۸ /۱۰/ ۱۳۲۸

بسم ا... الرحمن الرحیم

استنباطهایی برای داده های بریده شده

بوسیله

حسین ریاض الشمس

پایان نامه

ارائه شده به دانشکده تحصیلات تکمیلی به عنوان بخشی از فعالیتهای
تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته

آمار

از

دانشگاه شیراز

شیراز، ایران

ارزیابی و تصویب شده توسط کمیته پایان نامه با درجه:

امضاء اعضاء کمیته پایان نامه:

دکتر فریبرز حیدری، استاد یار بخش آمار(رئیس کمیته) ....

دکتر عبدالرسول برهانی، استادیار بخش آمار ....

دکتر احمدرضا سلطانی، استاد بخش آمار ....

مهر ۷۸

تقدیم به:

مادر مهربانم

سپاسگزاری

با سپاس و درود خداوندی را که جان را فکرت آموخت و روان را با دم وجودش زندگی بخشید تا آدمیان ناچیز خاکی را تا به مرتبه اعلی فکرت رساند که کائنات هستی را قدرت ادراک آدمی جایگاهش باشد. در این برهه از زمان که قدرت ادراک آدمی مرز زمان و مکان را در نوردیده چه خوش است آنان را که درجهت تعقل ره پویند. و چه بزرگی، شکیبایی، پشتکار و از خود گذشتگی است افرادی چونان دکتر فریبرز حیدری را که شیوه تعقل را پایه گزارند. و زبان مرا یارای قدردانی از چنان استاد فرزانه نباشد. همانقدر سپاس دارم که توانم که نه تنها علم آمار، که سخاوت، انسانیت، محسنات و تعقل را وارث ایشان باشم. سپاس دارم از اساتید فرزانه جناب دکتر احمدرضا سلطانی و دکتر عبدالرسول برهانی که در امر نگارش پایان نامه مرا یاری نمودند. وسپاس دارم یگانه مرد بزرگوار پروفسور جواد بهبودیان استاد مشاور بنده. و از کلیه اساتید که در دوران تحصیل چراغ بر افروخته شان روشنی بخش راه من بوده.

چکیده

استنباطهایی برای داده های بریده شده

توسط

حسین ریاض الشمس

در سانسور تصادفی دیده شده که براورد درستنمایی ماکزیمم، (MLE) منحنـی بقاء از فرض پارامتری بودن توزیـع متغـیر سانسـور تـاثیر نمـی پذیـرد. بـرآورد روش کاپلان مایر در ۱۹۵۸ یک برآورد MLE هم برای هردو مدل نا پارامتری و هـم مـدل نیمه پارامتری می باشد. در داده های تصادفی بریده شده برآورد حـدی ضربـی کـه توسط لیندل بل در ۱۹۷۱ ارائه شده برآوردی MLE برای مدل نا پارامتری اسـت. و آن مدل نیمه پارامتری را که در آن مکانیزم برش پارامتری باشد را در بر نمی گیرد.

فرض کنید X,Y دومتغیر تصادفی مثبت باشند. متغیر تصادفی X ، (متغیر مورد بررسی) از چپ بوسیله Y ، (متغیر برش) بریده شده (یا Y از راست بوسیله X بریده شده) اگـر زوج (X,Y) وقتـی مشـاهده شـود کـه X>Y، چنانچـه توزیـع Y از یـک خانواده پارامتری معلوم باشد مدل نیمه پارامتری نامیده می شود و چنانچه توزیـع Y کاملا معلوم باشـد مـدل ناپـارامتری اسـت. بریدگـی در بسـیاری از مطالعـات سـتاره شناسی، همه گیرشناسی، پزشکی، علوم اجتماعی، جرم شناسی وغیره، رخ می دهـد. برای برآورد منحنی بقاء روشـهایی ازقبیـل ناپـارامتری، نیمـه پـارامتری، رگرسـیونی، فرایندهای مارکف ارائه شده است. بسیاری از تحقیقات بر اساس ماکزیمم کردن تابع درستنمایی بنا نهاده شده و مانیز به طور گسترده چنیـن روشـهایی را بکـار خواهیـم برد. همچنین فرض مشخص بودن تابع توزیع متغیر برش برآوردهای بهتری از حـالت نا مشخص تابع توزیع بدست می دهد. اگر متغیر برش از خـانواده پـارامتری باشـد در روش درستنمایی ماکزیمم پارامترها را نیز برآورد می کنیم.

