: اول، این امکان وجود دارد که آزمون والد به یک سوال، بسته به نوع سوال، چند پاسخ مختلف بدهد یا به‌عبارتی، آماره آزمون والد

نسبت به تبدیل‌ها پایا نیست. به‌عنوان مثال، $R = ۱$ معادل است با $\log R = ۰$، اما آماره والد به‌دست آمده برای $R = ۱$، همان آماره به‌دست آمده برای $\log R = ۰$ نیست. ولی آزمون نسبت درست‌نمایی دقیقا جواب‌های یکسانی برای $R = ۱$ و $\log R = ۰$ یا هر تبدیلی از آن به ما می‌دهد؛ دوم آزمون والد با استفاده از دو تقریب (خطای استاندارد و توزیع کی‌دو) عمل می‌کند، ولی آزمون نسبت درست‌نمایی با استفاده از یک تقریب (توزیع کی‌دو).

آماره آزمون دیگری نیز وجود دارد که می‌تواند جایگزین خوبی برای آزمون والد و آزمون نسبت درست‌نمایی باشد، که به آماره آزمون امتیاز (رائو) معروف است (سریباری و مارسلو، ۲۰۱۲).

با توجه به آن‌که توزیع مجانبی آماره‌های آزمون ذکرشده نرمال است، فاصله اطمینان تقریبی $۱۰۰(۱-\alpha)$٪ برای $\boldsymbol{\beta}_i$، $i = ۱, \ldots, k$، به صورت

$$\hat{\beta}_i \mp \Phi^{-۱}(۱ - \frac{\alpha}{۲}) se(\hat{\beta}_i),$$

به‌دست می‌آید. به‌طور مشابه، یک فاصله اطمینان مجانبی $۱۰۰(۱-\alpha)$٪ برای پارامتر دقت ϕ به صورت

$$\hat{\phi} \mp \Phi^{-۱}(۱ - \frac{\alpha}{۲}) se(\hat{\phi}),$$

حاصل می‌شود. یک فاصله اطمینان تقریبی $۱۰۰(۱-\alpha)$٪ برای μ را نیز می‌توان به‌صورت (L, U) نوشت که در آن

$$L = g^{-۱}(\hat{\eta} - \Phi^{-۱}((۱ - \frac{\alpha}{۲}) se(\hat{\eta})), \qquad U = g^{-۱}(\hat{\eta} + \Phi^{-۱}((۱ - \frac{\alpha}{۲}) se(\hat{\eta})),$$

به‌طوری که $\hat{\eta} = x'\hat{\boldsymbol{\beta}}$ و $se(\hat{\eta}) = \sqrt{x'\widehat{cov}(\hat{\boldsymbol{\beta}})x}$.

شواهد عددی ارایه‌شده توسط نویسندگان مختلف، نشان می‌دهد، معمولا، استنباط‌های مبتنی بر آماره‌های آزمون اصلاح‌شده در مقایسه با آماره‌های آزمون نسبت درست‌نمایی، در حجم‌های نمونه کوچک، دقیق‌تر هستند. به منظور بهبود دقت در استنباط‌های حاصل برای نمونه‌های کوچک، می‌توان از روش‌های

بازنمونه‌گیری[۵] (ریزو، ۲۰۰۲) مانند بوت‌استرپ[۶] (افرون، ۱۹۷۹) استفاده کرد.

۵.۲ تحلیل باقی‌مانده‌ها برای مدل رگرسیونی بتا

در این بخش، به تحلیل باقی‌مانده‌ها می‌پردازیم و انواع باقی‌مانده‌ها را معرفی می‌کنیم.

انواع مختلفی از باقی‌مانده‌ها، برای مدل رگرسیونی بتا در دسترس هستند. هدف از تحلیل باقی‌مانده‌ها، ارزیابی نیکویی برازش مدل و شناسایی مشاهدات بی‌قاعده است. یکی از انواع باقی‌مانده‌های مورد استفاده، باقی‌مانده‌های پیرسونی است که فراری و سریباری (۲۰۰۴) آن را باقی‌مانده‌های معمولی استانداردشده نامیدند، و به صورت زیر تعریف می‌شوند:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}}, \qquad i = ۱, \cdots, n, \tag{۱۵.۲}$$

که در آن $\widehat{var}(y_i) = \frac{\widehat{\mu}_i(۱-\widehat{\mu}_i)}{۱+\widehat{\phi}}$، $\widehat{\mu}_i = g^{-۱}(x_i'\widehat{\beta})$ و $\hat{\beta}$ و $\hat{\phi}$ به‌ترتیب برآوردهای ماکسیمم درست‌نمایی پارامترهای β و ϕ می‌باشند.

برای مدل رگرسیونی بتا، مشابه باقی‌مانده‌های پیرسونی (۱۵.۲) می‌توان باقی‌مانده‌های انس‌کمب و انحراف را که استفاده از آن‌ها در مدل‌های خطی تعمیم‌یافته معمول هستند، تعریف کرد.

باقی‌مانده‌های جدید دیگری برای این مدل رگرسیونی معرفی شده‌اند. اسپینهیرا و همکاران (۲۰۰۸) باقی‌مانده‌های r_i^* را پیشنهاد کردند که به‌صورت زیر تعریف می‌شوند:

$$r_i^* = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\phi \upsilon_i}}, \qquad i =, \cdots, n,$$

که در آن $\upsilon_i = \{\frac{d\psi}{d\mu_i}(\mu_i\phi) + \frac{d\psi}{d\mu_i}((۱-\mu_i)\phi)\}$. همچنین با توجه به این‌که

$$var(y_i^*) = \frac{d\psi}{d\mu_i}(\mu_t\phi) + \frac{d\psi}{d\mu_i}((۱-\mu_i)\phi) = \upsilon_i,$$

باقی‌مانده‌های موزون جدید استانداردشده نیز به صورت

$$r_i^w = \phi^{\frac{۱}{۲}} r_i^* = \frac{y_i^* - \hat{\mu}_i^*}{\sqrt{\upsilon_i}},$$

[۵] Resampling Methods
[۶] Bootstrap

تعریف می‌شوند (اسپینهیرا و همکاران، ۲۰۰۸). همه انواع این باقی‌مانده‌ها در بسته **betareg** (سریباری و زیلیز، ۲۰۱۰) در نرم‌افزار R قابل محاسبه هستند.

۶.۲ مطالعه شبیه‌سازی

در این بخش، ویژگی‌های مجانبی برآوردگرهای پارامترهای مدل رگرسیونی بتا را که در این فصل، به‌طور نظری، بیان کردیم، برای حجم‌های نمونه کوچک و متوسط مورد ارزیابی قرار می‌دهیم. شبیه‌سازی‌ها در محیط نرم افزار R و با کمک بسته **betareg** اجرا شده‌اند. لازم به ذکر است که این بسته از تابع $optim()$ برای حصول برآوردها استفاده می‌کند و معمولا از روش عددی $BFGS$ (نکدال و رایت، ۱۹۹۹)، که یکی از روش‌های شبه نیوتن (بیرد و همکاران، ۱۹۹۵) است، بهره می‌برد. نحوه برازش یک مدل رگرسیونی بتا با استفاده از بسته **betareg** را در فصل ۴ به‌طور کامل تشریح خواهیم کرد. ویژگی‌های مجانبی برآوردگرها را برای حالتی که مقدار واقعی پارامتر دقت $\phi = ۱/۵$ است، با طراحی دو مثال شبیه‌سازی، بررسی می‌کنیم.

۱.۶.۲ مثال اول

برای مثال اول شبیه‌سازی، در مدل (۱.۲)، فرض کردیم پیشگوی خطی به‌صورت

$$logit(\mu_i) = \eta_i = \beta_۰ + \beta_۱ x_{۱i}, \qquad i = ۱, \cdots, n, \tag{۱۶.۲}$$

باشد، که در آن مقادیر واقعی $\beta_۰ = ۱$ و $\beta_۱ = ۲$ انتخاب شدند. مقادیر متغیر تبیینی $x_۱$ را از توزیع $N(۰, ۱)$ تولید کردیم. همچنین برای بررسی ویژگی‌های مجانبی برآوردگرها، پنج حجم نمونه $n = ۳۰, ۵۰, ۱۰۰, ۲۰۰, ۴۰۰$ را در نظر گرفتیم. نتایج را بر اساس ۱۰۰۰ مجموعه داده شبیه‌سازی شده از پیشگوی خطی (۱۶.۲) به‌دست آوردیم. جدول ۱.۲ نتایج را نشان می‌دهد. برآورد نهایی هر پارامتر، با میانگین‌گیری از برآوردهای حاصله در ۱۰۰۰ مجموعه مذکور به‌دست آمده است.

با توجه به نتایج جدول، واضح است که با افزایش حجم نمونه، مقدار MSE برآوردها در حال کاهش

جدول ۱.۲: نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱.۲) با پیشگوی خطی (۱۶.۲)

حجم نمونه	پارامترها	مقدار واقعی	برآورد	MSE	اریبی
۳۰	β_0	۱	۰/۹۱۸	۰/۰۶۷	−۰/۰۸۱
	β_1	۲	۱/۸۱۰	۰/۱۳۳	−۰/۱۸۹
	ϕ	۱/۵	۱/۶۰۰	۰/۲۷۰	۰/۱۰۰
۵۰	β_0	۱	۰/۹۲۴	۰/۰۳۳	−۰/۰۷۵
	β_1	۲	۱/۸۸۸	۰/۰۶۰	−۰/۱۱۱
	ϕ	۱/۵	۱/۶۲۷	۰/۱۴۹	۰/۱۲۷
۱۰۰	β_0	۱	۰/۸۹۲	۰/۰۲۷	−۰/۱۰۷
	β_1	۲	۱/۸۳۹	۰/۰۵۱	−۰/۱۶۰
	ϕ	۱/۵	۱/۴۸۹	۰/۰۵۰	−۰/۰۱۰
۲۰۰	β_0	۱	۰/۸۹۳	۰/۰۱۸	−۰/۱۰۶
	β_1	۲	۱/۸۴۲	۰/۰۳۷	−۰/۱۵۷
	ϕ	۱/۵	۱/۴۸۳	۰/۰۲۶	−۰/۰۱۶
۴۰۰	β_0	۱	۰/۸۸۹	۰/۰۱۶	−۰/۱۱۰
	β_1	۲	۱/۸۴۰	۰/۰۳۱	−۰/۱۵۹
	ϕ	۱/۵	۱/۴۶۰	۰/۰۱۳	−۰/۰۳۹

است. از طرفی، مقادیر اریبی برآوردگرها ناچیز هستند. بنابراین، می‌توان سازگاری برآوردگرها را در حالت حجم‌های نمونه کوچک و متوسط نیز نتیجه گرفت.

۲.۶.۲ مثال دوم

برای مثال دوم، در مدل (۱.۲)، پیشگوی خطی را به‌صورت

$$logit(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \qquad (۱۷.۲)$$

در نظر گرفتیم، که در آن مقادیر واقعی $\beta_0 = ۲$، $\beta_1 = ۱$ و $\beta_2 = ۰/۵$ انتخاب شدند. مقادیر متغیر تبیینی x_1 را از توزیع $N(۰, ۱)$ و مقادیر x_2 را از توزیع برنولی با احتمال موفقیت ۰/۶ تولید کردیم. همچنین برای بررسی ویژگی‌های مجانبی برآوردگرهای ضرایب رگرسیونی در کوچک نمونه، پنج

حجم نمونه $n = ۳۰, ۵۰, ۱۰۰, ۲۰۰, ۴۰۰$ را در نظر گرفتیم. نتایج را بر اساس ۱۰۰۰ مجموعه داده شبیه‌سازی‌شده به‌دست آوردیم. جدول ۲.۲ نتایج را نشان می‌دهد. روند تغییرات مقادیر اریبی و MSE برآوردگرها، مشابه مثال اول هستند و در این‌جا نیز می‌توان سازگاری برآوردگرها را نتیجه گرفت.

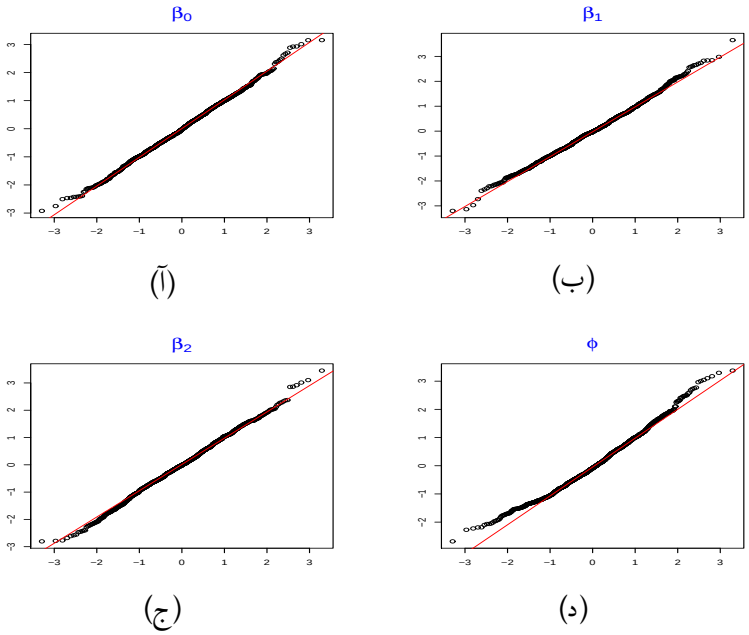
جدول ۲.۲: خلاصه نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱.۲) با پیشگوی خطی (۱۷.۲)

حجم نمونه	پارامترها	مقدار واقعی	برآورد	MSE	اریبی
۳۰	β_0	۲	۱/۹۲۵	۰/۰۵۳	−۰/۰۷۴
	β_1	۱	۰/۷۹۷	۰/۰۵۴	−۰/۲۰۲
	β_2	۰/۵	۰/۴۰۳	۰/۰۵۸	−۰/۰۹۶
	ϕ	۱/۵	۱/۵۸۹	۰/۱۰۱	۰/۰۸۹
۵۰	β_0	۲	۱/۹۰۷	۰/۰۴۳	−۰/۰۹۲
	β_1	۱	۰/۸۰۸	۰/۰۴۷	−۰/۱۹۱
	β_2	۰/۵	۰/۴۱۷	۰/۰۴۱	−۰/۰۸۲
	ϕ	۱/۵	۱/۵۵۸	۰/۰۶۸	۰/۰۵۸
۱۰۰	β_0	۲	۱/۹۱۶	۰/۰۳۱	−۰/۰۸۳
	β_1	۱	۰/۷۹۸	۰/۰۴۸	−۰/۲۰۱
	β_2	۰/۵	۰/۳۹۷	۰/۰۳۴	−۰/۱۰۲
	ϕ	۱/۵	۱/۵۴۴	۰/۰۴۸	۰/۰۴۴
۲۰۰	β_0	۲	۱/۹۰۵	۰/۰۲۴	−۰/۰۹۴
	β_1	۱	۰/۷۹۴	۰/۰۴۶	−۰/۲۰۵
	β_2	۰/۵	۰/۴۰۳	۰/۰۲۴	−۰/۰۹۶
	ϕ	۱/۵	۱/۵۳۲	۰/۹۵۴	۰/۰۲۲
۴۰۰	β_0	۲	۱/۹۰۱	۰/۰۲۱	−۰/۰۹۸
	β_1	۱	۰/۷۹۷	۰/۰۴۴	−۰/۲۰۲
	β_2	۰/۵	۰/۴۰۵	۰/۰۱۹	−۰/۰۹۴
	ϕ	۱/۵	۱/۵۱۶	۰/۰۲۰	۰/۰۱۶

با توجه به قضیه ۱.۳.۲، توزیع مجانبی برآوردگرها، نرمال است. برای بررسی این ویژگی در کوچک نمونه، از نمودارهای چندک-چندک [۷] استفاده کردیم. شکل ۲.۲، نمودارهای چندک-چندک پارامترهای برآوردشده در ۱۰۰۰ مجموعه داده شبیه‌سازی‌شده را برای پیشگوی خطی (۱۷.۲)، نمایش می‌دهد. در

[۷] Quantile-Quantile Plot

این شکل، نمودارهای چندک-چندک برای ضرایب رگرسیونی β_0، β_1، β_2 و ϕ، به ترتیب در نمودارهای (آ)، (ب)، (ج) و (د) برای $n = 400$ و ۱۰۰۰ مجموعه داده شبیه‌سازی‌شده، گزارش شده‌اند. با توجه به این نمودارها، می‌توان نرمال مجانبی بودن برآوردگرها را نتیجه گرفت، زیرا توزیع نرمال تقریب کاملا مناسبی برای توزیع برآوردگرها می‌باشد. با توجه به نتایج به‌دست آمده، می‌توان عنوان کرد که روش برازش پیشنهادی قابل اعتماد است.



