فصل اول:
مقدمات

در این فصل تعاریف و مقدمات اولیه برای مدل‌های خطی، مدل‌های خطی با خطای اندازه‌گیری، برآوردگرهای استوار به‌ویژه برآورد M، آنالیز بقا، برآوردگر کاپلان مایر، داده‌های سانسورشده و انواع سانسور ارائه می‌شود.

۱-۱- مدل خطی

یکی از کاربردی‌ترین روش‌ها برای تحلیل داده‌ها در بین ابزارهای آماری، تحلیل رگرسیونی است. تحلیل رگرسیونی،روشی کارآمد برای بررسی و مدل‌سازی ارتباط بین متغیرها است که از این مدل های رگرسیونی در توصیف داده‌ها، برآورد پارامترهای مجهول، پیش‌گویی و کنترل استفاده می شود.

در بیشتر موارد، پاسخ یک آزمایش به چندین متغیر مستقل مثلا k متغیر مستقل، وابسته است. در این صورت یک مدل خطی رابطه‌ای به صورت زیر را در نظر می‌گیرد:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (1.1.1)$$

که n اندازه نمونه می‌باشد. متغیرهای $(x_1, \dots, x_k)$ را متغیرهای توضیحی و متغیر تصادفی قابل مشاهده y را متغیر پاسخ می‌نامند.

متغیر تصادفی غیرقابل مشاهده ε متغیر خطا تلقی می‌شود، بدین معنی که به عنوان متغیری تصادفی، اندازه ناتوانی مدل در برازش دقیق داده‌ها را اندازه‌گیری می‌کند. این خطا ممکن است به دلیل عدم حضور برخی از متغیرهای مؤثر، خطاهای تصافی مربوط به مشاهدات و اندازه‌گیری‌ها و غیره صورت پذیرد.

همچنین فرض می‌شود که خطاها دارای توزیع نرمال با میانگین صفر و واریانس نامعلوم σ^2 و ناهمبسته باشند.

پارامترهای σ^2 و $(\beta_0, \beta_1, \dots, \beta_k)$ مجهول هستند و باید با استفاده از داده‌ها برآورد شوند. فرض می-شوددادهها عبارتند از $(y_1, y_2, \dots, y_n)$ که در آن y_i پاسخ متناظر با k سطح از متغیرهای مستقل $(x_{i1}, x_{i2}, \dots, x_{ik})$ است. یعنی بنابر معادله (۱،۱،۱) می‌توان نوشت:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ , i = 1, \dots, n \qquad (2.1.1)$$

آن‌گاه هدف ما به دست آوردن برآوردهای برای $\beta_0, \beta_1, \dots, \beta_k$به ترتیب به نام‌های $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ و در نتیجه به دست آوردن رابطه زیر است.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \qquad (3.1.1)$$

که در آن $\hat{y}$ نشان دهنده مقدار برآورد شده y به ازای مقادیر $(x_1, \dots, x_k)$ است. در این صورت معادله (۳،۱،۱) به عنوان معادله پیش بینی کننده می‌تواند مورد استفاده قرار گیرد.

معمول‌ترین روش در برآورد پارامترهای یک مدل خطی، استفاده از روش "کمترین مربعات معمول (OLS)" است که روشی بسیار سودمند و کارا است.

پایه و اساس روش کمترین مربعات به Gauss و Legendre باز می‌گردد. این روش (و تعمیم‌های آن) به دلیل راحتی محاسبات و جواب‌های بسته مبتنی برآن مورد توجه بسیاری از آماردانان است.

برآوردهای $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ را به گونه‌ای برمی‌گزینیم که مجموع توان دوم انحراف‌ها را کمینه کند، یعنی آن‌ها را به گونه‌ای به دست می‌آوریم که در معادله زیر هنگامی که به ترتیب جایگزین $\beta_0, \beta_1, \dots, \beta_k$ می‌شوند، کمترین مقدار ممکن را تولید کنند.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_{21} x_{i2} - \cdots - \beta_k x_{ik})^2 \qquad (4.1.1)$$

برآوردهای $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ با مشتق گرفتن از معادله (۴,۱,۱) نسبت به $\beta_0, \beta_1, \dots, \beta_k$ و مساوی صفر قرار دادن آن‌ها به دست می‌آیند. ملاحظه می‌شود که برای حل این معادله های نرمال بهتر است که از روش ماتریسی استفاده شود. می توان رابطه (۱,۱,۱) را به فرم ماتریسی زیرر در نظر گرفت.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} x_{11} x_{12} & \dots & x_{1k} \\ x_{2o} x_{21} x_{22} & \dots & x_{2k} \\ \vdots \quad \vdots \quad \vdots & & \vdots \\ x_{n0} x_{n1} x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

بطوری‌که $x_{10} = x_{20} = \cdots = x_{n0} = 1$

فرم ماتریسی را می توان بصورت زیر نوشت.