فهرست مطالب

عنوان | صفحه

عنوان صفحه

۱-فصل اول
بریدگی

۱- ۱-مقدمه و خلاصه

تجزیه و تحلیل داده های بریده شده در سالهای اخیر یک موضوع کاربردی همراه با پشتوانه قوی ریاضی به صورت یک ابزار بوسیله آماردانان کاربردی در آمده. البته مباحثی که جدیدتر از آن تحت عنوان سانسور تاخیری بوسیله کوپاس و حیدری ۱۹۹۷ مطرح شده است بریدگی را نیز در بر می گیرد.

هدف این فصل معرفی بریدگی و داده های بریده شده است. همچنین طرح مسائلی که در آنالیز بقاء مطرح است.

۱- ۲-تاریخچه

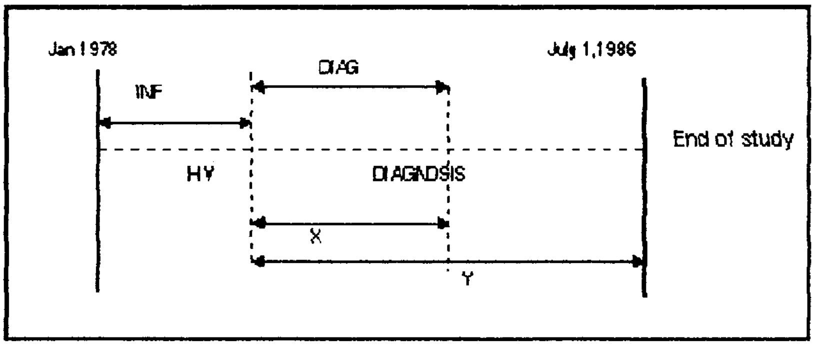
بریدگی اولین بار توسط لیندن بل (۱۹۷۱)در یک بررسی نجومی معرفی شد که البته می توان آنرا به عنوان روش ناپارامتری در نظر گرفت. اولین بحث تئوری بر اساس نسبت درستنمایی ناپارامتری را آوین (۱۹۹۰ ، ۱۹۸۸) ارائه کرده، و از آن به بعد نظرات را متوجه خود نمود. براورد ناپارامتری درستنایی ماکزیمم برای بقاء بوسیله گانگ لی (۱۹۹۵) بحث شده و برآورد احتمال بقاء و احتمال برش در حالت نیمه پارامتری بوسیله مئی چانگ وانگ (۱۹۸۹) و فاصله اطمینان برای احتمال بقاء و احتمال برش بوسیله گانگ لی و همکاران(۱۹۹۷) بدست آمده.

۱- ۳-تعاریف

فرض کنید Y, X دو متغیر تصادفی مثبت باشند که به ترتیب نشان دهنده طول عمر (X) و زمان برش (Y) مربوط به یک موضوع مورد بررسی باشند. زمان طول عمر X را میگوییم از چپ بوسیله Y بریده شده (بنابراین Y از راست بوسیله X بریده شده) اگر زوج (X, Y) تنها وقتی قابل مشاهده باشند که $X>Y$. ویا میگوییم X از راست توسط Y بریده شده اگر زوج (X, Y) تنها وقتی قابل مشاهده باشند که $X<Y$.