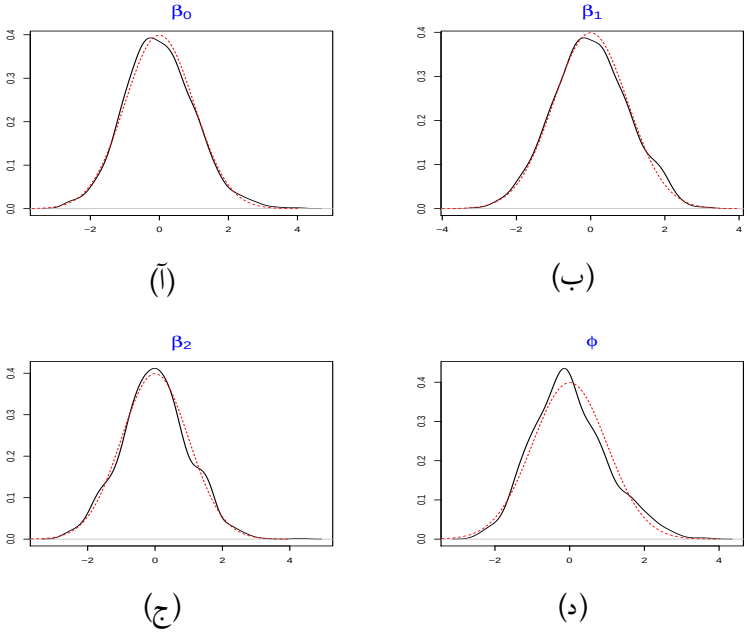
(آ) (ب)

(ج) (د)

شکل ۲.۲: نمودارهای چندک-چندک برآوردهای پارامترهای مدل (۱.۲)، با پیشگوی خطی (۱۷.۲) بر اساس ۱۰۰۰ مجموعه داده شبیه‌سازی‌شده

شکل ۳.۲، نمودارهای چگالی برآوردهای استانداردشده (منحنی توپر) را به همراه چگالی نرمال استاندارد (منحنی نقطه چین)، نشان می‌دهد. در این شکل منحنی‌های (آ)، (ب)، (ج) و (د) به ترتیب مربوط به برآوردگرهای پارامترهای β_0، β_1، β_2 و ϕ می‌باشند. این منحنی‌ها نیز تایید دیگری بر مناسب بودن توزیع نرمال به‌عنوان تقریب توزیع برآوردگرها است.

دقت داشته باشید که با توجه به شکل‌های ۲.۲ (د) و ۳.۲ (د)، برای همگرا شدن توزیع $\hat{\phi}$ به نرمال،



شکل ۳.۲: نمودارهای چگالی برآوردهای پارامترهای استانداردشده مدل (۱۷.۲) (منحنی توپر) به همراه چگالی نرمال استاندارد (نقطه چین)

به حجم نمونه بالاتری نیاز است. اما برای ضرایب رگرسیونی، در نمونه‌های کوچک نیز قابل قبول است.

در تکمیل مطالعه شبیه‌سازی، بحث انتخاب بهترین مدل (آزمون فرضیه پارامترهای رگرسیونی) را نیز دنبال کردیم. برای بررسی عملکرد انتخاب مدل درست، از مدل رگرسیونی بتا با پیشگوی خطی (۱۶.۲) یک مجموعه داده به حجم ۵۰ تولید کردیم و سپس بر روی این داده‌های تولیدشده دو مدل (۱۶.۲) و (۱۷.۲) را برازش دادیم. نتایج این دو مدل برازش‌شده در جدول ۳.۲ گزارش شده‌اند. برای برازش مدل (۱۷.۲)، مقادیر x_2 را جداگانه از توزیع برنولی با احتمال موفقیت ۰/۷ تولید کردیم.

با توجه به نتایج، واضح است که روش برآورد پیشنهادی و آماره‌های آزمون مجانبی که در بخش‌های قبلی مطرح شدند، در تشخیص مدل واقعی خوب عمل می‌کنند. برای انتخاب مدل، از معیار انتخاب مدل AIC نیز استفاده کردیم. مقادیر AIC برای دو مدل (۱۶.۲) و (۱۷.۲) در جدول ۴.۲ گزارش شده‌اند. با توجه به این مقادیر، مشاهده می‌کنید که این معیار نیز مدل واقعی (۱۶.۲) را انتخاب می‌کند.

جدول ۳.۲: نتایج برازش بر روی مجموعه داده شبیه‌سازی‌شده از مدل (۱۶.۲)

	پارامتر	برآورد	خطای استاندارد	آماره والد	p-مقدار
درست	β_0	۱٫۰۳۳	۰٫۰۹۱	۱۱٫۲۶	$< ۰٫۰۰۱$
	β_1	−۱٫۸۵۲	۰٫۱۱۶	−۱۵٫۸۸	$< ۰٫۰۰۱$
	ϕ	۱٫۶۴۸	۰٫۱۷۸	۹٫۲۴۱	$< ۰٫۰۰۱$
نادرست	β_0	۱٫۰۰۳	۰٫۲۷۶	۳٫۶۳۴	$< ۰٫۰۰۱$
	β_1	−۱٫۹۶۵	۰٫۲۲۷	−۸٫۶۵۳	$< ۰٫۰۰۱$
	β_2	−۰٫۰۴۹	۰٫۳۳۵	−۰٫۱۴۷	۰٫۲۰۶
	ϕ	۱٫۵۲۳	۰٫۳۳۵	−۰٫۱۴۷	$< ۰٫۰۰۱$

جدول ۴.۲: مقایسه AIC برای دو مدل (۱۶.۲) و (۱۷.۲)

	df	AIC
مدل (۱۶.۲)	۳	−۱۵۲۵٫۳۱۱
مدل (۱۷.۲)	۴	−۱۰۹۷٫۴۳۷

فصل ۳
مدل رگرسیونی بتا با پارامتر دقت متغیر

در فصل قبل مدل رگرسیونی بتا را برای حالتی که پارامتر دقت، ϕ، ثابت بود، مورد بررسی قرار دادیم. در این فصل، مدل را برای حالتی که پارامتر ϕ متغیر و خود تابعی از متغیرهای تبیینی باشد، تعمیم می‌دهیم. همچنین به برازش مدل با استفاده از روش درست‌نمایی ماکسیمم می‌پردازیم. در ادامه با مطالعه شبیه‌سازی، ویژگی‌های مجانبی برآوردگرها را برای این مدل تعمیم‌یافته، مورد ارزیابی قرار می‌دهیم.

۱.۳ معرفی مدل

همان‌طور که در فصل قبل بیان کردیم، متغیر پاسخ y دارای توزیع بتا به‌صورت زیر است:

$$f(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱}. \qquad (۱.۳)$$

اکنون فرض می‌کنیم پارامتر دقت ϕ نیز تابعی از مشاهدات است، به‌طوری که متغیرهای تبیینی $(z_۱, \cdots, z_n)$ بر روی آن تاثیر دارند. فرض کنید میانگین و پارامتر دقت برای y_i به‌صورت زیر مدل‌بندی شوند:

$$g_۱(\mu_i) = \eta_{۱i} = x_i'\beta, \qquad g_۲(\phi_i) = \eta_{۲i} = z_i'\theta, \qquad (۲.۳)$$

که در آن‌ها $\beta = (\beta_۱, \cdots, \beta_k)'$ و $\theta = (\theta_۱, \cdots, \theta_h)'$، به ترتیب بردارهای پارامترهای نامعلوم برای توابع میانگین و دقت هستند، که $\beta \in \mathbb{R}^k$، $\theta \in \mathbb{R}^h$ و $k + h < n$. همچنین توابع $\eta_{۱i}$ و $\eta_{۲i}$،

پیشگوهای خطی هستند و $x_{i1}, \cdots, x_{ik}$ و $z_{i1}, \cdots, z_{ih}$ متغیرهای تبیینی معلوم می‌باشند. البته برخی یا تمام متغیرهای x_i و z_i می‌توانند یکسان باشند. فرض می‌کنیم ماتریس‌های X و Z، ماتریس‌های طرح متناظر با متغیرهای تبیینی $x_1, \cdots, x_k$ و $z_1, \cdots, z_h$ باشند، که به‌صورت

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \qquad Z = \begin{bmatrix} z_{11} & \cdots & z_{1h} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nh} \end{bmatrix},$$

هستند. علاوه بر این، فرض می‌کنیم توابع پیوند $g_1 : (0,1) \to \mathbb{R}$ و $g_2 : (0,\infty) \to \mathbb{R}$، یکنوای اکید و دو بار مشتق‌پذیر هستند. لازم به ذکر است که برای در نظر گرفتن پارامتر عرض از مبدا، معمولا ستون‌های اول این دو ماتریس، بردار nبعدی $(1, \cdots, 1) = 1$ هستند.

توابع پیوند قابل استفاده برای $g_1(\cdot)$، همانند توابع پیوند معرفی‌شده در فصل ۲ هستند. برای $g_2(\cdot)$، نیز چند تابع پیوند مختلف را می‌توان معرفی کرد. مهم‌ترین آن‌ها، توابع پیوند لگاریتمی، $g_2(\phi) = \log \phi$، و ریشه دوم، $g_2(\phi) = \sqrt{\phi}$، هستند. در این فصل، شبیه‌سازی‌ها را بر اساس تابع پیوند لگاریتمی برای پارامتر دقت انجام می‌دهیم.

۲.۳ برازش مدل

تابع لگاریتم درست‌نمایی برای یک نمونه به حجم n، برای این مدل رگرسیون بتا به‌صورت

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi_i), \tag{۳.۳}$$

می‌باشد، که در آن

$$\begin{aligned} \ell_i(\mu_i, \phi_i) &= \log \Gamma(\phi_i) - \log \Gamma(\mu_i \phi_i) - \log \Gamma((1-\mu_i)\phi_i) \\ &\quad + (\mu_i \phi_i - 1) \log y_i + \{(1-\mu_i)\phi_i - 1\} \log(1 - y_i). \end{aligned} \tag{۴.۳}$$

توابع $\mu_i = g_1^{-1}(\eta_{1i})$ و $\phi_i = g_2^{-1}(\eta_{2i})$ که در رابطه (۲.۳) تعریف شده‌اند، به ترتیب توابعی از β و θ می‌باشند. برای برآورد پارامترها، مشابه فصل ۲، باید توابع امتیاز متناظر با β و θ را به‌دست آوریم.

قضیه ۱.۲.۳. * بردار پارامتر $\xi = (\beta', \theta')'$ را در نظر بگیرید. بردار امتیاز $(k+h)$بعدی $U(\xi)$ را به‌صورت $(U_\beta(\beta,\theta)', U_\theta(\beta,\theta)')'$ می‌نویسیم، که توابع امتیاز متناظر با β و θ عبارتند از

$$U_\beta(\beta,\theta) = \mathbf{X}'\boldsymbol{\phi} T_1 (y^* - \mu^*),$$

$$U_\theta(\beta,\theta) = \mathbf{Z}' T_2 v,$$

که در آن‌ها $T_1 = diag(\frac{d\mu_i}{d\eta_{1i}})$، $T_2 = diag(\frac{d\phi_i}{d\eta_{2i}})$. همچنین $v = (v_1, \cdots, v_n)'$، به‌طوری‌که

$$v_i = \mu_i(y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1-\mu_i)\phi) + \log(1-y_i),$$

و y^* و μ^* مشابه کمیت‌های تعریف‌شده در فصل ۲ هستند.

برهان. با استفاده از مشتق زنجیره‌ای، برای محاسبه تابع امتیاز β_r، با مشتق‌گیری از تابع لگاریتم درست‌نمایی (۳.۳)، برای $r = 1, \cdots, k$، داریم

$$U_r(\beta,\theta) = \frac{\partial \ell(\beta,\theta)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i} \frac{d\mu_i}{d\eta_{1i}} \frac{d\eta_{1i}}{\partial \beta_r}. \tag{۵.۳}$$

بنابراین برای محاسبه (۵.۳)، باید مشتق‌های جزیی مجموع سمت راست آن را به ترتیب محاسبه و جایگذاری کنیم. ابتدا $\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i}$ را به‌دست می‌آوریم. با توجه به رابطه (۴.۳)، داریم

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i} = -\phi_i \frac{d}{d\mu_i} \log\Gamma(\mu_i\phi_i) + \phi_i \frac{d}{d\mu_i} \log\Gamma((1-\mu_i)\phi_i) + \phi_i \log y_i - \phi_i \log(1-y_i).$$

با توجه به تعریف $\psi(z)$، در فصل دوم، داریم

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i} = \phi_i \left[\log \frac{y_i}{1-y_i} - \Big\{ \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i) \Big\} \right]. \tag{۶.۳}$$

با توجه به تعریف‌های $\mu_i^* = \psi(\mu_i\phi_i) - \psi((1-\mu_i)\phi_i)$ و $y_i^* = \log(\frac{y_i}{1-y_i})$، رابطه (۶.۳) به‌صورت $\phi_i(y_i^* - \mu_i^*)$ قابل بازنویسی است. بنابراین (۵.۳) را می‌توان به‌صورت

$$\frac{\partial \ell(\beta,\theta)}{\partial \beta_r} = \sum_{i=1}^{n} \phi_i(y_i^* - \mu_i^*) \frac{d\mu_i}{d\eta_{1i}} \frac{\partial \eta_{1i}}{\partial \beta_r}, \tag{۷.۳}$$

نوشت. با توجه به رابطه $\mu_i = g_1^{-1}(\eta_{1i})$، داریم

$$\frac{d\mu_i}{d\eta_{1i}} = \frac{1}{g_1'(\mu_i)}, \tag{۸.۳}$$

که در آن $g'_۱(\mu_i) = \frac{d}{d\mu_i} g_۱(\mu_i)$. از طرفی

$$\frac{d\eta_{۱i}}{d\beta_r} = x_{ir}. \tag{۹.۳}$$

در نهایت با جایگذاری روابط (۷.۳)، (۸.۳) و (۹.۳) در (۵.۳)، نتیجه می‌شود

$$U_r(\beta, \theta) = \sum_{i=۱}^{n} \phi_i \frac{(y_i^* - \mu_i^*)}{g'_۱(\mu_i)} x_{ir}.$$

بنابراین

$$U_{\boldsymbol{\beta}} = \begin{pmatrix} U_۱ \\ \vdots \\ U_k \end{pmatrix} = \begin{pmatrix} \sum_{i=۱}^{n} \phi_i \frac{(y_i^* - \mu_i^*)}{g'_۱(\mu_i)} x_{i۱} \\ \vdots \\ \sum_{i=۱}^{n} \phi_i \frac{(y_i^* - \mu_i^*)}{g'_۱(\mu_i)} x_{ik} \end{pmatrix} = \mathbf{X}' \boldsymbol{\phi} T_۱ (y^* - \mu^*).$$

به طور مشابه برای محاسبه تابع امتیاز θ_R، $R = ۱, \cdots, h$ ، داریم

$$U_R(\beta, \theta) = \frac{\partial \ell(\beta, \theta)}{\partial \theta_R}.$$

با استفاده از مشتق زنجیره‌ای، برای محاسبه تابع امتیاز θ_R، با مشتق‌گیری از تابع لگاریتم درست‌نمایی، برای $R = ۱, \cdots, h$، نتیجه می‌شود

$$U_R(\beta, \theta) = \frac{\partial \ell(\beta, \theta)}{\partial \theta_R} = \sum_{i=۱}^{n} \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i} \frac{d\phi_i}{d\eta_{۲i}} \frac{d\eta_{۲i}}{d\theta_R}. \tag{۱۰.۳}$$

