بسم الله الرحمن الرحيم

دانشگاه شیراز

دانشکده علوم

پایان نامه کارشناسی ارشد در رشته آمار ریاضی

بررسی سره یا ناسره بودن توزیع پسین و وجود ماکسیمم درستنمایی برای مدل های رگرسیونی با متغیرهای کمکی به طور تصادفی گمشده

توسط:

رویا نصیرزاده

استاد راهنما:

دکتر عبدالرسول برهانی حقیقی

شهریورماه ۱۳۸۶

به نام خدا

بررسی سره یا ناسره بودن توزیع پسین و وجود ماکسیمم درستنمایی برای مدل های رگرسیونی با متغیرهای کمکی به طور تصادفی گمشده

به وسیله‌ی :

رویا نصیرزاده

پایان‌نامه
ارائه‌شده به تحصیلات تکمیلی دانشگاه به‌عنوان بخشی
از فعالیت های لازم برای اخذ درجه کارشناسی ارشد

در رشته‌ی:

آمار ریاضی

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان‌نامه با درجه : عالی

دکتر عبدالرسول برهانی حقیقی، استادیار بخش آمار (رئیس کمیته).....

دکتر علیرضا نعمت اللهی، دانشیار بخش آمار.....

دکتر مینا توحیدی، استادیار بخش آمار.....

شهریورماه ۱۳۸۶

تقدیم به

پدر و مادر عزیزم، به خاطر حضورشان

تقدیم به

پویندگان دریای بی کران علم و دانش

سپاسگزاری

حمد و سپاس عالم علم و معلوم را که هر چه هست از اوست. خالق یکتا که انسان را مسیری نمود پر فراز و نشیب، عقلی را عطا فرمود تا بوسیله آن جویای علمی باشد که خود از آن آگاه است. اکنون که به یاری خداوند منان توانستم مرحله ای دیگر از زندگی را با موفقیت سپری نمایم، وظیفه خود می دانم از زحمات اساتید، بزرگواران و دوستان عزیز که همراهی و مساعدت آنان در به انجام رساندن این پایان نامه نقش بسزایی داشته، تقدیر و تشکر نمایم. بویژه از جناب آقای دکتر عبدالرسول برهانی حقیقی که افتخار راهنمایی این پایان نامه را به اینجانب دادند و با صبر و سعه صدر و دقت علمی فراوان مرا در این زمینه یاری نمودند، سپاسگزاری کنم. از اساتید گرامی جناب آقای دکتر علیرضا نعمت الهی و سرکار خانم دکتر مینا توحیدی که با پیشنهادات سازنده خود مرا در این راه یاری کردند، کمال تشکر و قدردانی را دارم. از جناب آقای دکتر محمد مهدی علیشاهی، دکتر فریبرز حیدری، دکتر امین قلمفرسا، دکتر احسان بهرامی و دکتر مهدی سلمانپور که همفکری صمیمانه آنان در جریان تحقیق، بخصوص در نوشتن برنامه های کامپیوتری بهره های بسیار را نصیب اینجانب نمودند، تقدیر و تشکر می نمایم. همچنین از تمام دانشجویان گرامی و دوستان عزیزم که مرا در طی دوران تحصیل یاری داده اند، کمال تشکر را دارم. در نهایت از کارکنان بخش آمار بویژه سرکار خانم ریحانی و خانم میرزایی و جناب آقای شیخ عطار تشکر می کنم و برای تمامی این عزیزان، آرزوی توفیق روزافزون دارم.

چکیده

بررسی سره یا ناسره بودن توزیع پسین و وجود ماکسیمم درستنمایی برای مدل های رگرسیونی با متغیرهای کمکی به طور تصادفی گمشده

به وسیله ی:

رویا نصیرزاده

در بسیاری از تحقیقات آماری با نرخ زیادی از داده های گمشده روبرو هستیم. بسیاری از آماردانان در تحقیقات خود، افراد با داده های گمشده را کنار می گذارند و فقط به نتیجه گیری با استفاده از داده های کامل می پردازند. این پایان نامه به بررسی داده های همه افراد، زمانی که داده ها به صورت تصادفی گمشده اند، می پردازد. در این پایان نامه آنالیز و تحلیل داده های گمشده از طریق آنالیز بیز انجام می شود و مدل رگرسیونی تعمیم یافته برای مدل بندی داده ها مورد استفاده قرار گرفته است. در ابتدا سره بودن توزیع پسین ضرایب رگرسیونی تحت مطالعه قرار می گیرد. سپس وجود ماکسیمم درستنمایی با ارائه یک قضیه مورد بررسی قرار می گیرد و در نهایت با استفاده از دو مثال به برآوردیابی ضرایب رگرسیونی در دو حالت استفاده از داده های کامل و داده های تمام افراد پرداخته خواهد شد و مشاهده می شود زمانی که از داده های تمام افراد استفاده می شود، برآورد بهتری برای ضرایب رگرسیونی بدست می آید. بنابراین این تحقیق مثبت بودن اثر داده های گمشده در مطالعه را نتیجه می دهد.

فهرست مطالب

عنوان ..... **صفحه**

فهرست جدول ها

فصل اول:
مقدمه و مروری بر مفاهیم پایه

فصل اول: مقدمه و مروری بر مفاهیم پایه

۱ – ۱ – مقدمه و تعاریف

متغیرهای کمکی گمشده در بیشتر تحقیقات آماری از جمله در بررسی های پزشکی – اجتماعی، مطالعه محیط زیست و مطالعات اقتصادی– اجتماعی دیده می شوند. بررسی مجموعه داده بزرگ، اغلب با نرخ زیادی از داده های گمشده همراه است که وجود این داده های گمشده ممکن است در تجزیه و تحلیل داده ها مشکل های جدی به وجود آورند.

در این پایان نامه نشان می دهیم که در آمار بیز ممکن است توزیع پسین ضرایب رگرسیونی برای داده های کامل ناسره باشد ولی برای داده های تمام افراد سره باشد. همچنین برآورد ضرایب رگرسیونی برای داده های تمام افراد بهتر از داده های کامل می باشد. توجه داشته باشید که این رویداد برای مجموعه داده های کم زمانی که نسبت داده های گمشده زیادند، امری کاملاً طبیعی است.

تعریف ۱ – ۱ – ۱:

یک مجموعه داده که تمام افراد و یا موضوع ها بانضمام داده هایی که بعضی از مولفه های آن ها، داده گمشده اند را **داده های تمام افراد**[1] گویند. دراین داده ها، به مجموعه داده هایی که تمام مولفه های آن ها مشاهده شده باشند **داده های کامل**[2] گویند.

مثال ۱ – ۱ – ۱:

فرض کنید که داده های یک مطالعه به صورت داده های جدول ۱ باشد.

---

1 All of the cases

2 Complete cases

i	x_{i1}	x_{i2}	x_{i3}	y_i
۱	۱	۰	۰	۰
۲	۱	۱	۰	۱
۳	۱	۰	۱	۰
۴	۱	۰	۱	۱
۵	۱	۰/۵	-	۱
۶	۱	۱	-	۱
۷	۱	-۱	-	۰
۸	۱	۱/۵	-	۰

جدول (۱)

آنگاه به کل داده های این جدول، داده های تمام افراد گویند و داده های کامل این مجموعه داده های زیر می باشند

i	x_{i1}	x_{i2}	x_{i3}	y_i
۱	۱	۰	۰	۰
۲	۱	۱	۰	۱
۳	۱	۰	۱	۰
۴	۱	۰	۱	۱

جدول (۲)

تعریف ۲ - ۱ - ۱:

گوئیم یک توزیع پیشین یا پسین **سره**[3] است، هرگاه این توزیع خصوصیات تابع چگالی را داشته باشد. در غیر اینصورت، گوئیم این توزیع **ناسره**[4] است.

بـه طور مثال، اگر θ دارای توزیع یکنواخت روی مجموعه اعداد حقیقی باشد، یعنی:

$$\pi(\theta) \propto 1, \quad \theta \in \Re.$$

آنگاه شرط برابر با یک بودن انتگرال تابع چگالی برقرار نمی باشد. یعنی:

$$\int_{-\infty}^{\infty} \pi(\theta)\, d\theta = \int_{-\infty}^{+\infty} d\theta = \infty$$

و می گوییم θ دارای توزیع یکنواخت ناسره می باشد.