بریدگی در بسیاری از موضوعها ممکن است رخ دهد مانند: ستاره شناسی، همه گیر شناسی، قابلیت اعتماد، مطالعات پزشکی، علوم اجتماعی، جرم شناسی، اقتصاد و غیره. به عنوان نمونه در مطالعات پزشکی وقتی می خواهیم طول زمان بقاء را بعد از

حمله یک ویروس به بدن بررسی کنیم، اگر X بیانگر زمان بین حمله ویروس به بدن تا مرگ باشد و زمان دنبال کردن بیماری Y واحد زمانی بعد از شروع حمله ویروس شروع شده باشد در اینجا به وضوح X از چپ بوسیله Y بریده شده است (X>Y) .
به عنوان مثال داده های مربوط به انتقال خون که به وسیله مرکز کنترل بیماریها در آتلانتا توسط کالب فلیش و لاولس در ۱۹۸۹ ارائه شده را در نظر بگیرید. متغیر مورد بررسی زمان کمون می باشد. متغیر مورد بررسی زمان کمون یعنی طول زمان از حمله ویروس HIV تا زمان تشخیص بیماری (X) میباشد. متغیر برش (Y) زمان از حمله ویروس به بدن تا پایان دوره مطالعه یعنی (۱ژولای ۱۹۸۶) میباشد. به انگاره زیر توجه کنید.



X=DIAG

$INF+Y=103 \Rightarrow Y=103-INF$

در اینجا داده هایی قابل مشاهده هستند که در آنها X<Y یعنی X زمان کمون از راست بوسیله Y زمان برش بریده شده.

۱-۴-طرح مسئله

اکنون فرض کنید G, F به ترتیب توابع توزیع غیر شرطی Y, X باشند. تابع

$$R(x)=1-F(x) \qquad (۱-۱)$$

در آنالیز بقاء به عنوان تابع بقاء و در قابلیت اعتماد به عنوان تابع قابلیت اعتماد یک سیستم شناخته می شود و برآورد این تابع همیشه به عنوان مسئله اصلی مطرح است. در واقع این تابع در آنالیز بقاء برابر احتمال اینست که طول عمر یک مولفه بعد از شروع بیماری بیشتر از x واحد زمانی باشد و در قابلیت اعتماد برابر احتمال اینست که سیستم بیشتر از یک زمان معین x کار کند.

همچنین تابع

$$\nu=p(X<Y) \qquad (۲-۱)$$

احتمال برش نام دارد. معیار ν در کاربرد نقش مهمی جهت دانستن درصد موارد مشاهده نشده دارد.

در فصلهای آتی براوردهایی برای ν ,$R(x)$ (یا $F(x)$)، همچنین فاصله اطمینان برای این معیارها ارائه خواهد شد.

۲-فصل دوم
مدلهای ریاضی

۲- ۱-مقدمه و خلاصه

فرض کنید زمان عمر X بوسیله Y از چپ بریده شده باشد. درحقیقت زوج مشاهده همتوزیع و مستقل $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ از تابع درستنمایی شرطی $L((X, Y)| X>Y)$ بدست آمده است. یعنی توزیع شرطی (X, Y) به شرط $X>Y$. مسئله برآورد احتمال برش $v=P(X<Y)$ و تابع بقاء $P(X\geq x)$ (ویا به عوض آن تابع توزیع $P(X\leq x)$) میباشد. فرض کنید G, F_0 به ترتیب توابع توزیع $Y,\ X$ باشند. همانطور که وانگ در ۱۹۸۹ بیان داشته F_0 با شرط

$$\inf\{t;G(t)>0\} > \inf\{t;F_0(t)>0\}$$

قابل برآورد نیست. در این حالت ما در حقیقت تابع توزیع شرطی زیر را برآورد می کنیم.

$$P(X<x|X>c)=\frac{P(X<x\cap X>c)}{P(X>c)}=\frac{F_0(x)-F_0(c)}{1-F_0(c)} \qquad x\geq c \quad (۱-۲)$$

که در آن $c\geq \inf\{t;G(t)>0\}$.