بنابراین برای محاسبه (۱۰.۳)، باید مشتق‌های جزیی مجموع سمت راست آن را به ترتیب محاسبه و جایگذاری کنیم. ابتدا $\frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i}$ را به‌دست می‌آوریم. با توجه به رابطه (۴.۳)، می‌توان نوشت

$$\begin{aligned} \frac{\partial \ell_i(\mu_i, \phi_i)}{\partial \phi_i} &= \frac{d}{d\phi_i} \log \Gamma(\phi_i) - \mu_i \frac{d}{d\phi_i} \log \Gamma(\mu_i \phi_i) - (۱ - \mu_i) \frac{d}{d\phi_i} \log \Gamma((۱ - \mu_i)\phi_i) \\ &\quad + \mu_i \log y_i + (۱ - \mu_i) \log(۱ - y_i) \\ &= \frac{d}{d\phi_i} \log \Gamma(\phi_i) - \mu_i \frac{d}{d\phi_i} \log \Gamma(\mu_i \phi_i) - \frac{d}{d\phi_i} \log \Gamma((۱ - \mu_i)\phi_i) \\ &\quad + \mu_i \frac{d}{d\phi_i} \log \Gamma((۱ - \mu_i)\phi_i) + \mu_i \log y_i + \log(۱ - y_i) - \mu_i \log(۱ - y_i). \end{aligned}$$

در نتیجه طبق تعاریفی که از قبل داشتیم، خواهیم داشت

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \phi_i} = \mu_i\left[\log\frac{y_i}{1-y_i} - \psi(\mu_i\phi_i) + \psi((1-\mu_i)\phi_i)\right] + \psi(\phi_i) - \psi((1-\mu_i)\phi_i) + \log(1-y_i). \tag{۱۱.۳}$$

با توجه به تعریف‌های μ_i^* و y_i^*، رابطه (۱۱.۳) به‌صورت

$$\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \phi_i} = \mu_i(y_i^* - \mu_i^*) + \psi(\phi_i) - \psi((1-\mu_i)\phi_i) + \log(1-y_i),$$

قابل بازنویسی است. بنابراین با توجه به تعریف v_i، رابطه (۱۰.۳) را می‌توان به‌صورت

$$\frac{\partial \ell(\beta,\theta)}{\partial \theta_R} = \sum_{i=1}^{n} v_i \frac{d\phi_i}{d\eta_{2i}}\frac{d\eta_{2i}}{d\theta_R}, \tag{۱۲.۳}$$

نوشت. با توجه به رابطه $\phi_i = g_2^{-1}(\eta_{2i})$، داریم

$$\frac{d\phi_i}{d\eta_{2i}} = \frac{1}{g_2'(\phi_i)}, \tag{۱۳.۳}$$

که در آن $g_2'(\phi_i) = \frac{d}{d\phi_i}g_2(\phi_i)$. از طرفی

$$\frac{d\eta_{2i}}{d\theta_R} = z_{iR}. \tag{۱۴.۳}$$

در نهایت با جایگذاری روابط (۱۳.۳) و (۱۴.۳) در (۱۲.۳) نتیجه می‌شود

$$U_R(\beta,\theta) = \sum_{i=1}^{n} \frac{v_i}{g_2'(\phi_i)} z_{iR}.$$

بنابراین

$$U_{\boldsymbol{\theta}} = \begin{pmatrix} U_1 \\ \vdots \\ U_h \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \frac{v_i}{g_2'(\phi_i)} z_{i1} \\ \vdots \\ \sum_{i=1}^{n} \frac{v_i}{g_2'(\phi_i)} z_{ih} \end{pmatrix} = \mathbf{Z}' T_2 v.$$

□

برای تکمیل فرآیند برازش مدل، نیازمند اطلاع از دقت برآوردگرهای به‌دست آمده هستیم. بنابراین باید ماتریس اطلاع فیشر برآوردگرها را به‌دست آوریم. در ادامه، قضیه ۲.۲.۳ نحوه محاسبه ماتریس اطلاع فیشر را برای مدل مورد نظر بیان می‌کند.

قضیه ۲.۲.۳. * ماتریس اطلاع فیشر برای پارامترهای مدل (۱.۳) به‌صورت $K(\xi) = P'WP$ می‌باشد که در آن

$$P = \begin{pmatrix} X & ۰ \\ ۰ & Z \end{pmatrix},$$

و W یک ماتریس $۲n \times ۲n$ به‌صورت

$$W = \begin{pmatrix} W_{\beta\beta} & W_{\beta\theta} \\ W_{\beta\theta} & W_{\theta\theta} \end{pmatrix},$$

است، به‌طوری‌که

$$\begin{aligned}
W_{\beta\beta} &= diag\left(\phi_i^۲ a_i \left(\frac{۱}{g_۱'(\mu_i)}\right)^۲\right), \\
W_{\beta\theta} &= diag\left(\phi_i\{\mu_i c_i - \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\}\left(\frac{۱}{g_۱'(\mu_i)}\right)\left(\frac{۱}{g_۲'(\phi_i)}\right)\right), \\
W_{\theta\theta} &= diag\left(b_i\left(\frac{۱}{g_۲'(\phi_i)}\right)^۲\right),
\end{aligned}$$

که در آن‌ها

$$\begin{aligned}
a_i &= \frac{d}{d\mu_i}\psi((۱-\mu_i)\phi_i) + \frac{d}{d\mu_i}\psi(\mu_i\phi_i), \\
b_i &= \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)(۱-\mu_i)^۲ + \frac{d}{d\phi_i}\psi(\mu_i\phi_i)\mu_i^۲ - \frac{d}{d\phi_i}\psi(\phi_i), \\
c_i &= \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i) + \frac{d}{d\phi_i}\psi(\mu_i\phi_i).
\end{aligned}$$

برهان. از رابطه (۵.۳) یک بار دیگر بر حسب β مشتق می‌گیریم. داریم

$$\begin{aligned}
\frac{\partial^۲\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_r\partial\beta_s} &= \sum_{i=۱}^{n}\frac{\partial}{\partial\mu_i}\left(\frac{\partial\ell_i(\mu_i,\phi_i)}{\partial\mu_i}\frac{d\mu_i}{d\eta_{۱i}}\right)\frac{d\mu_i}{d\eta_{۱i}}\frac{\partial\eta_{۱i}}{\partial\beta_s}x_{ir} \\
&= \sum_{i=۱}^{n}\left(\frac{\partial^۲\ell_i(\beta,\theta)}{\partial\mu_i^۲}\frac{d\mu_i}{d\eta_{۱i}} + \frac{\partial\ell_i(\beta,\theta)}{\partial\mu_i}\frac{\partial}{\partial\mu_i}\frac{d\mu_i}{d\eta_{۱i}}\right)\frac{d\mu_i}{d\eta_{۱i}}x_{ir}x_{is}.
\end{aligned}$$

از آن‌جایی که $E\left(\frac{\partial\ell_i(\mu_i,\phi_i)}{\partial\mu_i}\right) = ۰$، داریم

$$E\left(\frac{\partial^۲\ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_r\partial\beta_s}\right) = \sum_{i=۱}^{n}E\left(\frac{\partial^۲\ell_i(\mu_i,\theta_i)}{\partial\mu_i^۲}\right)\left(\frac{d\mu_i}{d\eta_{۱i}}\right)^۲ x_{ir}x_{is}.$$

با مشتق‌گیری از رابطه (۶.۳)، داریم

$$\begin{aligned}\frac{\partial^۲\ell_i(\mu_i,\phi_i)}{\partial\mu_i^۲} &= \frac{d}{d\mu_i}\left(\phi_i\left[\log\frac{y_i}{۱-y_i}-\Big\{\psi(\mu_i\phi_i)-\psi((۱-\mu_i)\phi_i)\Big\}\right]\right)\\ &= \phi_i\left(-\phi_i\frac{d}{d\mu_i}\psi(\mu_i\phi_i)-\phi_i\frac{d}{d\mu_i}\psi((۱-\mu_i)\phi_i)\right)\\ &= -\phi_i^۲\left(\frac{d}{d\mu_i}\psi(\mu_i\phi_i)-\frac{d}{d\mu_i}\psi((۱-\mu_i)\phi_i)\right).\end{aligned}$$

با توجه به تعریف a_i، می‌توان نوشت

$$\frac{\partial^۲\ell_i(\mu_i,\phi_i)}{\partial\mu_i^۲} = -\phi_i^۲ a_i.$$

همچنین بنابر (۸.۳)، داریم

$$\left(\frac{d\mu_i}{d\eta_{۱i}}\right)^۲ = \left(\frac{۱}{g_۱'(\mu_i)}\right)^۲.$$

بنابراین می‌توان نوشت

$$E\left(\frac{\partial^۲\ell(\beta,\theta)}{\partial\beta_r\partial\beta_s}\right) = -\sum_{i=۱}^{n}(\phi_i^۲ a_i)\left(\frac{۱}{g_۱'(\mu_i)}\right)^۲ x_{ir}x_{is}.$$

در نتیجه

$$-E\left(\frac{\partial^۲\ell(\beta,\theta)}{\partial\beta\partial\beta'}\right) = X'diag\left(\phi_i^۲ a_i\left(\frac{۱}{g_۱'(\mu_i)}\right)^۲\right)X = X'W_{\beta\beta}X.$$

به‌طور مشابه، برای محاسبه $W_{\beta\theta}$، با مشتق‌گیری از رابطه (۶.۳) نسبت به ϕ_i، داریم

$$\begin{aligned}\frac{\partial^۲\ell_i(\mu_i,\phi_i)}{\partial\mu_i\partial\phi_i} &= \frac{d}{d\phi_i}\left(\phi_i\left[\log\frac{y_i}{۱-y_i}-\Big\{\psi(\mu_i\phi_i)-\psi((۱-\mu_i)\phi_i)\Big\}\right]\right)\\ &= \log\frac{y_i}{۱-y_i}-\psi(\mu_i\phi_i)+\psi((۱-\mu_i)\phi_i)\\ &\quad -\phi_i\mu_i\frac{d}{d\phi_i}\psi(\mu_i\phi_i)+\phi_i(۱-\mu_i)\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\\ &= \phi_i\left[\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)-\mu_i\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)-\mu_i\frac{d}{d\phi_i}\psi(\mu_i\phi_i)\right]\\ &\quad +\psi((۱-\mu_i)\phi_i)-\psi(\mu_i\phi_i)+\log\frac{y_i}{۱-y_i}\\ &= \phi_i\left[-\mu_i\Big\{\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)+\frac{d}{d\phi_i}\psi(\mu_i\phi_i)\Big\}+\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\right]\\ &\quad +\psi((۱-\mu_i)\phi_i)-\psi(\mu_i\phi_i)+\log\frac{y_i}{۱-y_i}\end{aligned}$$

$$= -\phi_i \left[\mu_i a_i - \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\right] + \underbrace{\psi((۱-\mu_i)\phi_i) - \psi(\mu_i\phi_i) + \log\frac{y_i}{۱-y_i}}_{\frac{\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i}}{\phi_i}}.$$

از آن‌جایی‌که $E\left(\frac{\partial \ell_i(\mu_i,\phi_i)}{\partial \mu_i}\right) = ۰$، و بنابر (۶.۳) می‌توان نوشت

$$\begin{aligned} E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta_r\partial\theta_s}\right) &= \sum_{i=۱}^{n} E\left(\frac{\partial^۲\ell_i(\mu_i,\theta_i)}{\partial\mu_i\partial\theta_i}\right)\left(\frac{۱}{g_۱'(\mu_i)}\right)\left(\frac{۱}{g_۲'(\phi_i)}\right)x_{ir}z_{is} \\ &= \sum_{i=۱}^{n} -\phi_i\left[\mu_i c_i - \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\right]\left(\frac{۱}{g_۱'(\mu_i)}\right)\left(\frac{۱}{g_۲'(\phi_i)}\right)x_{ir}z_{is}. \end{aligned}$$

در نتیجه

$$\begin{aligned} -E\left(\frac{\partial^۲ \ell(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial\beta\partial\theta}\right) &= X'diag\left(\phi_i\left\{\mu_i c_i - \frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\right\}\left(\frac{d\mu_i}{d\eta_{۱i}}\right)\left(\frac{d\phi_i}{d\eta_{۲i}}\right)\right)Z \\ &= X'W_{\beta\theta}Z. \end{aligned}$$

برای محاسبه $W_{\theta\theta}$ با استفاده از رابطه (۱۱.۳)، می‌توان نوشت

$$\begin{aligned} \frac{\partial^۲\ell_i(\mu_i,\phi_i)}{\partial\phi_i^۲} &= \frac{d}{d\phi_i}\left(\mu_i\left[\log\frac{y_i}{۱-y_i} - \psi(\mu_i\phi_i) + \psi((۱-\mu_i)\phi_i)\right]\right) \\ &\quad + \frac{d}{d\phi_i}\left(\psi(\phi_i) - \psi((۱-\mu_i)\phi_i) + \log(۱-y_i)\right) \\ &= -\mu_i^۲\frac{d}{d\phi_i}\psi(\mu_i\phi_i) + \mu_i(۱-\mu_i)\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i) \\ &\quad + \frac{d}{d\phi_i}\psi(\phi_i) - (۱-\mu_i)\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i) \\ &= (۱-\mu_i)\left[\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i) - \mu_i\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i)\right] \\ &\quad -\mu_i^۲\frac{d}{d\phi_i}\psi(\mu_i\phi_i) + \frac{d}{d\phi_i}\psi(\phi_i) \\ &= -\mu_i^۲\frac{d}{d\phi_i}\psi(\mu_i\phi_i) - (۱-\mu_i)^۲\frac{d}{d\phi_i}\psi((۱-\mu_i)\phi_i) + \frac{d}{d\phi_i}\psi(\phi_i) \\ &= -b_i. \end{aligned}$$

حال می‌توان نوشت

$$E\left(\frac{\partial^۲\ell(\beta,\theta)}{\partial\theta_r\partial\theta_s}\right) = \sum_{i=۱}^{n} E\left(\frac{\partial^۲\ell_i(\mu_i,\theta_i)}{\partial\phi_i^۲}\right)\left(\frac{۱}{g_۲'(\phi_i)}\right)^۲ z_{ir}z_{is}.$$

بنابراین خواهیم داشت

$$-E\left(\frac{\partial^{۲}\ell(\beta,\theta)}{\partial\theta\partial\theta'}\right)=Z'diag\left(b_i\left(\frac{۱}{g'_{۲}(\phi_i)}\right)^{۲}\right)Z=Z'W_{\theta\theta}Z.$$

در نتیجه

$$\begin{aligned}K &= \begin{pmatrix}X'W_{\beta\beta}X & X'W_{\beta\theta}Z\\ Z'W_{\theta\beta}X & Z'W_{\theta\theta}Z\end{pmatrix}\\ &= \begin{pmatrix}X & ۰\\ ۰ & Z\end{pmatrix}'\begin{pmatrix}W_{\beta\beta} & W_{\beta\theta}\\ W_{\theta\beta} & W_{\theta\theta}\end{pmatrix}\begin{pmatrix}X & ۰\\ ۰ & Z\end{pmatrix}\\ &= P'WP.\end{aligned}$$

□

۳.۳ ویژگی‌های مجانبی و فاصله اطمینان

در این قسمت، ویژگی‌های مجانبی برآوردگرها را بررسی کرده و فواصل اطمینان برای آن‌ها را معرفی می‌کنیم. قضیه زیر توزیع مجانبی برآوردگرها را نشان می‌دهد.

قضیه ۱.۳.۳. (کسلا و برگر، ۲۰۰۲) تحت شرایط نظم، توزیع مجانبی برآوردگرهای ماکسیمم درستنمایی بردار پارامتر ξ، وقتی اندازه نمونه بزرگ باشد، و با فرض وجود و معکوس‌پذیر بودن $J(\xi)=\lim_{n\to\infty}\frac{K(\xi)}{n}$، عبارتست از

$$\sqrt{n}(\hat{\xi}-\xi)\sim\mathcal{N}_{k+h}\left(۰,J(\xi)^{-۱}\right).$$

با استفاده از قضیه ۱.۳.۳، فواصل اطمینان مجانبی برآوردگرها را می‌توان به‌صورت زیر محاسبه کرد. اگر $K(\hat{\xi})^{rr}$، rامین مولفه قطری ماتریس $K(\hat{\xi})^{-۱}$ و $K(\hat{\xi})^{RR}$، $(k+R)$امین مولفه قطری $K(\hat{\xi})^{-۱}$ باشند، آنگاه فواصل اطمینان مجانبی $۱۰۰(۱-\alpha)٪$ برای β_r و θ_R به‌صورت

$$\hat{\theta}_R\mp\Phi^{-۱}(۱-\frac{\alpha}{۲})\left(K(\hat{\xi})^{RR}\right)^{\frac{۱}{۲}},\qquad\hat{\beta}_r\mp\Phi^{-۱}(۱-\frac{\alpha}{۲})\left(K(\hat{\xi})^{rr}\right)^{\frac{۱}{۲}},$$

به‌دست می‌آیند.