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (5.1.1)$$

این مدل را یک مدل خطی گویند، زیرا نسبت به پارامترهای مدل، خطی است.

در این مدل خطی Y یک ماتریس $n \times 1$ X یک ماتریس $(p = k + 1)n \times p$ ، β یک ماتریس $p \times 1$ و ε یک ماتریس $n \times 1$ هستند.

آن‌گاه می‌توان معادله‌های نرمال را به صورت زیر نوشت:

$$(X'X)\boldsymbol{\beta} = X'Y \qquad (6.1.1)$$

زیرا

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon'\varepsilon} = (\boldsymbol{Y} - X\boldsymbol{\beta})'(\boldsymbol{Y} - X\boldsymbol{\beta})$$

$$= (\boldsymbol{Y'} - \boldsymbol{\beta'}X')(\boldsymbol{Y} - X\boldsymbol{\beta}) \quad = \boldsymbol{Y'Y} - \boldsymbol{Y'}X\boldsymbol{\beta} - \boldsymbol{\beta'}X'\boldsymbol{Y} + \boldsymbol{\beta'}X'X\boldsymbol{\beta}$$

چون $\beta'X'Y$ یک ماتریس 1×1 است در نتیجه با ترانهاده خود برابر است پس:

$$\boldsymbol{\beta'}X'\boldsymbol{Y} = \boldsymbol{Y'}X\boldsymbol{\beta}$$

و خواهیم داشت:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{Y'Y} - 2\,\boldsymbol{\beta'}X'\boldsymbol{Y} + \boldsymbol{\beta'}X'X\boldsymbol{\beta} \qquad (7.1.1)$$

با مشتق گرفتن از رابطه (۷,۱,۱) نسبت به بردار β و جایگزین کردن $\hat{\beta}$ به جای β و مساوی صفر قرار دادن آن، معادله‌های نرمال (۶,۱,۱) به دست می‌آید.

ماتریس‌های $X'X$ و $X'Y$ عبارتند از:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{i2} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}x_{ik} & \sum_{i=1}^{n} x_{i2}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

$$X'\boldsymbol{Y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

با فرض معکوس‌پذیر بودن ماتریس $X'X$ داریم:

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y} \qquad (8.1.1)$$

که در این صورت معادله پیش بینی کننده عبارت است از:

$$\hat{y} = \widehat{\boldsymbol{\beta}}'\boldsymbol{x} \qquad (9.1.1)$$

که در آن داریم:

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{bmatrix}$$

اما زمانی که داده پرت داشته باشیم روش کمترین مربعات معمولی جوابگو نیست، به همین دلیل به معرفی برآوردگرهای استوار می پردازیم.

۱-۲- انواع برآوردگرهای استوار

برآوردگرهای استوار برآوردهایی هستند که با استفاده از آن‌ها می‌توان حساسیت روش حداقل مربعات را نسبت به وجود داده‌های پرت کاهش داد.

برای این منظور روش کمی زیر را معرفی می‌کنیم:

می‌توان ε_i^2 را توسط تابع دیگری مانند $\rho(\varepsilon_i)$ جایگزین کرد. و با کمینه کردن $\sum_{i=1}^{n}\rho(\varepsilon_i)$ به برآوردگری استوار دست یافت. برآوردهای L_1، برآوردهای M و برآوردهای GM با این روش حاصل می‌شوند. که در این پایان نامه فقط به معرفی برآوردهای M می‌پردازیم.

تذکر. جایگزین کردن مجموع یا میانگین با کمیت‌های استوار نظیر آن‌ها مانند میانه یا میانگین پیراسته است. بر این اساس، روش‌هایی تحت عنوان LMS (کمترین میانه مربعات) یا LTS (کمترین میانگین پیراسته مربعات) معرفی شده‌اند.

۱-۲-۱- برآوردM

می‌توان در رابطه ی $S = \sum_{i=1}^{n} \varepsilon_i^2$ به جای ε_i^2 توابع دیگری مانند $\rho(\varepsilon_i)$ را قرار داد و برآوردهای پارامترها را به‌گونه‌ای یافت که کمیت زیر حاصل شود.

$$\min_{\beta} \sum_{i=1}^{n} \rho(\varepsilon_i) \qquad (1.2.1)$$

که ρ یک تابع حقیقی با ویژگی‌های زیر است:

الف. $\rho(0) = 0$

ب. تابع ρ متقارن است.

ج. تابع ρ پیوسته است.