---

[3] proper
[4] improper

نکته ۱ – ۱ – ۱:

اگر $\pi(\theta)$ توزیع پیشین برای θ و $\pi(\theta|y)$ توزیع پسین برای θ باشد. آنگاه $\pi(\theta|y)$ سره است اگر و تنها اگر

$$\int L(\theta|y)\,\pi(\theta)\,d\theta < \infty.$$

تعریف ۳ – ۱ – ۱:

الف – (بردار مثبت): بردار $\boldsymbol{v}=(v_1,\cdots,v_n)'$ را **مثبت** گوییم، اگر تمام مولفه های آن مقدار مثبت باشند. یعنی:

$$v_i > 0, \qquad \forall i=1,\cdots,n.$$

ب: بردار $\boldsymbol{v}=(v_1,\cdots,v_n)'$ از بردار $\boldsymbol{u}=(u_1,\ldots,u_n)'$ **بزرگتر** است، اگر:

$$v_i > u_i, \qquad \forall\, i=1,\cdots,n.$$

۲ – ۱ – مروری بر تحقیقات گذشته

نویسندگان متعددی، ویژگی توزیع پسین را مورد مطالعه قرار داده اند، به طور مثال، ابراهیم[5] و لاود[6] (۱۹۹۱) که شرط لازم و کافی را برای برازندگی توزیع پسین $\boldsymbol{\beta}$ روی کلاس مدل های رگرسیونی تعمیم یافته با استفاده از توزیع پیشین جفری[7] بیان کردند، مکلاخ[8] و ناتاراجان[9] (۱۹۹۵) شرط لازم و کافی را برای برازندگی توزیع پسین روی مدل های رگرسیونی آمیخته با متغیرهای پاسخ دوتایی بیان کردند، چن[10] و کسلا[11] و هوبرت[12] (۱۹۹۶) اثر توزیع پسین ناسره را مورد مطالعه قرار داده اند، گلفند[13] و ساها[14] (۱۹۹۹)، ابراهیم و یانوتسوس[15] (۱۹۹۹) وچن و شاو[16] (۲۰۰۱) نیز روی توزیع پسین $\boldsymbol{\beta}$ کار کرده اند، اما همه مطالعات روی داده های کامل انجام گرفته است. این پایان نامه به بررسی داده های تمام افراد می پردازد.

---

5 Ibrahim
6 laud
7 Jeffrey
8 McCulloch
9 Natarajan
10 Chen
11 Casella
12 Hobert
13 Gelfand
14 Sahu
15 Yiannoutsos
16 Shao

همچنین نویسندگان زیادی به بررسی وجود ماکسیمم درستنمایی روی مدل های رگرسیونی خطی پرداخته اند از جمله: هابرمن[17](۱۹۷۴)، شرط لازم وکافی را برای وجود ماکسیمم درستنمایی در مدل های لگ – خطی بیان کردند. سیلواپول[18] (۱۹۸۱) شرط لازم و کافی را برای وجود ماکسیمم درستنمایی در مدل های رگرسیونی دوتایی بیان کرد. ودربرن [19](۱۹۷۶)، آلبرت[20] و اندرسن[21] (۱۹۸۴)، مکلاخ وناتارجان (۱۹۹۵) نیز به بررسی وجود ماکسیمم درستنمایی پرداخته اند، در اینجا نیز همه مطالعات روی داده های کامل انجام گرفته است. برای بررسی این مهم، در این پایان نامه قضیه ای برای داده های تمام افراد ارائه می گردد.

۳ – ۱ – مکانیسم داده های گمشده

مکانیسم داده های گمشده از اهمیت زیادی برخوردار است. مطالعه داده های گمشده به طبیعت مقادیر وابسته در این مکانیسم وابسته است. فرض کنید که ماتریس $\mathbf{X}$، شامل تمام متغیرهای کمکی باشد. بنابراین می توان آن را به صورت $\boldsymbol{X}=(X_o,X_m)$ نمایش داد، که در آن X_o نشان دهنده مقادیر مشاهده شده و X_m نشان دهنده مقادیر گمشده است. .همچنین فرض کنید که ϕ مجموعه پارامترهای مجهول و $\boldsymbol{M}=(m_{ij})$ نشان دهنده ماتریسی باشد که وقتی x_{ij} مشاهده شده باشد مقدار صفر و در غیر این صورت مقدار یک را بپذیرد، آنگاه مکانیسم داده های گمشده به صورت $f(\boldsymbol{M}|\boldsymbol{X},\phi)$ می باشد. مکانیسم داده های گمشده می تواند به یکی از سه حالت زیر باشد:

الف – گمشدگی کاملاً تصادفی[22] (MCAR):

در این حالت، احتمال این که یک مشاهده گمشده باشد، به مقدار متغیرهای مشاهده شده و گمشده وابسته نمی باشد. در این حالت مقادیر مشاهده شده بطور موثر تشکیل یک نمونه تصادفی ساده از کلیه مقادیر حاصل از مطالعه تمام افراد می دهند. این نوع مکانیسم را به صورت زیر می توان نوشت:

برای هر X و فضای پارامتری ϕ: $$f(\boldsymbol{M}|\boldsymbol{X},\phi)=f(\boldsymbol{M}|\phi),$$

17 aberman
18 Silvapulle
19 Wedderburn
20 Albert
21 Anderson
22 Missing Completely at random

به مکانیسم بالا، گمشدگی کاملاً تصادفی می گویند و به طور خلاصه به صورت MCAR نمایش می دهند. در بررسی های نمونه ای به این مکانیسم، مکانیسم بدون پاسخ یکنواخت[۲۳] نیز می گویند.

هرچند کنار گذاشتن داده های گمشده باعث از دست دادن یک سری از اطلاعات می شود، اما زمانی که مکانیسم داده ها به صورت MCAR باشد، نتایج آنالیز داده های کامل و داده های تمام افراد، مشابه به هم است و بنابراین در این گونه مکانیسم ها، به آنالیز داده های کامل می پردازیم.

به طور مثال زمانی که یک نمونه آزمایشگاهی از بین رود، آنگاه نتیجه مشاهده گمشده کاملاً تصادفی به حساب می آید.

در بسیاری از موارد به نظر می آید که مکانیسم داده ها به صورت MCAR است اما در اصل MCAR نیست. به طور مثال اگر بیماری که تحت درمان است، بر اثر تصادف بمیرد، به نظر می رسد که به طور کاملاً تصادفی گمشده است. اما اگر همین بیمار تحت درمان روانپزشکی باشد، آنگاه مرگ آن بیمار بر اثر خودکشی بوده و این ناشی از ضعف روش درمانی است و بنابراین، این گمشدگی نمی تواند به صورت کاملاً تصادفی باشد.

ب - گمشدگی تصادفی[۲۴] (MAR):

فرض کنید، احتمال اینکه یک مشاهده گمشده باشد، به مقدار متغیرهای گمشده وابسته نباشد، یعنی گمشدگی تنها به X_o وابسته است و به X_m وابسته نباشد. این نوع مکانیسم را می توان به صورت زیر نوشت:

برای هر X_m و فضای پارامتری ϕ

$$f(M \mid X, \phi) = f(M \mid X_o, \phi),$$

به مکانیسم بالا، گمشدگی تصادفی گویند. این نوع گمشدگی به طور خلاصه به صورت MAR نمایش داده می شود

به طور مثال، اگر در حال انجام یک آزمایش، دستگاه خراب شود، آنگاه این گمشدگی به صورت MAR است.

توجه داشته باشید که در این نوع مکانیسم، اگر مقادیر مشاهده شده دو واحد برابر باشد، آنگاه آن دو واحد از نظر آماری دارای رفتارهای یکسانی می باشند (زمانی که متغیر، گمشده یا مشاهده باشد). به عنوان مثال اگر داده های ۲ واحد در یک مطالعه به صورت زیر باشد

---

[23] Uniform nonresponse
[24] Missing at random

	متغیرها					
واحد	۱	۲	۳	۴	۵	۶
۱	۱	۳	۴.۳	۳.۵	۱	۴.۶
۲	۱	۳	-	۳.۵	-	-

منظور از "-" ، داده های مشاهده نشده است.

جدول (۳)

همانطور که مشاهده می کنید، متغیرهای مشاهده شده برای هر دو واحد دارای مقادیر یکسانی هستند. تحت MAR متغیرهای ۳، ۵ و ۶ در واحد دوم با متغیرهای ۳، ۵ و ۶ در واحد اول همتوزیع می باشند. توجه داشته باشید که نمی توان گفت، این مقادیر برابر هستند.