۴.۳ مطالعه شبیه‌سازی

در این بخش با یک مطالعه شبیه‌سازی، سازگاری و نرمال مجانبی بودن برآوردگرهای مدل را برای حالتی که پارامتر دقت، ϕ، تابعی از متغیرهای تبیینی باشد، در کوچک نمونه، مورد بررسی قرار می‌دهیم. روند کار به‌همان صورتی است که در فصل قبل انجام شد، با این تفاوت که در این بخش پارامتر دقت مدل به‌صورت یک تابع است. یعنی باید برای پارامتر دقت، همانند پارامتر میانگین، μ، عمل کنیم. در نتیجه مطالعه شبیه‌سازی را با استفاده از تابع پیوند لوجیت برای میانگین و تابع پیوند لگاریتمی برای پارامتر دقت، انجام دادیم. مدل مورد نظر را به‌صورت

$$logit(\mu_i) = \eta_1 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \log(\phi_i) = \eta_2 = \theta_0 + \theta_1 z_{1i} + \theta_2 z_{2i}, \quad i = 1, \cdots, n, \tag{۱۵.۳}$$

در نظر گرفتیم، که مقادیر واقعی $(\beta_0, \beta_1, \beta_2) = (-1, 0/75, 1/5)$ و $(\theta_0, \theta_1, \theta_2) = (2, -1, 0/5)$ می‌باشند. مقادیر متغیرهای تبیینی x_1 و x_2 را به ترتیب از توزیع $N(0, 1)$ و توزیع برنولی با احتمال موفقیت ۰/۶، تولید کردیم. مقادیر متغیرهای تبیینی z_1 و z_2 را نیز به ترتیب از توزیع برنولی با احتمال موفقیت ۰/۵ و توزیع $N(1, 2)$ تولید کردیم. همچنین برای بررسی ویژگی‌های مجانبی برآوردگرها، پنج حجم نمونه $n = 30, 50, 100, 200, 400$ را در نظر گرفتیم. نتایج را بر اساس ۵۰۰ مجموعه داده شبیه‌سازی‌شده از پیشگوی خطی (۱۵.۳) به‌دست آوردیم. جدول ۱.۳ نتایج را گزارش می‌دهد.

با توجه به جدول ۱.۳، واضح است که با افزایش حجم نمونه، مقدار MSE برآوردها در حال کاهش است. همچنین، اریبی برآوردگرها، مقداری ناچیز می‌باشد. مقادیر اریبی به گونه‌ای هستند که به نظر می‌رسد، برآوردها نااریب‌اند. بنابراین، همانند فصل قبل در این مدل نیز می‌توان سازگاری برآوردگرها را در حالت حجم‌های نمونه کوچک و متوسط نتیجه گرفت.

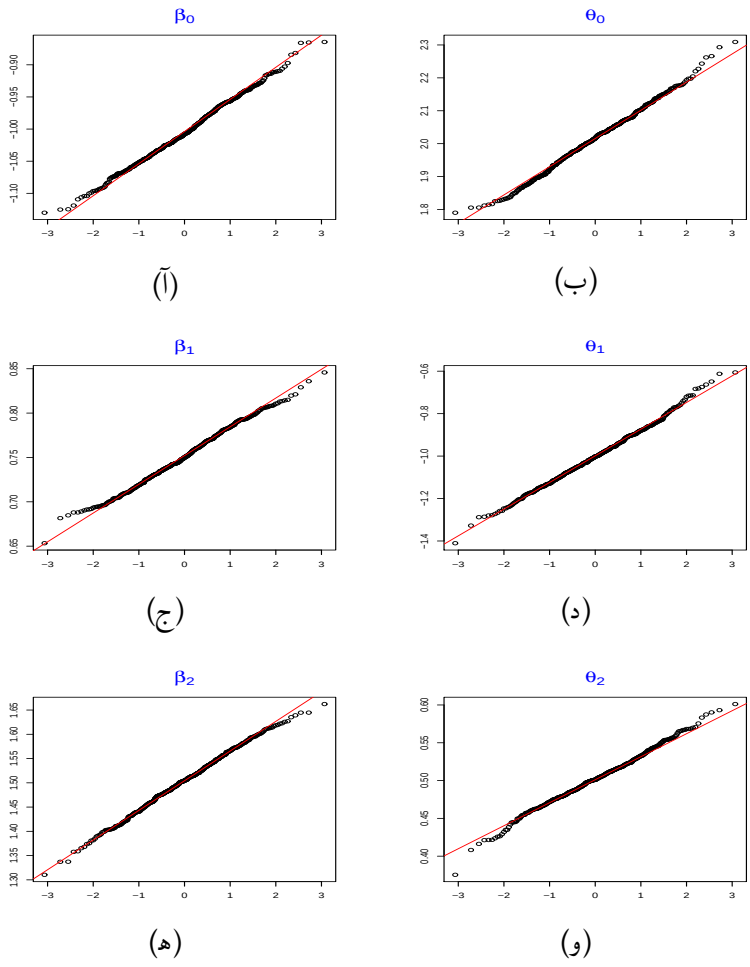
با توجه به قضیه ۱.۳.۳، برای بررسی نرمال مجانبی بودن برآوردگرهای پارامترهای مدل در کوچک نمونه، از نمودارهای چندک-چندک استفاده کردیم. شکل ۱.۳، نمودارهای چندک-چندک پارامترهای

جدول ۱.۳: نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱.۳) با پیشگوی خطی (۱۵.۳)

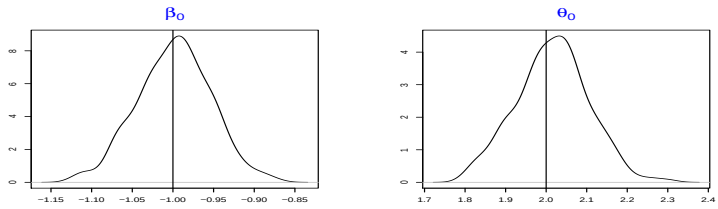
حجم نمونه	پارامترها	مقدار واقعی	برآورد	MSE	اریبی
۳۰	β_0	-1	$-1/0188$	۰/۰۴۰۷	$-0/0188$
	β_1	۰/۷۵	۰/۷۵۵۸	۰/۰۲۰۹	$-0/0058$
	β_2	۱/۵	۱/۵۳۳۱	۰/۰۵۹۱	$-0/0331$
	θ_0	۲	۲/۲۴۳۰	۰/۳۰۴۲	۰/۲۴۳۰
	θ_1	-1	$-1/1278$	۰/۴۴۹۶	$-0/1278$
	θ_2	۰/۵	۰/۵۵۶۹	۰/۰۲۶۲	۰/۰۵۶۹
۵۰	β_0	-1	$-0/9981$	۰/۰۲۱۸	$-0/0018$
	β_1	۰/۷۵	۰/۷۴۹۰	۰/۰۱۰۳	$-0/0009$
	β_2	۱/۵	۱/۴۹۸۳	۰/۰۳۴۹	$-0/0016$
	θ_0	۲	۲/۱۱۵۹	۰/۱۱۳۷	۰/۱۱۵۹
	θ_1	-1	$-1/0404$	۰/۱۷۱۴	$-0/0404$
	θ_2	۰/۵	۰/۵۲۶۲	۰/۰۱۴۶	۰/۰۲۶۲
۱۰۰	β_0	-1	$-1/0026$	۰/۰۰۹۰	$-0/0026$
	β_1	۰/۷۵	۰/۷۵۱۵	۰/۰۰۴۷	$-0/0015$
	β_2	۱/۵	۱/۴۹۵۱	۰/۰۱۵۳	$-0/0048$
	θ_0	۲	۲/۰۵۷۵	۰/۰۴۴۵	۰/۰۵۷۵
	θ_1	-1	$-1/0156$	۰/۰۷۱۵	$-0/0156$
	θ_2	۰/۵	۰/۵۱۱۹	۰/۰۰۵۲	۰/۰۱۱۹
۲۰۰	β_0	-1	$-1/0033$	۰/۰۰۴۴	$-0/0033$
	β_1	۰/۷۵	۰/۷۵۱۲	۰/۰۰۱۷	$-0/0012$
	β_2	۱/۵	۱/۵۰۴۰	۰/۰۰۷۲	$-0/0040$
	θ_0	۲	۲/۰۳۱۱	۰/۰۲۲۶	۰/۰۳۱۱
	θ_1	-1	$-1/0144$	۰/۰۳۶۰	$-0/0144$
	θ_2	۰/۵	۰/۵۰۵۶	۰/۰۰۲۲	۰/۰۰۵۶
۴۰۰	β_0	-1	$-0/9988$	۰/۰۰۲۲	$-0/0015$
	β_1	۰/۷۵	۰/۷۴۹۷	۰/۰۰۰۹	$-0/0015$
	β_2	۱/۵	۱/۴۹۹۶	۰/۰۰۳۸	$-0/0009$
	θ_0	۲	۲/۰۱۲۶	۰/۰۲۲۶	۰/۰۲۱۸
	θ_1	-1	$-0/9997$	۰/۰۱۷۱	$-0/0098$
	θ_2	۰/۵	۰/۵۰۳۴	۰/۰۰۱۰	۰/۰۰۲۵

برآوردشده در ۵۰۰ مجموعه داده شبیه‌سازی‌شده را برای مدل (۱۵.۳) نمایش می‌دهد. این شکل، نمودارهای چندک–چندک ضرایب β_0، β_1، β_2، θ_0، θ_1 و θ_2 را به ترتیب در (آ)، (ب)، (ج)، (د)، (هـ) و (و) برای حجم نمونه ۴۰۰ نمایش می‌دهد. همان‌گونه که دیده می‌شود، توزیع برآوردگرها به توزیع نرمال نزدیک است. شکل ۲.۳ نیز منحنی چگالی مقادیر برآوردشده از ۵۰۰ مجموعه داده شبیه‌سازی‌شده را برای ضرایب مدل (۱۵.۳) با $n = 400$، نمایش می‌دهد. این شکل، تایید دیگری بر خوب بودن تقریب نرمال برای توزیع برآوردگرهای پارامترهای مدل است. مقادیر واقعی پارامترها با خط‌های عمودی نمایش داده شده‌اند که تقریبا نزدیک به مقدار میانگین (برآورد پارامترها) هستند. با توجه به نتایج به‌دست آمده، می‌توان عملکرد مناسب روش برازش پیشنهادی را نتیجه گرفت.

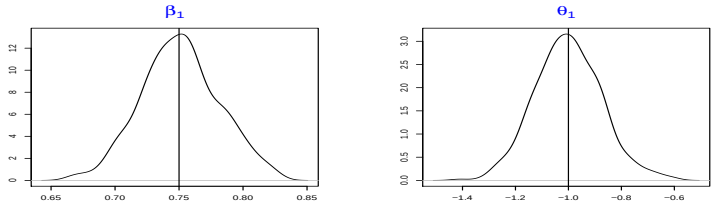
ارزیابی نیکویی برازش این مدل مشابه فصل دوم انجام می‌شود. بنابراین برای پرهیز از تکرار مطالب، از بحث در این مورد برای رگرسیون بتا با پارامتر دقت متغیر، خودداری کردیم.



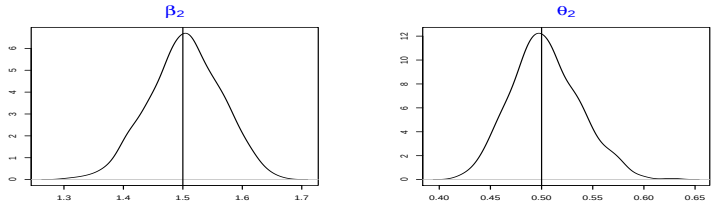
شکل ۱.۳: نمودارهای چندک-چندک برآورد ضرایب رگرسیونی میانگین و پارامتر دقت در مدل (۱۵.۳) برای $n = ۴۰۰$



(آ) (ب)



(ج) (د)



(ه) (و)

شکل ۲.۳: نمودارهای چگالی برای برآورد ضرایب رگرسیونی، به همراه نمایش مقدار واقعی آن در مدل (۱۵.۳) برای $n = ۴۰۰$

فصل ۴

کاربرد مدل رگرسیونی بتا و برازش آن با R

در این فصل، نحوه استخراج استنباط‌های مطرح‌شده در فصل‌های قبلی برای مدل‌های رگرسیونی بتا را با استفاده از بسته **betareg**، موجود در نرم‌افزار $R-۳۰۰۰$، تشریح می‌کنیم. سپس کاربرد مدل را با دو مثال واقعی، شامل داده‌های بنزین (پراتر، ۱۹۵۶) و داده‌های مقاومت بتن (رضایی و همکاران، ۱۳۹۱)، نمایش می‌دهیم.

۱.۴ مقدمه

می‌دانیم، برآوردگرهای ماکسیمم درست‌نمایی پارامترهای یک مدل رگرسیونی بتا، شکل تحلیلی و بسته‌ای ندارند. بنابراین آن‌ها را باید با روش‌های بهینه‌سازی عددی محاسبه کنیم. بسته **betareg** برای بهینه‌سازی عددی تابع درست‌نمایی از تابع $optim()$ که در هسته اصلی R تعبیه شده است، استفاده می‌کند. روش پیش‌فرض تابع $optim()$ برای بهینه‌سازی یک تابع، روش $BFGS$ می‌باشد که از جمله روش‌های شبه نیوتن است. اما سایر روش‌های عددی مانند نلدر-مید[۱] (نلدر و مید، ۱۹۶۵)، نیوتن-رافسون و غیره نیز قابل به‌کارگیری هستند.

تابع اصلی برازش مدل رگرسیونی بتا، $betareg()$ می‌باشد. همان‌طور که در فصل ۲ بیان کردیم،

[۱] Nelder-Mead Method

مدل رگرسیونی بتا عضوی از رده GLM است. بنابراین مشخصه این رده از مدل‌ها مانند پیشگوی خطی، تابع پیوند، پارامتر دقت و غیره را به ارث می‌برد.

تابع برازش مدل $betareg()$ و رده مرتبط با آن طوری طراحی شده‌اند، که تا حد امکان مشابه تابع $glm()$ استاندارد (تیم مرکزی توسعه R[۲]، ۲۰۰۹) که برای برازش GLM به‌کار می‌رود، باشد. یک تفاوت اساسی بین این دو تابع آن است که در $betareg()$ می‌توان دو معادله، برای میانگین و پارامتر دقت، تعریف کرد و در نتیجه دو ماتریس از متغیرهای تبیینی، دو پیشگوی خطی و دو مجموعه از ضرایب به‌دست آورد. روش استفاده از $betareg()$ در R به‌صورت زیر می‌باشد:

```
betareg(formula, data, subset, na.action, weights, offset,
link = "logit", link.phi = NULL, control = betareg.control(. . .),
model = TRUE, y = TRUE, x = FALSE, . . .)
```

خط اول، شامل مشخصات مدل استاندارد (چمبرز و هستی، ۱۹۹۲) است. از $formula$ برای مشخص کردن پیشگوهای خطی، مشابه تابع GLM، استفاده می‌شود. خط دوم شامل مشخصه‌های ویژه مربوط به مدل رگرسیونی بتا است. آخرین خط نیز برخی اجزا و مولفه‌های مقادیر خروجی را دربر دارد.

اگر $formula$ مثلا به صورت `y=x1+x2` باشد، آنگاه متغیرهای تبیینی $x_۱$ و $x_۲$ تنها برای مدل‌بندی میانگین رگرسیون بتا در نظر گرفته می‌شوند. در این حالت، ϕ ثابت فرض می‌شود. به عبارت دیگر، $z_i = ۱$ و $g_۲(\cdot)$ تابع پیوند همانی است که همان مدل معرفی‌شده در فصل ۲ می‌باشد. اما چنان‌چه $formula$ شامل دو بخش به‌صورت، مثلا، `y=x1+x2|z1+z2+z3`، باشد، در این حالت ϕ نیز تابعی از متغیرهای تبیینی $z_۱$ تا $z_۳$ است که مدل معرفی‌شده در فصل ۳ می‌باشد. برای مشاهده جزییات بیشتر در مورد صورت‌های مختلف تدوین $formula$ به بسته **Formula** (زیلیز و کرویسنت، ۲۰۱۰) مراجعه کنید.