توجه کنید که $F=F_0$ اگر G, F_0 دارای محمل یکسان باشند و

$$c=\inf\{t;G(t)>0\} \qquad (۲-۲)$$

در عمل اغلب c را برابر مقدار زیر می گیریم.

$$c=\min\{X_i, i=1 ... n\} \qquad (۳-۲)$$

برای جلوگیری از بحثهای جزئی فرض می کنیم $P(X=Y)=0$ و $P(X>Y)>0$.

وودروف در ۱۹۸۵ در بررسی یک مسئله نجومی با استفاده از توزیعهای تجربی مربوط به $X_1, X_2, ..., X_n$ و $Y_1, Y_2, ..., Y_n$ در حالی که G, F کاملاً مجهول فرض شده بودند مدلهای ناپارامتری برای برآورد G, F ارائه نمود.

ولی ما فرض را بر معلوم بودن G می گذاریم چرا که در آنالیز بقاء همانطور که در فصلهای آینده خواهیم دید می توانیم توزیع متغیر Y (G) یعنی زمان حمله بیماری تا تشخیص بیماری را بدست و یا در مثال بخش ۱-۲ توزیع زمان از حمله بیماری تا پایان مطالعه کامل مشخص است. و یک چنین فرض اضافی منجر به

برآوردهای بهتری خواهد شد تا زمانی که G کاملاً ناشناخته اســت. چـرا کـه در ایـن حالت از اطلاعات موجود در متغیر برش Y استفاده می کنیم.

در بسیاری موارد توزیع برش G متعلق به یک خانواده پارامتری

$$k=\{G(t, \theta), \theta\in\Theta\subset R^{q}\}$$

می باشد. که در آن θ می تواند پارامتر q بعدی متعلق به Θ باشد. و برای هر θ ، $G(t, \theta)$ یک تابع توزیع شناخته شده است. این حالت مدل نیمـه پـارامتری نـامیده می شود. مدل نیمه پارامتری اولین بار بوسیله وانــگ در ۱۹۸۹ معرفــی شـده و از آن برای برآورد درستنمایی ماکزیمم نیمه پارامتری نام بـرده مـی شـود. براوردهـایی کـه بوسیله ماکزیمم کردن تابع درستنمایی بدست می ایند توجه زیادی را به خـود جلـب کرده اند که ما نیز از این امکانات به طور وسیع استفاده خواهیم کرد.

به طور کلی در مورد تجزیه و تحلیل داده های بریده شده از مدلهای رگرسیونی، پارامتری، نیمه پارامتری و نا پارامتری استفاده شده است.

۲-۲-توابع توزیع، مدل ناپارامتری

فرض کنید X متغیر تصادفی طول عمر و Y متغیر برش به ترتیب با توزیعهـای $G(y), F(x)$ باشند. و $G(y)$ به پارامتر بستگی نداشته باشد.

تعاریف:

$$F(x)=P(X\le x)$$

$$G(y)=P(Y\le y)$$

$$\alpha=P(Y<X) \qquad (۲-۴)$$

$$\nu=P(X<Y) \qquad (۲-۵)$$

$$H_*(x, y)=(X\le x, Y\le y|Y<X) \qquad (۲-۶)$$

$$F_*(x)=P(X\le x|Y<X) \qquad (۲-۷)$$

$$G_*(y)=P(Y\le y|Y<X) \qquad (۲-۸)$$

G_*, F_* به ترتیب توابع توزیع کناری شرطی Y, X به شرط $Y<X$ میباشند. و $H_*(x, y)$ تابع توزیع شرطی (X, Y) به شرط $X>Y$ است.