ج- گمشدگی غیر تصادفی[۲۵] (MNAR):

فرض کنید مکانیسم داده ها به صورت MAR یا MCAR نباشد، یعنی:

برای فضای پارامتری ϕ

$$f(\boldsymbol{M} \mid \boldsymbol{X}, \phi) = f(\boldsymbol{M} \mid X_o, X_m, \phi).$$

همان طور که مشاهده می کنید در این حالت گمشدگی به X_o و X_m وابسته است. در این حالت گفته می شود گمشدگی به صورت غیر تصادفی (غیر قابل اغماض) می باشد. این نوع گمشدگی به طور خلاصه به صورت MNAR نمایش داده می شود.

زمانی که به آنالیز داده ها با استفاده از روش تابع درستنمایی می پردازیم. این نوع گمشدگی بیشترین مساله را در آنالیز داده های گمشده سبب می شود. (برای اطلاع بیشتر می توان به منابع [۴] و [۲۵] رجوع کرد.)

توجه داشته باشیدکه مکانیسم مورد مطالعه در این پایان نامه به صورت MAR است.

۴ – ۱ – معرفی داده های تحت مطالعه

در این پایان نامه، سه دسته داده را تحت مطالعه قرار خواهیم داد. دسته داده اول که با نام مثال ۱ - ۱ - ۱، معرفی کردیم، داده هایی مصنوعی هستند. اما دو دسته داده دیگر، داده های واقعی هستند که در این بخش به معرفی آن ها می پردازیم:

---

[25] Missing Not at random

۱ - ۴ - ۱ - داده های سلامت تنفسی (شش شهر):

در سال ۱۹۸۴، فریس[۲۶]، ویر[۲۷]، داکری[۲۸]، اسپیرو[۲۹] و اسپیزر[۳۰] به مطالعه سلامت تنفسی در ۲۳۹۴ کودک ۱۱ ساله پرداختند، که این مطالعه در شش شهر انجام شده و به همین دلیل بـه داده های شش شهر مشهور است. نتایج این مطالعه در جدول (۴) آمده است.

در این جدول x_{i1}، نشان دهنده عرض از مبدا در مدل رگرسیونی است که همواره برابر با یک است. x_{i2}، نوع شهر محل اقامت است. اگر کودک در شهر پر جمعیت زندگی کنند، x_{i2} مقـدار یک و در غیر این صورت مقدار صفر را می گیرد. اگر z_{i3} را تعداد دفعات سیگار کشیدن مادران آن ها در یک روز در نظر بگیریم، آنگاه x_{i3} به صورت زیر ساخته می شود:

$$x_{i3}=\begin{cases}۰, & z_{i3}=۰\\ ۱, & z_{i3}=۱\\ ۲, & z_{i3}\geq ۲\end{cases}$$

و y_i نشاندهنده خس خس کردن کودکان است. اگر i امین بچه خس خـس کنـد، y_i مقـدار یک و در غیر این صورت، مقدار صفر را می گیرد.

متغیر x_{i1} همواره مشاهده شده است، متغیر x_{i2} در ٪ ۳۲/ ۸ و متغیر x_{i3} در ٪ ۳/۳ از موردهـا گمشده است و به طور کلی ٪ ۱/ ۳۵ از موردها دارای مولفه گمشده می باشند.

برای $[y_i \mid x_{i1}, x_{i2}, x_{i3}]$ یک مدل لجستیک به صورت زیر در نظر می گیریم:

$$f(y_i \mid x_{i1}, x_{i2}, x_{i3})=\frac{\exp[y_i(\beta_1+\beta_2 x_{i2}+\beta_3 x_{i3})]}{[1+\exp(\beta_1+\beta_2 x_{i2}+\beta_3 x_{i3})]}.$$

که در آن $\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3)'$، بردار ضرایب رگرسیونی است. برای (x_{i2},x_{i3}) توزیع توأم را بـه فرم $[x_{i2},x_{i3}]=[x_{i2}][x_{i3}\mid x_{i2}]$ در نظر می گیریم، فرض می کنیم $[x_{i2} \mid \alpha_{2.}]$ ها، متغیرهـای تصادفی مستقل همتوزیع برنولی با احتمال موفقیت $\exp(\alpha_{2.})/[1+\exp(\alpha_{2.})]$ باشـند کـه در آن $\alpha_{2.}$ پـارامتری مجهـول و شـاخص ایـن توزیـع اسـت. همچنـین فـرض مـی کنـیم کـه $[z_{i3} \mid x_{i2}, a_{2.}]$ ها، متغیرهای تصادفی مستقل همتوزیع پواسن بـا میـانگین $\exp(\alpha_{2.}+\alpha_{21}x_{i2})$ باشند که در آن $\boldsymbol{\alpha}_2=(\alpha_{2.},\ \alpha_{21})'$ بردار پارامترهای مجهول این توزیع است.