[۲] R Development Core Team

تابع پیوند پیش‌فرض برای میانگین $g_1(\mu) = logit(\mu)$ و برای پارامتر دقت $g_2(\phi) = \log(\phi)$ است. تابع پیوند برای میانگین μ را می‌توان با انتخاب $link$ ، و برای پارامتر دقت ϕ توسط $link.phi$ تغییر داد. توابع مجاز برای μ همه توابع پیوند مجاز در $glm()$ شامل $logit$، $probit$، $cloglog$، $cauchit$، log و $loglog$، برای خانواده دوجمله‌ای، هستند و برای ϕ، توابع همانی، لگاریتمی و ریشه دوم می‌باشند.

جدول ۱.۴: توابع و روش‌های موجود در بسته `betareg`

تابع	توصیف
`summary()`	خلاصه‌ای از مدل رگرسیونی برازش‌شده
`coef()`	استخراج ضرایب برآوردشده مدل
`vcov()`	ماتریس کوواریانس برآوردگرها
`predict()`	پیش‌بینی داده‌های جدید
`fitted()`	برازش میانگین‌ها برای داده‌های مشاهده‌شده
`residuals()`	استخراج باقی‌مانده‌ها
`estfun()`	تخمین مشاهدات و برآورد پارامترها
`terms()`	استخراج مولفه‌های مدل
`model.matrix()`	استخراج مولفه‌های ماتریس مدل
`model.frame()`	استخراج مولفه‌های اصلی مدل
`loglik()`	استخراج لگاریتم درستنمایی برازش‌شده
`plot()`	رسم نمودارهای مناسب مدل
`coeftest()`	آزمون والد
`waldtest()`	آزمون والد
`linear.hypothesis()`	فرضیه خطی آزمون‌های والد
`lrtest()`	آزمون نسبت درستنمایی
`AIC()`	آماره‌های اطلاع $(AIC, BIC, \cdots)$

در جدول ۱.۴، مجموعه توابعی که در بسته **betareg** و بسته‌های مرتبط با آن موجودند و برای استخراج استنباط‌ها در مدل رگرسیونی بتا مفیدند، فهرست شده است.

شایان ذکر است که تابع $summary()$ آماره‌های آزمون والد برای همه ضرایب را گزارش می‌دهد، اما جدولی مشابه جدول $Anova$ برای بررسی معنی‌داری کل مدل فراهم نمی‌کند. برای داشتن چنین

جدولی، می‌توان از توابع $coeftest()$، $waldtest()$ در بسته **lmtest** (زیلیز و هترن، ۲۰۰۲) و $linear.hypothesis()$ در بسته **car** (فاکس، ۲۰۰۲) استفاده کرد. برای اجرای آزمون‌های نسبت درست‌نمایی نیز می‌توان از تابع $lrtest()$ در بسته **lmtest** استفاده کرد.

۲.۴ کاربرد مدل در دو مثال واقعی

این بخش شامل دو مثال کاربردی از مدل رگرسیون بتا است که در فصل‌های دوم و سوم مورد بحث قرار گرفتند.

۱.۲.۴ داده‌های بنزین

مثال اول مربوط به داده‌های بنزین است که توسط پراتر (۱۹۵۶) گردآوری شدند. در این مثال، به دنبال مدل‌بندی نسبت نفت خام تبدیل‌شده به بنزین بعد از فرآیند تقطیر و شکنش[۳] هستیم. این مجموعه داده توسط اتکینسن (۱۹۸۵) با استفاده از رگرسیون خطی مورد تحلیل قرار گرفته شده است. فراری و سریباری (۲۰۰۴) برای مدل‌بندی این نوع داده‌ها از دو متغیر تبیینی درجه حرارت کلی برای تبخیر نفت خام ($temp$) و درجه حرارت در هر مرحله از آزمایش ($batch$) استفاده کردند. مراحل آزمایش شامل ۹ مرحله مختلف هستند. این مجموعه داده‌ها شامل ۳۲ مشاهده است.

پارامتر دقت ثابت

ابتدا مدل را با فرض ثابت بودن پارامتر دقت ϕ در نظر گرفتیم. پیشگوی خطی برای μ_i به‌صورت

$$logit(\mu_i) = \beta_0 + \boldsymbol{\beta}\mathbf{batch}_i + \beta_{10} temp_i, \qquad i = 1, \cdots, 32, \qquad (1.4)$$

که $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_9)$ است. داده‌ها در بسته **betareg** با نام $GasolineYield$ قابل دسترس هستند. با دستورات زیر می‌توان مدل (۱.۴) را بر روی داده‌ها برازش داد:

[۳] Fractionation

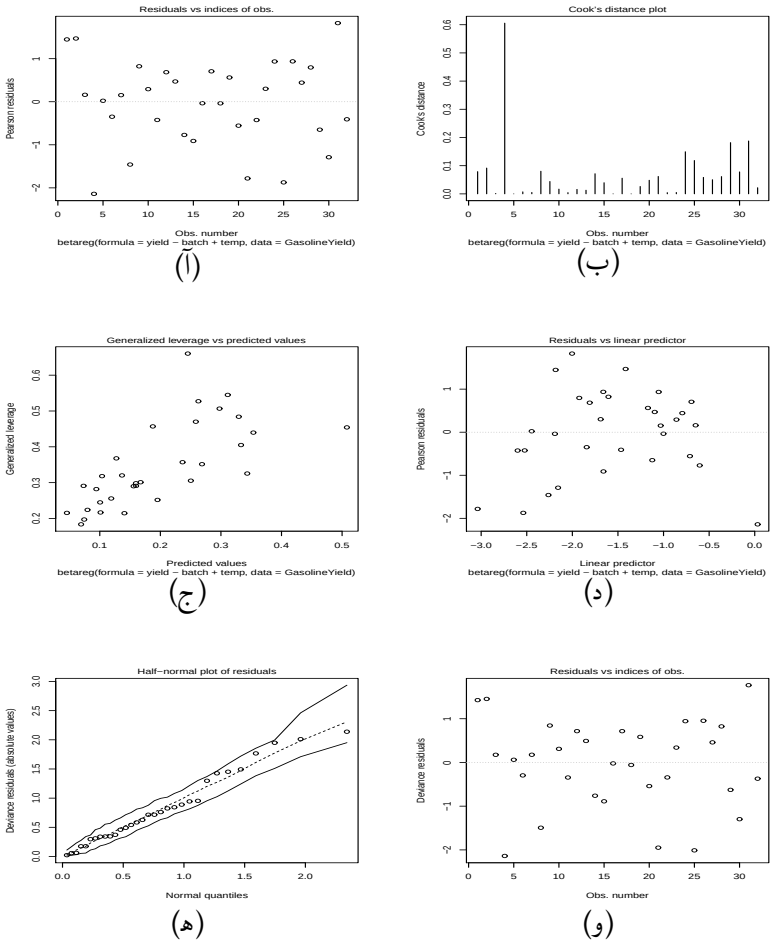
```
data("GasolineYield", package = "betareg")

gy_logit <- betareg(yield ~ batch + temp, data = GasolineYield)

summary(gy_logit)
```

جدول ۲.۴: پارامترهای برآوردشده با استفاده از داده‌های بنزین با ϕ ثابت

پارامتر	برآورد	آماره والد	خطای استاندارد	مقدار p−
عرض از مبدا	−۶/۱۵۹	−۳۳/۷۸۴	۰/۱۸۲	< ۰/۰۰۱
$batch_۱$	۱/۷۲۷	۱۷/۰۶۷	۰/۱۰۱	< ۰/۰۰۱
$batch_۲$	۱/۳۲۲	۱۱/۲۱۸	۰/۱۱۷	< ۰/۰۰۱
$batch_۳$	۱/۵۷۲	۱۳/۵۴۲	۰/۱۱۶	< ۰/۰۰۱
$batch_۴$	۱/۰۵۹	۱۰/۳۵۳	۰/۱۰۲	< ۰/۰۰۱
$batch_۵$	۱/۱۳۳	۱۰/۹۵۲	۰/۱۰۳	< ۰/۰۰۱
$batch_۶$	۱/۰۴۰	۹/۸۰۹	۰/۱۰۶	< ۰/۰۰۱
$batch_۷$	۰/۵۴۳	۴/۹۸۲	۰/۱۰۹	< ۰/۰۰۱
$batch_۸$	۰/۴۹۵	۴/۵۵۳	۰/۱۰۸	< ۰/۰۰۱
$batch_۹$	۰/۳۸۵	۳/۲۵۳	۰/۱۱۸	۰/۰۰۱
$temp$	۰/۰۱۰	۲۶/۵۷۷	$۱۰^{-۴}$	< ۰/۰۰۱
ϕ	۴۴۰/۳	۴/۰۰۲	۱۱۰/۰	< ۰۰۰۱

نتایج برازش مدل (۱.۴)، در جدول ۲.۴ گزارش شده‌اند. میزان $R^۲$ به‌دست آمده برای این مدل برابر با ۰/۹۶۱۷ می‌باشد. نمودارهای شناسایی نقاط موثر[۴] در شکل ۱.۴ نشان داده شده‌اند. این نقاط از نظر تفسیر و ماهیت، شبیه به همان نقاط موثر در مدل‌های رگرسیون خطی هستند (میرز و مونت گمری، ۱۹۹۷). شکل ۱.۴، شامل (آ) نمودار پراکنش باقی‌مانده‌های پیرسونی در مقابل شماره مشاهدات، (ب) نمودار فاصله‌های کوک، (ج) نمودار پراکنش مقادیر پیش‌بینی‌شده در مقابل مقادیر نافذ[۵]، (د) نمودار پراکنش باقی‌مانده‌ها در مقابل مقادیر پیشگوی خطی، (هـ) نمودار نیم‌نرمال باقی‌مانده‌ها و (و) نمودار باقی‌مانده‌های انحراف در مقابل شماره مشاهدات می‌باشد. بررسی شکل ۱.۴ نشان می‌دهد که مشاهده چهارم دارای اندازه فاصله کوک بزرگی است. در نتیجه، می‌توان این مشاهده را به عنوان یک نقطه موثر بر شمرد. این مشاهده دارای مقدار پاسخ برابر ۰/۴۵۷ (با مقدار برآورد ۰/۵۰۸) است.

[۴] Influence Points
[۵] Leverage



شکل ۱.۴: نمودارهای شناسایی نقاط موثر برای مدل رگرسیونی بتا، با تابع پیوند لوجیت برای داده‌های بنزین

با توجه به این نتیجه، مدل رگرسیونی بتا را بدون مشاهده چهارم برازش دادیم. در برازش جدید برآورد پارامترهای رگرسیونی β تغییر قابل توجهی نکردند، اما برآورد پارامتر دقت از ۴۴۰/۳ به ۵۷۷/۸ افزایش یافت. برآورد انحراف معیارهای مجانبی همه برآوردها نیز قابل اغماض بودند. دستورات زیر برازش مدل بدون مشاهده چهارم را اجرا می‌کنند:

```
gy_logit4 <- update(gy_logit, subset = -4)
coef(gy_logit, model = "precision")
coef(gy_logit4, model = "precision")
```

پارامتر دقت متغیر

در قسمت قبل، مقدار ϕ ثابت فرض شد. مدل را برای حالتی که ϕ تابعی از متغیرهای تبیینی باشد نیز اجرا کردیم. فرض کردیم

$$logit(\mu_i) = \beta_0 + \boldsymbol{\beta}\mathbf{batch}_i + \beta_{10} temp_i \qquad \log(\phi_i) = \theta_0 + \theta_1 temp_i. \tag{۲.۴}$$

برازش این مدل در R به‌صورت زیر است:

```
gy1_logit <- betareg(yield ~ batch + temp|temp, data = GasolineYield,
+    link.phi="log")
summary(gy1_logit)
```

نتایج برازش این مدل در جدول ۳.۴ گزارش شده است. همان‌طورکه در جدول ۳.۴ مشاهده می‌کنید، معنی‌داری ضرایب بیشتر شده و متغیر $temp$ بر روی دقت مدل موثر است. جدول ۴.۴ ملاک AIC برای دو مدل با ϕ ثابت و متغیر را نشان می‌دهد. با توجه به این دو مقدار، مدل (۲.۴)، یعنی مدلی با پارامتر دقت متغیر، برای مدل‌بندی این داده‌ها ترجیح داده می‌شود.

جدول ۳.۴: پارامترهای برآوردشده با استفاده از داده‌های بنزین با ϕ متغیر

پارامتر	برآورد	آماره والد	خطای استاندارد	p −مقدار
عرض از مبدا	−۵/۹۲۳	−۳۲/۲۷۵	۰/۱۸۳	< ۰/۰۰۱
$batch_۱$	۱/۶۰۱	۲۵/۰۸۷	۰/۰۶۳	< ۰/۰۰۱
$batch_۲$	۱/۲۹۷	۱۳/۰۹۰	۰/۰۹۹	< ۰/۰۰۱
$batch_۳$	۱/۵۶۵	۱۵/۶۹۴	۰/۰۹۹	< ۰/۰۰۱
$batch_۴$	۱/۰۳۰	۱۶/۲۷۶	۰/۰۶۳	< ۰/۰۰۱
$batch_۵$	۱/۱۵۴	۱۷/۵۸۲	۰/۰۶۵	< ۰/۰۰۱
$batch_۶$	۱/۰۱۹	۱۵/۳۶۴	۰/۰۶۶	< ۰/۰۰۱
$batch_۷$	۰/۶۲۲	۹/۴۸۱	۰/۰۶۵	< ۰/۰۰۱
$batch_۸$	۰/۵۶۴	۹/۳۸۱	۰/۰۶۰	< ۰/۰۰۱
$batch_۹$	۰/۳۵۹	۵/۳۵۴	۰/۰۶۷	< ۰/۰۰۱
$temp$	۰/۰۱۰	۲۳/۷۵۱	$۱۰^{-۴}$	< ۰/۰۰۱
عرض از مبدا	۱/۳۶۴	۱/۱۱۳	۱/۲۲۵۷	۰/۲۶۶
$temp$	۰/۰۱۴	۴/۰۲۷	۰/۰۰۳	< ۰/۰۰۱

جدول ۴.۴: مقایسه AIC برای دو مدل (۱.۴) و (۲.۴)

	df	AIC
مدل (۱.۴)	۱۲	−۱۴۵/۵۹۵۱
مدل (۲.۴)	۱۳	−۱۴۷/۹۵۴۱

۲.۲.۴ میزان مقاومت بتن

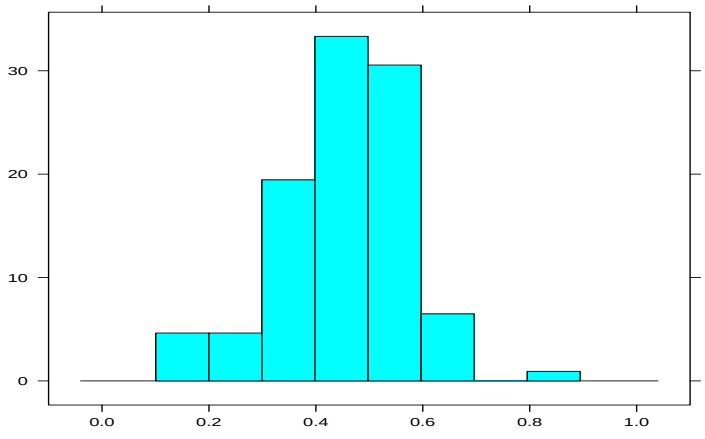
مثال بعدی، مربوط به بررسی عوامل موثر بر مقاومت بتن تولیدشده است. ابتدا، به اختصار، در مورد ساخت بتن (رضایی و همکاران، ۱۳۹۱) و کاربرد این ماده در صنعت صحبت می‌کنیم.