قضیه --۱۱۲-۱ : چنانچه X, Y دو متغیر تصادفی مثبت و مستقل باشند آنگاه روابط زیر برقرار است.

$$\alpha=\int_0^{\infty} G(x)dF(x) \qquad (۹-۲)$$

$$=\int_0^{\infty} (1-F(v))dG(v) \qquad (۱۰-۲)$$

$$\nu=\int_0^{\infty} [1-G(x)]dF(x) \qquad (۱۱-۲)$$

$$F_*(x)=\alpha^{-1}\int_0^{x} G(u)dF(u)$$ توزیع: (۱۲-۲)

$$dF_*(x)=\alpha^{-1}G(x)dF(x)$$ چگالی: (۱۳-۲)

$$G_*(y)=\alpha^{-1}\int_0^{y} (1-F(v))dG(v)$$ توزیع: (۱۴-۲)

$$dG_*(y)=\alpha^{-1}(1-F(y))dG(y)$$ چگالی: (۱۵-۲)

$$H_*(x, y)=\alpha^{-1}\int_0^{x}\int_0^{y} I(v<u)dG(v)dF(u) \qquad (۱۶-۲)$$

$$=\alpha^{-1}\int_0^{y} (F(x)-F(v))dG(v)$$

$$dH_*(x, y)=\alpha^{-1}dF(x)dG(y)\, I(x>y) \qquad (۱۷-۲)$$

البته فرمول مربوط به $G_*(y)$ جهت کامل بودن مطلب در اینجا آورده شده وگرنه از این فرمول استفاده ای نخواهیم کرد.

هر کدام از عبارتهای بالا را می توان به چندین راه ثابت کرد ما در اینجا یک راه حل ساده به منظور قابل فهم بودن مطلب و همچنین یک راه حل ریاضی دقیقتر را بیان خواهیم کرد.

اثبات: راه حل اول.

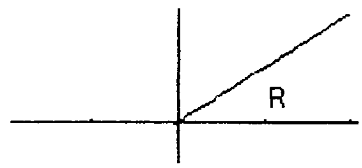
$$\alpha = P(Y < X) = \iint_R dF(x,y)dydx$$

$$= \iint_R dG(y)dF(x)$$

که در آن R ناحیه انترالگیری مطابق شکل زیر است.

$$R=\{(x, y) \mid 0\le y\le x,\ 0\le x<\infty\} \quad (۱۸\text{-}۲)$$

$$R=\{(x, y) \mid 0\le y<\infty,\ y\le x<\infty\} \quad (۱۹\text{-}۲)$$



طبق ناحیه (۲-۱۸)

$$\alpha=\int_0^{\infty}\int_0^{x} dG(x)dF(x)=\int_0^{\infty} G(x)dF(x)$$

ویا طبق ناحیه (۲-۱۹)

$$\alpha=\int_0^{\infty}\int_y^{\infty} dF(x)dG(y)=\int_0^{\infty} (1-F(y))dG(y)$$

که (۲-۹) و (۲-۱۰) ثابت شده.

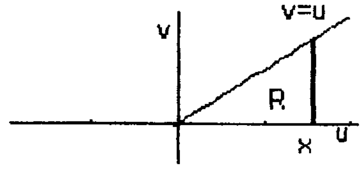
فرمول مربوط به v نیز به همین ترتیب با کمی تغییرات بدست می آید.

می توان $F_*(x)$ و $G_*(y)$ را مستقیما با استفاده از تعریف آنها بـه شـرح زیـر محاسـبه نمود.

$$F_*(x)=P(X\le x|Y<X)=\frac{P(X\le x, Y<X)}{P(Y<X)}$$
$$=\alpha^{-1}P(X\le x, Y<X)$$
$$=\alpha^{-1}\iint_R dG(v)dF(u)$$

که در آن

$$R=\{(u, v) \mid 0\le u\le x,\ 0\le v\le u\}$$



$$\Rightarrow F_*(x) = \alpha^{-1}\int_0^x\int_0^u dG(v)dF(u)$$
$$= \alpha^{-1}\int_0^x G(u)dF(u)$$

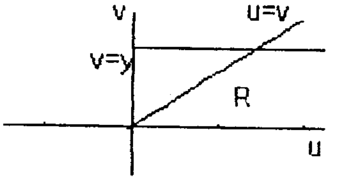
که رابطه (۲-۱۲) را ثابت می کند. رابطه (۲-۱۳) به راحتی از این رابطـه نتیجـه مـی شود.