همچنین فرض می کنیم که همه ضرایب رگرسیونی و پارامترهـای شـاخص، مـستقل از هـم و دارای توزیع پیشین یکنواخت روی $\Re$ باشند.

---

[۲۶] Ferris
[۲۷] ware
[۲۸] Dockery
[۲۹] Spiro
[۳۰] Speizer

y_i	x_{i1}	x_{i2}	x_{i3}	تعداد
۰	۱	۰	۰	۴۱۸
۰	۱	۰	۲	۲۱۷
۰	۱	۱	۱	۳
۰	۱	-	-	۱۸
۰	۱	-	۱	۶
۰	۱	۱	-	۲۴
۱	۱	۰	۰	۱۲۷
۱	۱	۰	۲	۷۱
۱	۱	۱	۱	۲
۱	۱	-	-	۸
۰	۱	-	۲	۲۲۳
۱	۱	۱	-	۱۰
۰	۱	۰	۱	۹
۰	۱	۱	۰	۳۲۳
۰	۱	۱	۲	۱۹۸
۰	۱	-	۰	۳۶۹
۱	۱	-	۲	۷۵
۱	۱	۰	۱	۱
۱	۱	۱	۰	۱۰۶
۱	۱	۱	۲	۸۱
۱	۱	-	۰	۸۶
۰	۱	۰	-	۱۹

جدول (۴): داده های سلامت تنفسی

منظور از "-" ، داده های مشاهده نشده است

۲ – ۴ – ۱ – داده های مرحله III سرطان پوست

یکی دیگر از داده های تحت مطالعه، داده های مرحله III سرطان پوست است. که به وسیله گروه ECOG[31] جمع آوری شده است. این مطالعه شامل یک نمونه تصادفی ۲۸۶ بیمار می باشد که به طور تصادفی درون یکی از دو دسته درمانی قرار گرفته اند. دسته اول درمان به وسیله انترفرون[32] (IFN_α - 2b) و دسته دوم شامل درمان خودکار(بدون درمان دارویی) می باشد. که خلاصه ای از نتایج در جدول(۵) آمده است.

---

[31] Eastern Cooperative Oncology Group

[32] **interferon alfa - 2b** ، فعالیت های آنتی تومور را در گسترش سرطان پوست نشان می دهد و براین اساس به عنوان ماده محرک کمک درمانی در جراحی به کار می رود.

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	δ_i	y_i
۱	۱	۲/۳۰۳	۱	۱	۰	۹/۶۳۰
۲	۱	۱/۲۰۹	۱	۰	۱	./۳۴۸
۳	۱	۱/۳۳۵	۱	۱	۱	۰/۴۷۹
۴	۱	-۰/۷۷۷	۰	۰	۱	۲/۱۳۱
۵	۱	./۵۸۸	۰	۱	۰	۷/۸۳۶
۶	۱	۱/۳۴۳	۰	۰	۱	۲/۷۲۳
۷	۱	۱/۵۵۸	۱	۱	۰	۸/۴۵۷
۸	۱	-۰/۶۷۳	۰	۱	۱	۰/۵۵۱
۹	۱	./۷۸۸	۱	۰	۰	۸/۲۴۶
۱۰	۱	-./۷۱۳	۰	۰	۰	۸/۰۲۷

جدول (۵): داده های سرطان پوست

در این جدول x_{i1}، نشان دهنده عرض از مبدا در مدل رگرسیونی است که همواره برابر با یک است. x_{i2}، لگاریتم ضخامت برسلو[۳۳] بر حسب میلی متر و x_{i3}، سطح ابتدائی است که مقدار یک را زمانی که گسترش سرطان، سطحی و ظاهری است و در غیر این صورت مقدار صفر را می پذیرد. x_{i4}، نوع درمان است که مقدار یک را برای درمان انترفرون و مقدار صفر را برای مشاهدات بدون دارو قرار می دهیم. x_{i2} و x_{i3} داری مقادیر گمشده می باشند و x_{i4} برای همه موارد، مشاهده شده است. استفاده از لگاریتم در ضخامت برسلو، به این دلیل است که به یک تقریب نرمال برای متغیرهای کمکی پیوسته برسیم. این دو متغیر روی هم رفته در ٪ ۱۷/۶ از موردها، دارای عضو گمشده می باشند. متغیر پاسخ ، y_i، زمان رهایی از بیماری بر حسب سال می باشد که پیوسته و از سمت راست سانسور شده می باشد. δ_i، شاخص سانسور بودن می باشد که اگر i امین مورد از بیماری رهایی یابد، برابر با یک و در غیر این صورت، برابر با صفر می باشد.