بتن در مفهوم وسیع به هر ماده یا ترکیبی که از یک ماده چسبنده با خاصیت سیمانی شدن تشکیل شده باشد، گفته می‌شود (اره‌پناهی، ۱۳۸۵). این ماده چسبنده عموما حاصل فعل و انفعال سیمان‌های هیدرولیکی و آب می‌باشد. امروزه چنین تعریفی از بتن شامل طیف وسیعی از محصولات می‌شود. بتن ممکن است از انواع مختلف سیمان و نیز پوزولان‌ها[۶]، سرباره کوره‌ها[۷]، گوگرد، مواد افزودنی، پلیمرها، الیاف و غیره تهیه شود (استلا و ماروسین، ۱۹۹۶). همچنین در نحوه ساخت آن ممکن است حرارت، بخار آب، اتوکلاو[۸]، خلا، فشارهای هیدرولیکی و متراکم‌کننده‌های مختلف استفاده شوند. با توجه به گسترش و پیشرفت علم و پیدایش تکنولوژی‌های فراوان در قرن اخیر، شناخت بتن و خواص آن نیز توسعه قابل ملاحظه‌ای داشته است، به نحوی که امروزه شاهد کاربرد انواع مختلف بتن با مصالح مختلف هستیم که هر یک خواص و کاربری مخصوص به خود را داراست. در حال حاضر انواع مختلفی از سیمان‌ها که شامل پوزولان‌ها، سولفورها[۹]، پلیمرها، الیاف‌های مختلف و افزودنی‌های متفاوتی هستند، تولید می‌شوند (شفرد و وسکی، ۲۰۱۳). همچنین می‌توان گفت تولید انواع بتن با استفاده از حرارت، بخار، اتوکلاو، تخلیه هوا، فشار هیدرولیکی ویبره و قالب انجام می‌گیرد. بتن، به‌طورکلی، محصولی است که از اختلاط آب با سیمان آبی و سنگدانه‌های مختلف در اثر واکنش آب با سیمان، در شرایط محیطی خاصی، حاصل می‌شود. در دسترس بودن مصالح آن، دوام نسبتاً زیاد و نیاز به ساخت و سازهای فراوان سازه‌های بتنی چون ساختمان‌ها، سازه‌ها، سدها، پل‌ها، تونل‌ها و راه‌ها، این ماده را بسیار پرمصرف نموده است. امروزه با پیشرفت علم و تکنولوژی، مشخص شده است که صرف توجه به مقاومت به‌عنوان یک

[۶] Pozzolana
[۷] Furnace Slag
[۸] Autoclave
[۹] Sulphur

معیار برای طراحی بتن، برای محیط‌های مختلف و کاربردهای مختلف، نمی‌تواند جواب‌گوی مشکلاتی باشد که در درازمدت در سازه‌های بتنی ایجاد می‌گردند. چند سالی است که مسأله دوام بتن در محیط‌های مختلف مورد توجه قرار گرفته است. مشاهده خرابی‌هایی با عوامل فیزیکی و شیمیایی در بتن‌ها در اکثر نقاط جهان و با شدتی بیشتر در کشورهای در حال توسعه، افکار و اذهان را به سمت طرح بتن‌هایی با دوام لازم سوق داده است. در این راستا، در پاره‌ای از کشورها، دستورالعمل‌ها و استانداردهایی برای طراحی بتن با عملکرد بالا تهیه شده و طراحان و مجریان در بعضی از این کشورهای پیشرفته ملزم به رعایت این دستورالعمل‌ها گشته‌اند.

داده‌هایی که مورد تحلیل قرار گرفتند، شامل ۱۰۸ مشاهده است. این داده‌ها توسط رضایی و همکاران (۱۳۹۱) گزارش شده‌اند. عواملی که در این مجموعه داده مورد بررسی قرار گرفتند، شامل: ماسه، آب، میکروسیلیس[۱۰] و سیمان هستند. متغیر پاسخ نیز مقاومت بتن تولیدشده به درصد است. شکل ۲.۴ نمودار بافت‌نگار متغیر پاسخ را نشان می‌دهد. با توجه به این نمودار، مشاهده می‌کنید که بیشتر داده‌ها در زیرفاصله (۰/۲, ۰/۶) پراکنده شده‌اند. بنابراین، بنابر مطالبی که در فصل‌های گذشته مطرح کردیم، مدل مناسب ممکن برای این داده‌ها مبتنی بر توزیع بتا است.



شکل ۲.۴: نمودار بافت‌نگار متغیر پاسخ درصد مقاومت بتن

[۱۰] Silicafume

پارامتر دقت ثابت

برای ϕ ثابت، مدل مورد نظر را به‌صورت

$$logit(\mu_i) = x_i'\boldsymbol{\beta}, \qquad (۳.۴)$$

در نظر گرفتیم، که در آن (ماسه، آب، میکروسیلیس، سیمان، ۱)= x_i . دستورات زیر مدل (۳.۴) را بر روی داده‌ها برازش می‌دهد:

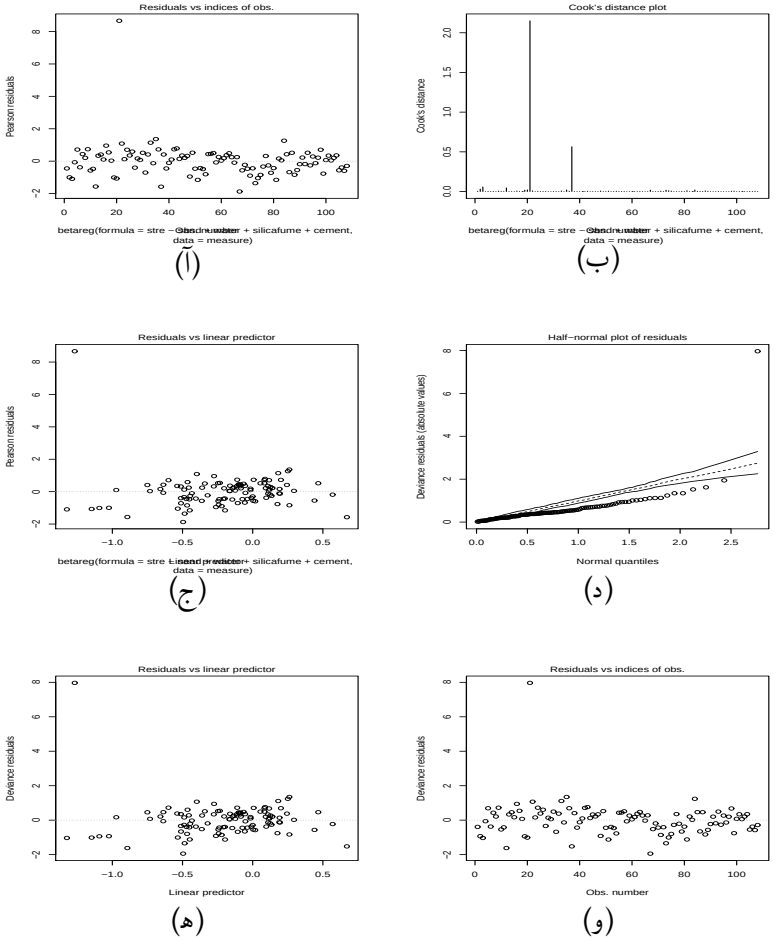
```
gy_logit = betareg(stre ~ sand + water + silicafume + cement,
 data = measure)
summary(gy_logit)
```

جدول ۵.۴: پارامترهای برآوردشده با استفاده از داده‌های ساخت بتن با ϕ ثابت

ضرایب	برآورد	آماره والد	خطای استاندارد	p −مقدار
عرض از مبدا	−۹٫۶۶۶۵	−۸٫۸۶۶	۱٫۰۹۰۳	< ۰٫۰۰۰۱
ماسه	۰٫۰۰۶۸	۷٫۴۰۹	۰٫۰۰۰۹	< ۰٫۰۰۰۱
آب	−۰٫۰۰۴۳	−۲٫۷۴۹	۰٫۰۰۱۵	۰٫۰۰۵۹
میکروسیلیس	۰٫۰۱۵۰	۶٫۳۵۴	۰٫۰۰۲۳	< ۰٫۰۰۰۱
سیمان	۰٫۰۱۰۶	۹٫۶۰۲	۰٫۰۰۱۱	< ۰٫۰۰۰۱
ϕ	۲۹٫۸۹۶	۷٫۴۶۶	۴٫۰۰۴	< ۰٫۰۰۰۱

جدول ۵.۴، نتایج مدل برازش‌شده را برای حالتی که ϕ ثابت باشد نشان می‌دهد. همان‌طور که در این جدول مشاهده می‌کنید، همه متغیرهای تبیینی معنی‌دار هستند و متغیر تبیینی آب نیز بر سختی بتن تاثیر معکوس دارد. از طرفی، بیشترین تاثیر بر میانگین سختی بتن مربوط به متغیر میکروسیلیس است. با اجرای دستورات زیر نیز می‌توان نمودارهای شناسایی مدل را رسم کرد، که نتیجه در شکل ۳.۴ نمایش داده شده است.

```
gy_logit <- betareg(stre ~ sand + water + silicafume + cement,
```



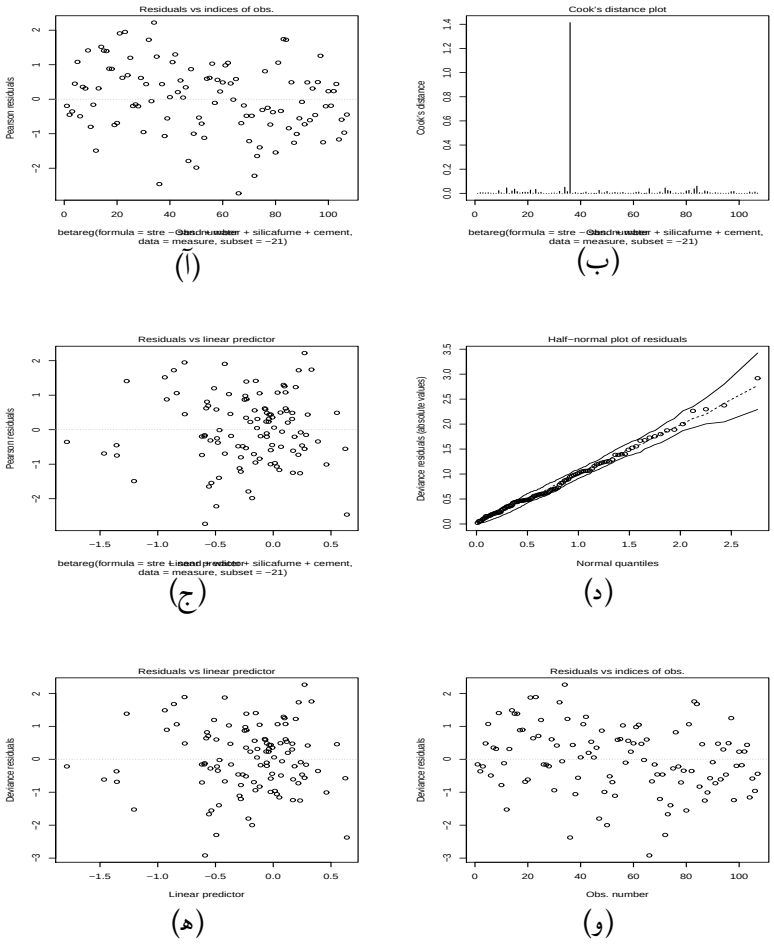
شکل ۳.۴: نمودارهای شناسایی مدل رگرسیونی بتا با تابع پیوند لوجیت برای داده‌های مقاومت بتن

```
 data = measure)

plot(gy_logit, which = 1, type = "pearson")

plot(gy_logit, which = 2, type = "pearson")

plot(gy_logit, which = 4, type = "pearson")

plot(gy_logit, which = 5, type = "deviance", sub.caption = "")

plot(gy_logit, which = 4, type = "deviance", sub.caption = "")

plot(gy_logit, which = 1, type = "deviance", sub.caption = "")
```

همان‌طور که در شکل ۳.۴ مشاهده می‌کنید، در این داده‌ها مشاهده ۲۱ام نسبت به بقیه پرت محسوب می‌شود و کیفیت مدل برازش‌شده نیز مناسب نیست. بنابراین، تصمیم گرفتیم آن را حذف کنیم و نمودارهای شناسایی مدل را دوباره رسم کردیم. شکل ۴.۴ این نمودارها را پس از حذف مشاهده ۲۱ام نشان می‌دهد. ملاحظه می‌کنید که نیکویی مدل برازش‌شده، مورد تایید می‌باشد.

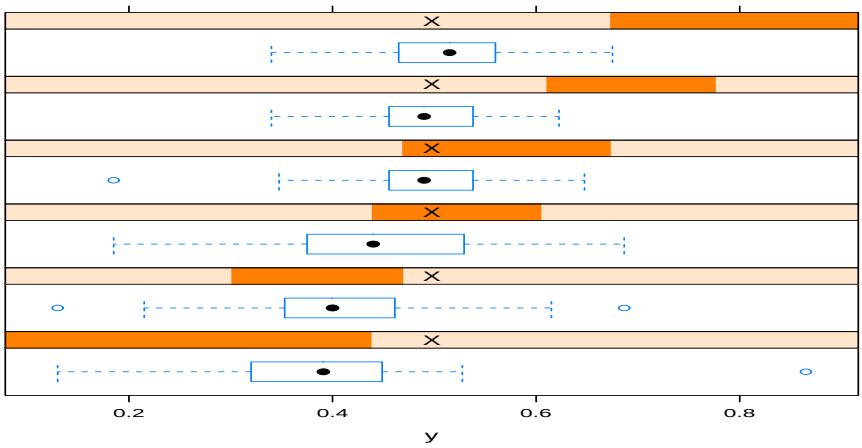
پارامتر دقت متغیر

در این حالت، مدل را زمانی که ϕ تابعی از متغیرهای تبیینی باشد، بر روی داده‌ها برازش دادیم. به منظور تعیین دقیق مدل، ابتدا با استفاده از یک تحلیل اکتشافی به دنبال متغیرهای تبیینی گشتیم که در پیشگوی خطی مربوط به پارامترهای μ و ϕ می‌توانند حضور داشته باشند. برای پارامتر μ از همه متغیرها استفاده کردیم. برای تعیین نامزدهای ممکن برای حضور در ϕ، از نمودارهای جعبه‌ای متغیر پاسخ در دسته‌هایی از متغیرهای تبیینی بهره بردیم. شکل‌های ۵.۴ تا ۸.۴، این نمودارها را برای همه متغیرهای تبیینی نمایش می‌دهند. با توجه به این شکل‌ها، مشاهده می‌کنید که تغییرات متغیر پاسخ در دسته‌هایی از متغیرهای ماسه، سیمان و آب به یک اندازه نبوده و متفاوت هستند. اما این تغییرات در دسته‌های مختلف متغیر میکروسیلیس تقریبا یکسان باقی مانده است. بنابراین، پیشگوی خطی مربوط به ϕ، متغیرهای تبیینی ماسه، آب و سیمان را به‌عنوان نامزدهای اولیه دارد.