برای رابطه (۲-۱۴) داریم.

$$G_*(y) = P(Y \le y | Y < X) = \frac{P(Y \le y, Y < X)}{P(Y < X)}$$
$$= \alpha^{-1}\iint_R dF(u)dG(v)$$

که در آن:

$$R=\{(u, v) \mid 0\le v\le y,\ \ v\le u<\infty\}$$



$$\Rightarrow\ G_*(y) = \alpha^{-1}\int_0^y\int_v^\infty dF(u)dG(v)$$
$$= \alpha^{-1}\int_0^y (1-F(v))dG(v)$$

که همان رابطه (۲-۱۴) می باشد و رابطه (۲-۱۵) به سادگی از این رابطه بدست مـی آید..

اکنون جهت محاسبه $H_*(x, y)$ داریم:

$$H_*(x,y) = P(X \le x, Y \le y | Y < X) = \frac{P(X \le x, Y \le y, Y < X)}{P(Y < X)} \quad (۲۰-۲)$$

$$P(X \le x, Y \le y, Y < X) = \iint_R dG(v)dF(u) \quad \text{و}\ (۲۱-۲)$$

با استفاده از ناحیه زیر

$$R=R_1\cup R_2$$

$$R_1=\{(u, v) \mid 0\le u\le y ,\ 0\le v\le u\}$$

$$R2=\{(u, v) \mid y\le u\le x ,\ 0\le v\le y\}$$

داریم:

$$P(X \le x, Y \le y, Y < X) = \iint_{R1} dG(v)dF(u) + \iint_{R2} dG(v)dF(u)$$

$$= \int_0^y\int_0^u dG(v)dF(u) + \int_y^x\int_0^y dG(v)dF(u)$$

$$= \int_0^y G(u)dF(u) + \int_y^x G(u)dF(u)$$

$$= \int_0^y G(\min(u, y))dF(u) \qquad (۲۲-۲)$$

به سادگی می توان عبارت احتمال (۲-۲۰) را به فرم زیر نوشت.

$$P(X\le x ,\ Y\le y ,\ Y<X) = \int_0^x\int_0^y I(v\le u)\, dF(u)dG(v) \quad (۲۳-۲)$$

$$= \int_0^y (F(x-F(v))\, dG(v) \qquad (۲۴-۲)$$

با استفاده از رابطه (۲-۲۰) و (۲-۲۳) و (۲-۲۴) فرمول (۲-۱۶) مربوط بـه $H_*(x,\ y)$ بدست می اید..

البته توجه کنید که با داشتن H_* می توان $G_* ,\ F_*$ را با توجه به رابطـه زیـر بدسـت آورد.

$$F_*(x)=P(X\le x \mid Y<X) = \lim_{x\to\infty} H_*(x, y) = H_*(x, \infty)$$

$$G_*(y)=P(Y\le y \mid Y<X) = \lim_{x\to\infty} H_*(x, y) = H_*(\infty, y)$$

وبه راحتی روابط (۱-۱۲) و (۲-۱۴) را نتیجه خواهد داد.

اثبات: راه حل دوم.

می توان با استفاده از قضیه زیر اثبات دقیقتر بر پایه ریاضی ارائه داد.

قضیه ۲-۲: اگر Y, X بردارهای مستقل با توزیعهای ν, μ در R^k, R^j باشند آنگاه:

$$P[(X,Y) \in B] = \int_{R^j} P[(x,Y) \in B]\mu(dx) \qquad : B \in \Re^{j+k} \qquad (۲۵-۲)$$

$$P[X \in A,(X,Y) \in B] = \int_A P[(x,Y) \in B]\mu(dx) \ : \ A \in \Re^j, B \in \Re^{j+k} \quad (۲۶-۲)$$