برای داده های از سمت راست سانسور شده $[y_i \mid x_i, \beta]$ یک مدل رگرسیونی نمایی به صورت زیر در نظر می گیریم:

$$f(y_i \mid \delta_i, x_i, \beta) = \exp\{\delta_i (x_i' \beta)\} \exp\{-y_i \exp(x_i' \beta)\}, \quad (۱.۴.۲.۱)$$

---

[۳۳] **Breslow thickness**: ضخامت برسلو؛ یک روش پیشگویی از روی علایم پوست است. اولین بار الکساندر برسلو از این شاخص استفاده کرد، وی مشاهده کرد که وقتی ضخامت تومور افزایش پیدا می کند، شانس زنده بودن کاهش می یابد. به طور مثال وقتی که ضخامت پوست کمتر از mm ۷۶ / ۰ باشد، شانس زنده بودن تا پنج سال بعد، برابر با ٪۹۷ است و وقتی بیشتر از mm ۸ باشد، شانس به ٪۳۲ کاهش می یابد.

فرض کنید، به ازای $i=1,\cdots,n$ ، $\boldsymbol{x}_i$ یک بردار ۱×۴ به صورت زیر باشد:

$$\boldsymbol{x}_i=(x_{i1},x_{i2},x_{i3},x_{i4})'$$

و بردار ضرایب رگرسیونی به صورت $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_4)'$ باشد.

به دلیل اینکه به ازای همه i ها x_{i1} و x_{i4} مشاهده شده می باشند و تنها x_{i2} و x_{i3} دارای عضو گمشده هستند، بنابراین نیازی به تعیین مدل برای x_{i1} و x_{i4} نداریم. در نتیجه توزیع متغیرهای کمکی به صورت زیر می باشد:

$$f(x_{i2},x_{i3}\mid x_{i4},\boldsymbol{\alpha})=f(x_{i3}\mid x_{i2},x_{i4},\boldsymbol{\alpha}_3)\,f(x_{i2}\mid x_{i4},\boldsymbol{\alpha}_2),\qquad \forall i=1,\cdots,n.$$

همچنین یک توزیع نرمال برای $f(x_{i2}\mid x_{i4},\boldsymbol{\alpha}_2)$ در نظر می گیریم. یعنی:

$$(x_{i2}\mid x_{i4},\ \boldsymbol{\alpha}_2)\ \sim\ N(\mu_1+\mu_2 x_{i4}\ ,\sigma^2)\,, \qquad (۱.۴.۲.۲)$$

زمانی که $\boldsymbol{\alpha}_2=(\mu_1,\mu_2,\sigma^2)'$ بردار پارامترهای مجهول برای این توزیع باشد.

همچنین فرض کنید که، $[x_{i3}\mid x_{i2},\ x_{i4}]$ دارای مدل رگرسیون لجستیک به صورت زیر باشد

$$f(x_{i3}\mid x_{i2},x_{i4},\ \boldsymbol{\alpha}_3)=\frac{\exp\{x_{i3}(\alpha_{30}+\alpha_{31}x_{i2}+\alpha_{32}x_{i4})\}}{1+\exp(\alpha_{30}+\alpha_{31}x_{i2}+\alpha_{32}x_{i4})}\,, \qquad (۱.۴.۲.۳)$$

که در آن $\boldsymbol{\alpha}_3=(\alpha_{30},\alpha_{31},\alpha_{32})'$ بردار پارامترهای مجهول برای این توزیع باشد.

همچنین فرض کنید که:

$$\pi(\boldsymbol{\beta}\,,\boldsymbol{\alpha})\equiv\pi(\,\boldsymbol{\beta}\,,\mu_1,\mu_2,\,\sigma^2,\,\boldsymbol{\alpha}_3)\ \propto\ \frac{1}{\sigma^2}. \qquad (۱.۴.۲.۴)$$