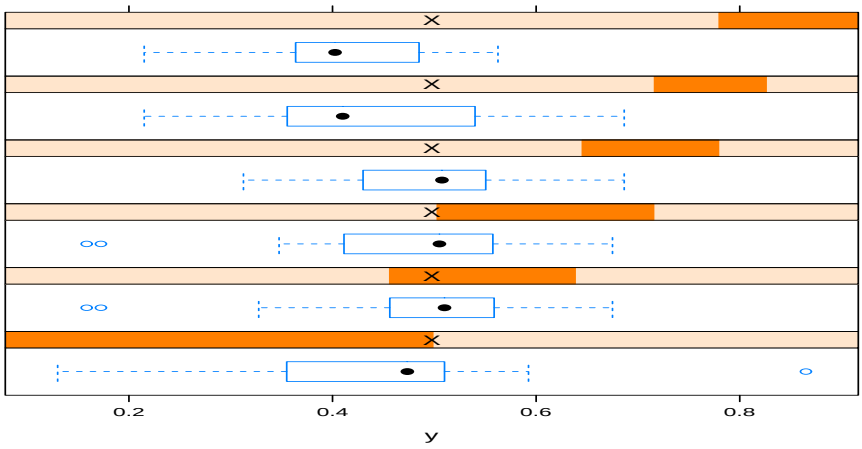


(آ) (ب) (ج) (د) (ه) (و)

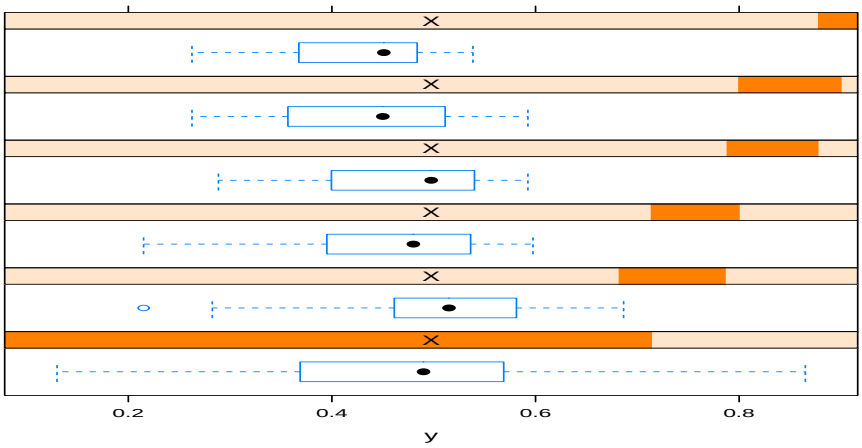
شکل ۴.۴: نمودارهای شناسایی مدل رگرسیونی بتا با تابع پیوند لوجیت، پس از حذف مشاهده ۲۱ام برای داده‌های مقاومت بتن



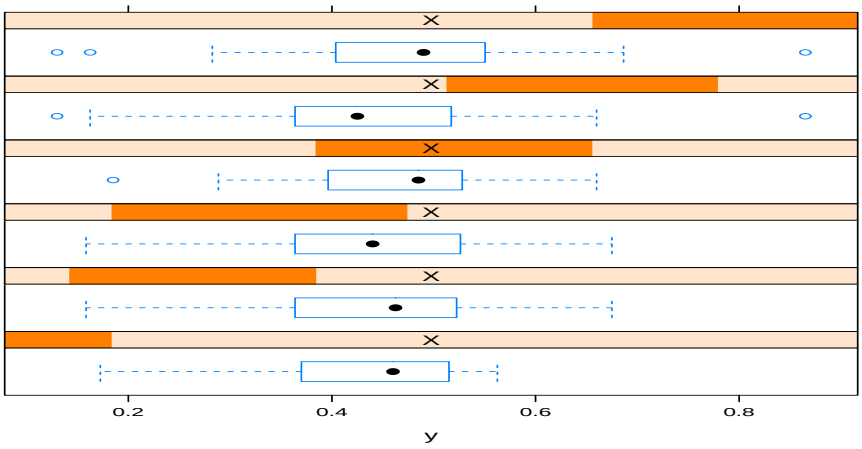
شکل ۵.۴: نمودار جعبه‌ای میزان مقاومت بتن در دسته‌های مختلفی از سیمان



شکل ۶.۴: نمودار جعبه‌ای میزان مقاومت بتن در دسته‌های مختلفی از ماسه



شکل ۷.۴: نمودار جعبه‌ای میزان مقاومت بتن در دسته‌های مختلفی از آب



شکل ۸.۴: نمودار جعبه‌ای میزان مقاومت بتن در دسته‌های مختلفی از میکروسیلیس

با توجه به توضیحات داده‌شده، مدل نهایی را زمانی که ϕ، تابعی از متغیرهای سیمان، آب و ماسه باشد، برازش دادیم. جدول ۶.۴ برآورد پارامترها و همچنین خطای استاندارد و معنی‌داری آن‌ها را برای مدل نهایی نشان می‌دهد. بنابراین، مدل نهایی برای داده‌های مقاومت بتن به صورت

$$logit(\mu_i) = x_i'\beta, \qquad \log(\phi_i) = z_i'\theta, \tag{۴.۴}$$

می‌باشد، که در آن

$$x_i = (\text{۱، ماسه، آب، سیمان}), \qquad z_i = (\text{۱، ماسه، آب، سیمان، میکروسیلیس}),$$
$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4), \qquad \theta = (\theta_0, \theta_1, \theta_2, \theta_3).$$

جدول ۶.۴: پارامترهای برآوردشده با استفاده از داده‌های ساخت بتن برای حالتی که ϕ تابعی از متغیرهای تبیینی است

ضرایب	پارامتر	برآورد	آماره والد	خطای استاندارد	p − مقدار
عرض از مبدا	β_0	−۱۱٫۰۸	−۹٫۹۸۹	۱٫۱۰۹	< ۰٫۰۰۱
ماسه	β_1	۰٫۰۰۷	۹٫۲۴۸	۰٫۰۰۱	< ۰٫۰۰۱
آب	β_2	−۰٫۰۰۵	−۵٫۲۶۴	۰٫۰۰۱	< ۰٫۰۰۱
میکروسیلیس	β_3	۰٫۰۱۵	۹٫۶۶۱	۰٫۰۰۱	< ۰٫۰۰۱
سیمان	β_4	۰٫۰۱۲	۱۳٫۴۴۶	۰٫۰۰۱	< ۰٫۰۰۱
عرض از مبدا	θ_0	−۱۰٫۸۸۹	−۳٫۴۲۶	۳٫۱۷۸۴	< ۰٫۰۰۱
ماسه	θ_1	۰٫۰۰۶	۲٫۳۴۶	۰٫۰۰۲	۰٫۰۱۸
آب	θ_2	۰٫۰۱۶	۲٫۷۶۸	۰٫۰۰۵	۰٫۰۰۵
سیمان	θ_3	۰٫۰۱۸	۵٫۲۷۵	۰٫۰۰۳	< ۰٫۰۰۱

با توجه به نتایج جدول ۶.۴ می‌توان گفت:

۱) همه ضرایب معنی‌دار هستند.

۲) تنها متغیر تبیینی آب بر سختی بتن تاثیر معکوس دارد.

۳) بیشترین تاثیر بر میانگین سختی بتن مربوط به متغیر میکروسیلیس است.

۴) برای همه متغیرهای تبیینی حاضر در پیشگوی خطی مربوط به ϕ، با افزایش سطح متغیرها میزان دقت مدل بیشتر می‌شود.

۳.۴ نتیجه‌گیری و پیشنهادات برای آینده تحقیق

در این پایان‌نامه، یک مدل رگرسیونی مناسب برای پاسخ‌هایی که به‌طور پیوسته در فاصله $(۰,۱)$ تغییر می‌کنند، ارایه شد. این وضعیتی است که متخصصان هنگام مدل‌بندی نرخ‌ها، نسبت‌ها و درصدها با آن مواجه می‌شوند. پذیره اساسی برای این نوع داده‌ها، آن است که متغیر پاسخ از توزیع بتا پیروی می‌کند. همان‌طور که معروف است، توزیع بتا برای مدل‌بندی داده‌ها در فاصله واحد بسیار انعطاف‌پذیر است، زیرا تابع چگالی بتا با پارامترهای شکل متفاوت، منحنی‌های متفاوتی را تولید می‌کند.

شبیه‌سازی‌های انجام‌شده نشان دادند که ویژگی‌های مجانبی برآوردگرهای پارامترهای مدل (در هر دو حالتی که پارامتر دقت ثابت یا متغیر باشد) حتی برای نمونه‌های کوچک و متوسط نیز قابل استفاده هستند.

تعمیم مدل به حالتی که پارامتر دقت نیز مدل‌بندی می‌شود، ما را قادر می‌سازد تا از دنیای مدل‌های رگرسیونی بر حسب میانگین توزیع، خارج شویم و مدل‌های رگرسیونی را برای سایر ویژگی‌های یک توزیع تعمیم دهیم. این تعمیم به سهم خود جالب توجه است.

انعطاف بالای مدل رگرسیونی بتا در کنار اجرای ساده آن به کمک بسته **betareg** در R موجب شده است تا در سال‌های اخیر افراد زیادی از این مدل استفاده کنند. به‌عنوان مثال، می‌توان به پاولینو (۲۰۰۱)، اسمیتسون و ورکویین (۲۰۰۶)، کلی و همکاران (۲۰۰۷)، اسپینهیرا و همکاران (۲۰۰۸)، هان (۲۰۰۸)، والیس و همکارن (۲۰۰۹)، باکلی (۲۰۰۲)، کیسچنیک و مک‌کالا (۲۰۰۳) و گارسیا و همکاران (۲۰۱۱) اشاره کرد:

با توجه به مطالب نظری و شبیه‌سازی در این پایان‌نامه، می‌توان به موضوعات زیر به عنوان آینده تحقیق اشاره کرد.

۱) مدل‌هایی که در این پایان‌نامه معرفی شدند، متکی بر پذیره استقلال بین مشاهدات هستند. بنابراین زمانی که بین داده‌ها وابستگی وجود دارد، مانند داده‌های سری زمانی، داده‌های طولی و داده‌های فضایی، باید مدل‌های معرفی‌شده را طوری تعمیم داد تا این وابستگی لحاظ شود. مدل‌های آمیخته خطی تعمیم‌یافته[۱۱] ($GLMM$) (سیرل و همکاران، ۲۰۰۱) می‌توانند برای این تعمیم انتخاب مناسبی باشند.

۲) در مدل‌های معرفی‌شده در این پایان‌نامه، پیشگوها برای میانگین و پارامتر دقت مدل، به‌صورت توابعی خطی از متغیرهای تبیینی در نظر گرفته شدند. در کاربردهای مختلفی شکل واقعی رابطه با متغیرهای تبیینی خطی نیست. به‌عنوان مثال، مدل‌های غیرخطی (سیماس و همکاران، ۲۰۱۰)، مدل‌های ناپارامتری (گرین و سیلورمن، ۲۰۰۰) و مدل‌های نیمه‌پارامتری (راپرت و همکاران، ۱۹۹۰) می‌توانند عملکرد بهتری داشته باشند. می‌توان مدل‌های رگرسیون بتای معرفی‌شده در این پایان‌نامه را برای این حالت‌ها تعمیم داد.

۳) رهیافت مورد استفاده در این پایان‌نامه، برای استنباط، مبتنی بر درست‌نمایی است. تحلیل مدل‌های رگرسیونی بتا در دیدگاه بیزی نیز می‌تواند جالب توجه باشد. از طرفی، هر دو پیشنهاد ۱) و ۲) را از دیدگاه بیزی می‌توان مورد بررسی قرار داد.

[۱۱] Generalized Linear Mixed Models

پیوست آ

دستورهای لازم برای بازتولید نتایج و استنباط‌های مثال‌های پایان‌نامه

در این پیوست، دستورهای مورد نیاز برای بازتولید نتایج برخی از مثال‌های دو فصل دوم و سوم، با استفاده از نرم‌افزار R، گزارش شده‌اند.

۱) دستورات مربوط برای تولید شکل ۱.۲ به‌صورت زیر می‌باشند:

برای حالت (آ):

```
f=function(x,a,b){
  dbeta(x,a,b)
}
curve(f(x,0.25,4.75),c(0,1),xlim=c(0,1),ylim=c(0,12),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,1.25,3.75),c(0,1),col="red",xlim=c(0,1),ylim=c(0,12),main=' ',
xlab=' ',ylab=' ')
```

```
par(new=T)
curve(f(x,2.5,2.5),c(0,1),col="blue",xlim=c(0,1),ylim=c(0,12),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,3.75,1.25),c(0,1),col="red",xlim=c(0,1),ylim=c(0,12),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,4.75,0.25),c(0,1),xlim=c(0,1),ylim=c(0,12),main=' ',
xlab=' ',ylab=' ')
```

برای حالت (ب):

```
curve(f(x,0.75,14.25),c(0,1),xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,1.5,13.5),c(0,1),col="red",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,3.75,11.25),c(0,1),col="green",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,7.5,7.5),c(0,1),col="blue",xlim=c(0,1),ylim=c(0,15),main='',
xlab=' ',ylab=' ')
```

```
par(new=T)
curve(f(x,11.25,3.75),c(0,1),col="green",xlim=c(0,1),ylim=c(0,15),main='',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,13.5,1.5),c(0,1),col="red",xlim=c(0,1),ylim=c(0,15),main='',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,14.25,0.75),c(0,1),xlim=c(0,1),ylim=c(0,15),main='',
xlab=' ',ylab=' ')
```

برای حالت (ج):

```
curve(f(x,2.5,47.5),c(0,1),xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,5,45),c(0,1),col="red",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,12.5,37.5),c(0,1),col="green",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,25,25),c(0,1),col="blue",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
```

```
par(new=T)
curve(f(x,37.5,12.5),c(0,1),col="green",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,45,5),c(0,1),col="red",xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,47.5,2.5),c(0,1),xlim=c(0,1),ylim=c(0,15),main=' ',
xlab=' ',ylab=' ')
```

و برای حالت (د)

```
curve(f(x,5,95),c(0,1),xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,10,90),c(0,1),col="red",xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,25,75),c(0,1),col="green",xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,50,50),c(0,1),col="blue",xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
```

```
par(new=T)
curve(f(x,75,25),c(0,1),col="green",xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,90,10),c(0,1),col="red",xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
par(new=T)
curve(f(x,95,5),c(0,1),xlim=c(0,1),ylim=c(0,20),main=' ',
xlab=' ',ylab=' ')
```

۲) دستورات مربوط برای استخراج نتایج مثال شبیه‌سازی اول فصل دوم:

```
########## Loding Required Pakages
library(Formula)
library(betareg)
########## Some initial real values
n = 50
beta=c(1,2)
phi=1.5
########## Generating data
gen.data = function(beta,X){
  eta = X%*%beta
  mu = exp(eta)/(1+exp(eta))
```

```
  alpha = mu*phi
  gamma = (1-mu)*phi
  y = rbeta(n,alpha,gamma)
  data = data.frame(y,X)
  index = which(data$y==1)
  data = data[-index,]
  return(data)
}
########## Fitting beta regression model
M=1000
param = matrix (NA, nrow = M, ncol = 3)
betareg.fitting = function(n,M,beta,phi){
  for (j in 1:M){
  x1 = rep(1,length(n))
  x2 = rnorm(n)
  X = cbind(x1,x2)
  data = gen.data(beta,X)
  try({
    fit = betareg(y ~ x2, data = data, start = c(1, 2, 1.5))
    param[j,] = as.vector(unlist(fit$coefficients))
    })
  }
```

```
  return(param)
}
sim = betareg.fitting(n,M,beta,phi)
estimate = apply(sim, 2, function(x) mean(x, na.rm = TRUE))
estimate
############## Computing MSE
beta1=c(1,2,phi)
Mse=numeric(3)
for (k in 1:3){
mse=function(x) mean((x-beta1[k])^2)
Mse[k]=mse(sim[,k])
}
############## Computing bias
b1=estimate[1]-1
b2=estimate[2]-2
b3=estimate[3]-1.5
```

۳) دستورات لازم برای استخراج نتایج مثال شبیه‌سازی فصل سوم:

```
########## Some initial real values
n = 400
beta=c(-1,.75,1.5)
theta=c(2,-1,0.5)
```

```
##########Generating data
gen.data2 = function(beta,theta,X,Z){
  eta1 = X%*%beta
  mu = exp(eta1)/(1+exp(eta1))
  eta2 = Z%*%theta
  phi = exp(eta2)
  alpha = mu*phi
  gamma = (1-mu)*phi
  y = rbeta(n,alpha,gamma)
  data = data.frame(y,X,Z)
  return(data)
  }
########## Fitting beta regression model with random precision
M=500
param = matrix (NA, nrow = M, ncol = 6)
betareg.fitting = function(n,M,beta,theta){
    for (j in 1:M){
    x1 = rep(1,length(n))
    x2 = rnorm(n)
    x3 = rbinom(n,1,0.6)
    X = cbind(x1,x2,x3)
    z1 = rep(1,length(n))
```

```
    z2 = rbinom(n,1,0.5)

    z3 = rnorm(n,1,2)

    Z= cbind(z1,z2,z3)

    data = gen.data2(beta,theta,X,Z)

    try({

     fit = betareg(y ~ x2 + x3| z2 +z3, data = data, link.phi= "log",

        hessian = TRUE,start = c(-1, 0.75, 1.5, 2, -1, 0.5))

        param[j,] = as.vector(unlist(fit$coefficients))

        })

    }

    return(param)

}

sim = betareg.fitting(n,M,beta,theta)

q=which(is.na(sim[1:500,])==T)

sim=sim[-q,]

estimate = apply(sim, 2, function(x) mean(x, na.rm = TRUE))

estimate

Es=estimate

############## Computing MSE

beta1=c(-1, 0.75, 1.5, 2, -1, 0.5)

Mse=numeric(6)

for (k in 1:6){
```

```
mse=function(x) mean((x-beta1[k])^2)

Mse[k]=mse(sim[,k])

}

Mse

############## Computing bias

b1=Es[1]+1

b2=Es[2]-0.75

b3=Es[3]-1.5

t1=Es[4]-2

t2=Es[5]+1

t3=Es[6]-0.5

############### Computing density-plot

plot(density(f1),main="",xlab="",ylab="")

title("beta0",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=-1)

plot(density(f2),main="",xlab="",ylab="")

title("beta1",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=0.75)

plot(density(f3),main="",xlab="",ylab="")

title("beta2",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=1.5)

plot(density(f4),main="",xlab="",ylab="")
```

```
title("theta0",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=2)

plot(density(f5),main="",xlab="",ylab="")

title("theta1",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=-1)

plot(density(f6),main="",xlab="",ylab="")

title("theta2",cex.main =3,   font.main= 1, col.main= "blue")

abline(v=0.5)

############### Computing variances

var1=mean((sim[,1]-mean(sim[,1]))^2)

var2=mean((sim[,2]-mean(sim[,2]))^2)

var3=mean((sim[,3]-mean(sim[,3]))^2)

var4=mean((sim[,4]-mean(sim[,4]))^2)

var5=mean((sim[,5]-mean(sim[,5]))^2)

var6=mean((sim[,6]-mean(sim[,6]))^2)

##########################################

f1=(sim[,1]-mean(sim[,1]))/sqrt(var1)

f2=(sim[,2]-mean(sim[,2]))/sqrt(var2)

f3=(sim[,3]-mean(sim[,3]))/sqrt(var3)

f4=(sim[,4]-mean(sim[,4]))/sqrt(var4)

f5=(sim[,5]-mean(sim[,5]))/sqrt(var5)

f6=(sim[,6]-mean(sim[,6]))/sqrt(var6)
```

```
d=seq(-4,4,0.01)

par(mfrow=c(3,2))

############### Computing qq-plot

qqnorm(f1,main="",xlab="",ylab=""); qqline(f1, col = 2)

title("beta0",cex.main = 3,   font.main= 1, col.main= "blue")

qqnorm(f2,main="",xlab="",ylab=""); qqline(f2, col = 2)

title("beta1",cex.main = 3,   font.main= 1, col.main= "blue")

qqnorm(f3,main="",xlab="",ylab=""); qqline(f3, col = 2)

title("beta2",cex.main = 3,   font.main= 1, col.main= "blue")

qqnorm(f4,main="",xlab="",ylab=""); qqline(f4, col = 2)

title("teta0",cex.main = 3,   font.main= 1, col.main= "blue")

qqnorm(f5,main="",xlab="",ylab=""); qqline(f5, col = 2)

title("teta1",cex.main = 3,   font.main= 1, col.main= "blue")

qqnorm(f6,main="",xlab="",ylab=""); qqline(f6, col = 2)

title("teta2",cex.main = 3,   font.main= 1, col.main= "blue")
```

مراجع

[۱] اره‌پناهی، م. (۱۳۸۵)، مکانیک خاک و تکنولوژی بتن، چاپ پنجم، انتشارات فرا آموزش.

[۲] رضایی، م. (۱۳۹۱)، تخمین خواص بنیادین بتن‌های hsc با استفاده از روش‌های آماری و شبکه‌های عصبی مصنوعی، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی سهند.

[۳] نیرومند، ح. (۱۳۸۷)، تحلیل رگرسیون خطی، چاپ اول، موسسه چاپ و انتشارات دانشگاه فردوسی مشهد.

[4] Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, Wiley, New York.

[5] Atkinson, A.C. (1985), *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford University Press, New York.

[6] Bai-Ni, G. & Feng, Q. (2010), Some Properties of The PSI and Polygamma Functions, *Hacettepe Journal of Mathematics and Statistics*, 39, 219–231.

[7] Barndorff-Nielsen, O.E. (1983), On a Formula for the Distribution of the Maximum Likelihood Estimator, *Biometrika*, 70, 343–365.

[8] Bickel, P.J. & Doksum, K.J. (2001), *Mathematical Statistics*, Prentice-Hall, New Jersey.

[9] Buckley, J. (2002). Estimation of Models with Beta-Distributed Dependent Variables: A Replication and Extension of Paolino (2001), *Political Analysis*, 11, 1–12.

[10] Byrd, R.H., Lu, P., Nocedal, J. & Zhu, C. (1995), A Limited Memory Algorithm for Bound Constrained Optimization, *SIAM Journal of Scientific Computing*, 16, 1190–1208.

[11] Casella, G. & Berger, R.L. (2002), *Statistical Inference*, 2nd ed., New York.

[12] Chambers, J.M. & Hastie, T.J. (1992), *Statistical Models in S*, Chapman and Hall, London.

[13] Cribari-Neto, F. & Marcela P.F.Q. (2012), On Testing Inference in Beta Regressions, *Journal of Statistical Computation and Simulation*, 5, 0740–540.

[14] Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, 7, 1-26.

[15] Espinheira, P., Ferrari, S.L.P. & Cribari-Neto, F. (2008), On Beta Regression Residuals, *Journal of Applied Statistics*, 35, 407–419.

[16] Ferrari, S.L.P. & Cribari-Neto, F. (2004), Beta Regression for Modelling Rates and Proportion, *Journal of Applied Statistics*, 7, 799-815.

[17] Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*, Sage Publications, Thousand Oaks, CA.

[18] Garcia, C., Garcia, J. & Van Dorp, J.R. (2011), Modeling Heavy-Tailed, Skewed and Peaked Uncertainty Phenomena with Bounded Support, *Statistical Methods and Applications*, 20, 463–486.

[19] Green, P.J. & Silverman, B.W. (2000), *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, Florida.

[20] Hahn, E.D. (2008), Mixture Densities for Project Management Activity Times: A Robust Approach to PERT, *European Journal of Operational Research*, 188, 450–459.

[21] Hosmer, D.W., Lemeshow, J.S. & Rodney, X.S. (2013), *Applied Logistic Regression*, 3th ed., Wiley, New York.

[22] Kelly, G., Garabed, R., Branscum, A., Perez, A. & Thurmond, M. (2007), Prediction Model for Sequence Variation in the Glycoprotein Gene of Infectious Hematopoietic Necrosis Virus in California, USA, *Diseases of Aquatic Organisms*, 78, 97–104.

[23] Kieschnick, R. & McCullough, B.D. (2003), Regression Analysis of Variates Observed on (0,1): Percentages, Proportions, and Fractions, *Statistical Modeling*, 3, 193–213.

[24] Meyer, J.P., Stanley, D.J., Herscovitch, L. & Topolnytsky, L. (2002), Affective, Continuance, and Normative Commitment to the Organization: A Meta-analysis of Antecedents, Correlates, and Consequences, *Journal of Vocational Behavior*, 61, 20–52

[25] McCullagh, P. & Nelder, J.A. (1989), *Generalized Linear Models, 2nd ed.*, Chapman and Hall, London.

[26] McCullach, C.E. & Searle, S. (2001), *Generalized, Linear, and Mixed Models*, Department of Statistical and Biometrics, Cornell University.

[27] Myers, R.H. & Montgomery, D.C. (1997), A Tutorial on Generalized Linear Models, *Jurnal of Quality Technology*, 29, 274-291

[28] Nelder, J.A. & Mead, R. (1965), A Simplex Algorithm for Function Minimization, *Computer Journal*, 7, 308–313.

[29] Nelder, J.A. & Wedderburn, R.W.M. (1972), Generalized Linear Models, *Journal of the Royal Statistical Society, A*, 135, 370–384.

[30] Nocedal, J., Wright, S.J. (1999), *Numerical Optimization*, Springer-Verlag, New York.

[31] Paolino, P. (2001), Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables, *Political Analysis*, 9, 325–346.

[32] Prater, N.H. (1956), *Estimate Gasoline Yields from Crudes*, Petroleum Refiner, 35, 236-238.

[33] R Development Core Team. (2009), R Foundation for Statistical Computing, Vienna, Austria, *R: A Language and Environment for Statistical Computing.*, ISBN 3-900051-07-0.

[34] Rizzo, M.L. (2002), *Statistical Computing With R*, Chapman & Hall, London.

[35] Ruppert, D., Wand, M.P. & Carroll, R.J. (1990), Semiparametric Regression, *Combridge University Press,* London.

[36] Shepherd, S. & Woskie, S. (2013), Controlling Dust From Concrete Saw Cutting, *Journal of Occupational and Enviromental Hygiene* 10, 64-70.

[37] Simas, A.B., Barreto-Souza, W. & Rocha, A.V. (2010), Improved Estimators for a General Class of Beta Regression Models, *Computational Statistics and Data Analysis*, 54, 348-366.

[38] Skovgaard, I.M. (2001), Likelihood Asymptotics, *Scandinavian Journal of Statistics*, 28, 3–32.

[39] Smithson, M. & Verkuilen, J. (2006), A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables, *Psychological Methods*, 11, 54–71.

[40] Stella, L. & Marusin, S. (1996), Ancient Concretest Structures, *Concrete International*, 56-68.

[41] Strang, G. (2003), Introduction to Linear Algebra 3th ed., *Wellesley Cambridge Press*.

[42] Wald, A. (1943), Test of Statistical Hypotheses Concerning Several Parameters When the Number of Obsevations is Large, *Transcations of the American Mathematical Society*, 54, 426–482.

[43] Wallis, E., Mac Nally, R. & Lake, S. (2009), Do Tributaries Affect Loads and Fluxes of Particulate Organic Matter, Inorganic Sediment and Wood? Patterns in an Upland River Basin in South-Eastern Australia, *Hydrobiologia*, 636, 307–317.

[44] Washington, S.P., Karlaftis, M.G. & Mannering, F.L. (2011), *Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed.*, Chapman and Hall, United States.

[45] Yanqing, Y. & Wang, X. (2011), Comparsion of Wald, Score, and Likelihood Ratio Tests for Response Adaptive Designs, *Journal of Statistial Theory and Applications*, 10, 553–569

[46] Zeileis, A., Croissant, Y. (2010), Extended Model Formulas in R: Multiple Parts and Multiple Responses, *Journal of Statistical Software*, 34, 1-13.

[47] Zeileis, A., Hothorn, T. (2002), Diagnostic Checking in Regression Relationships, *R News*, 2, 7-10.

واژه‌نامه فارسی به انگلیسی

اتوکلاو ..... Autoclove

از ریخت‌افتادگی‌های حجمی ..... Display Considerable Size Distortions

آزمون والد ..... Wald's Test

آماره کی‌دو پیرسن ..... Chi-Square Pearson Statistic

اریبی ..... Bias

انحراف ..... Deviance

انحراف مقیاس‌بندی شده ..... Scaled Deviance

بازنمونه‌گیری ..... Resampling Methods

باقی‌مانده انس‌کمب ..... Anscombe Residual

باقی‌مانده پیرسن ..... Pearson Residual

برگشت ..... Regression

پارامتر شکل ..... Shape Parameter

پزولان‌ها ..... Pozzolana

تابع امتیاز ..... Score Function

تابع پیوند ..... Link Function

تابع پیوند لگاریتمی ..... Log Link Function

تابع پیوند متعارف ..... Canonical Link Function

تابع چندگاما ..... Polygamma Function

تابع سه‌گاما ..... Trigamma Function

توزیع چوله ..... Skewed Distribution

تیم مرکزی توسعه R ..... R Development Core Team

دترمینان ..... Determinant

دقت ..... Precision

روش بوت‌استرپ ..... Bootstrap Method

روش نلدر–مید ..... Nelder-Mead Method

سازگاری ..... Consistency

سرباره کوره ..... Furnace Slag

سولفور ..... Sulphur

شبه نیوتن ..... Quasi-Newton

شکنش ..... Fractionation

فاصله‌های کوک ..... Cook Distances

کمترین توان‌های دوم موزون ..... Minimum Weighted Least Squar

کهاد ..... Minor

لانه‌ای ..... Nested

ماتریس اطلاع فیشر ..... Fisher's Information Matrix

ماتریس الحاقی ..... Adjoint Matrix

ماتریس ترانهاده ..... Transpose Matrix

ماتریس معین مثبت ..... Positive Definite Matrix

ماتریس نیمه معین مثبت ..... Positive Semi Matrix

ماتریس همسازه ..... Cofactor Matrix

مدل‌های آمیخته خطی تعمیم‌یافته ..... Generalized Linear Mixed Models

مدل‌های خطی تعمیم‌یافته ..... Generalized Linear Models

مدل‌های لگ خطی ..... Log-Linear Models

معکوس ماتریس ..... Inverse of a Matrix

معیار اطلاع آکاییک ..... Akaike Information Criterion

مقادیر نافذ ..... Leverage

موثر ..... Influence

میانگین توان دوم خطا ..... Mean Squared Error

میکروسیلیس ..... Silicafume

نمودار چندک-چندک ..... Quntile-Quntile Plot

واژه‌نامه انگلیسی به فارسی

ماتریس الحاقی ..... Adjoint Matrix

معیار اطلاع آکاییک ..... Akaike Information Criterion

باقی‌مانده انس‌کمب ..... Anscombe Residual

اتوکلاو ..... Autoclove

اریبی ..... Bias

روش بوت‌استرپ ..... Bootstrap Method

تابع پیوند متعارف ..... Canonical Link Function

آماره کی‌دو پیرسن ..... Chi-Square Pearson Statistic

ماتریس همسازه ..... Cofactor Matrix

سازگاری ..... Consistency

فاصله‌های کوک ..... Cook Distances

دترمینان ..... Determinant

انحراف ..... Deviance

از ریخت‌افتادگی‌های حجمی ..... Display Considerable Size Distortions

ماتریس اطلاع فیشر ..... Fisher's Information Matrix

شکنش ..... Fractionation

سرباره کوره ..... Furnace Slag

مدل‌های آمیخته خطی تعمیم‌یافته ..... Generalized Linear Mixed Models

مدل‌های خطی تعمیم‌یافته ..... Generalized Linear Model

موثر ..... Influence

معکوس ماتریس ..... Inverse of a Matrix

مقادیر نافذ ..... Leverage

تابع پیوند ..... Link Function

مدل‌های لگ خطی ..... Log-Linear Models

تابع پیوند لگاریتمی ..... Log Link Function

میانگین توان دوم خطا ..... Mean Squared Error

کهاد ..... Minor

کمترین توان‌های دوم موزون ..... Minimum Weighted Least Squar

روش نلدر–مید ..... Nelder-Mead Method

لانه‌ای ..... Nested

باقی‌مانده پیرسن ..... Pearson Residual

تابع چندگاما ..... Polygamma Function

ماتریس معین مثبت ..... Positive Definite Matrix

ماتریس نیمه معین مثبت ..... Positive Semi Matrix

پزولان‌ها ..... Pozzolana

دقت ..... Precision

شبه نیوتن ..... Quasi-Newton

نمودار چندک–چندک ..... Quntile-Quntile Plot

تیم مرکزی توسعه R ..... R Development Core Team

برگشت ..... Regression

بازنمونه‌گیری ..... Resampling Methods

انحراف مقیاس‌بندی شده ..... Scaled Deviance

تابع امتیاز ..... Score Function

پارامتر شکل ..... Shape Parameter

میکروسیلیس ..... Silicafume

توزیع چوله ..... Skewed Distribution

سولفور ..... Sulphur

ماتریس ترانهاده ..... Transpose Matrix

تابع سه‌گاما ..... Trigamma Function

آزمون والد ..... Wald's Test

Surname: Valipur Pashakolaei Name: Miaad

Title: Beta Regression for Modeling Rates and Proportions Data

Supervisor: Dr. Hossein Baghishani
Advisor: Dr. Mohammad Arashi

Degree: Master of Science Subject: Statistics
Field: Regression Models

Technology University of Shahrood Faculty of Mathematical Sciences
Date: January 2014 Number of pages: 101

Keywords: Asymptotic Distribution, Beta Regression, Fisher's Information Matrix, Generalized Linear Model, Link Function, Score Function

Abstract

In the most applications, researchers are imterested in finding out a relation between a response and some covariates. For such aim, a regression model is usually implemented. In many practical situation, the response variable is continuous and restricted to the interval (0,1); e.g. proportions, rates and percentages. For examples: in econometrics, peaple are interested in finding out the relation between growth rate, unemployment rate, and GDP percentage with several other economic variables. To modeling such responses, the common models are logistic and probit regressions. However, the proportions and rates are usually concentrated in a specific subinterval. In other words, such responses are skewed. Therefore, the logistic and probit models are not appropriate for modeling these responses. Regarding to these constraints, the appropriate and efficient model for this kind of data is the novel and beta regression model. In this thesis, we first define the proposed model and then estimate the parameters by maximum likelihood approach. We establish the asymptotic properties of the extracted estimators as well. The efficiency of beta regression model and its asymptotic properties are explored by a few simulation studies. Finally, we apply the proposed model to analyze two examples; first example is on gas percent output from crude oil, and the second

is on concrete resistivity.

Technology University of Shahrood
Faculty of Mathematical Sciences

Dissertation Submitted in Partial
Fulfillment of The Requirements For The
Degree of Master of Science in
Statistics

Beta Regression for Modeling Rates and Proportions Data

Supervisor
Dr. Hossein Baghishani
Advisor
Dr. Mohammad Arashi
by
Miaad Valipur Pashakolaei

January 2014