د. اگر $0 \leq \varepsilon_i \leq \varepsilon_j$ آنگاه $\rho(\varepsilon_i) \leq \rho(\varepsilon_j)$ است.

ه. فرض کنید $a = sup\, \rho(\varepsilon_i)$ باشد، آنگاه $0 < a < \infty$ است.

و. اگر $0 \leq \varepsilon_i \leq \varepsilon_j$ و $\rho(\varepsilon_i) < a$ آنگاه $\rho(\varepsilon_j) < a$ است.

تذکر. $\rho(\varepsilon_i) \neq \varepsilon_i^2$ *می‌باشد.*

۱-۳- مدل رگرسیون خطی با خطای اندازه‌گیری

تجزیه و تحلیل مدل رگرسیونی، هنگامی که برخی متغیرها را نتوان دقیقا مشاهده یا اندازه‌گیری نمود، از مدت-ها پیش به عنوان یک مسئله مهم در برخی از بخش‌های کاربردی آمار شناخته شده است.

مشکل عمده در این مدل‌ها عدم وجود برآوردهای مناسب (نااریب و ساگار)، برای پارامترهای مدل می‌باشد و تحقیقات بیش‌تر در این زمینه بر مبنای اضافه نمودن فرض‌های مناسب و بدست آوردن برآوردهای مناسب می-باشد.

اگر بخواهیم رابطه میان دو متغیر را بررسی نمائیم، روش معمول استفاده از یک مدل رگرسیونی است. برای قابل شناسایی بودن مدل، لازم است فرضیاتی در نظربگیریم، و اگر هر یک از این فرضیات برقرار نباشد نتایج حاصله اعتبار نخواهند داشت. از جمله فرض‌های هر مدل رگرسیونی عدم وابستگی بین متغیرهای خطا و متغیرهای مستقل مدل می‌باشد. تحت این فرض، به‌راحتی و با استفاده از روش‌های موجود ، می‌توان مدل را کاملا تجزیه و تحلیل و پارامترهای آن را برآورد نمود.

اما در بسیاری از مواقع این فرض برقرار نبوده و بین متغیرهای خطا و متغیرهای مستقل وابستگی وجود دارد. این مشکل زمانی بوجود می‌آید که متغیر مستقل را فقط با خطا بتوان مشاهده نمود. در این صورت در مدل یک متغیر خطای دیگر نیز ظاهر می‌شود. این مدل‌ها را مدل‌های رگرسیونی با خطا در متغیرها می‌نامند.

مهمترین مشکل این مدل‌ها این است که در روش‌های از قبیل حداقل مربعات و ماکزیمم درستنمایی مستقیما نمی‌توان استفاده نمود و برآوردهای مناسب برای مدل وجود نخواهند داشت، مگر آن‌که فرضیاتی بر مدل اضافه شود. اما در عمل بسیاری از این فرضیات کاربردی ندارند. اما بهرحال روش‌های مختلفی برای تجزیه و تحلیل این مدل‌ها موجود می‌باشد. برخی فقط جنبه تئوری دارند و برخی دیگر از جنبه عملی کاربردهای بسیاری دارند.

مدل رگرسیونی زیر را در نظر بگیرید:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\xi_i = X_i + u_i$$

در این مدل $(X_i, \beta_0 + \beta_1 X_i)$مقادیر غیر قابل مشاهده و (ξ_i, Y_i) مقادیر قابل مشاهده می‌باشند. همچنین (β_0, β_1) پارامترهای مدل، (u_i, ε_i) متغیرهای تصادفی خطا که دارای توزیع مستقل با میانگین صفر و واریانس $(\sigma_u^2, \sigma_\varepsilon^2)$ می‌باشند. متغیر قابل مشاهده ξ_i را متغیر آشکار و متغیر غیر قابل مشاهده u_i را متغیر پنهان می نامند.

تذکر. زمانی که خطای اندازه‌گیری نداشته باشیم، مدل رگرسیونی تبدیل به مدل رگرسیون خطی معمول می‌شود، در این صورت ξ با X برابر است.

برای روشن شدن مطلب مثالی را ارائه می‌دهیم:

در این مثال رابطه بین میزان محصول ذرت و میزان نیتروژن موجود در خاک را در نظر می‌گیریم. فرض کنید که رابطه بین تولید ذرت و میزان نیتروژن به صورت یک مدل رگرسیون خطی معمولی است، X_i، میزان نیتروژن خاک، Y_i، میزان محصول ذرت و ضریب β نشان دهندهٔ رابطه بین این دو می‌باشد. به عبارتی با افزایش میزان نیتروژن موجود در خاک میزان تولید محصول هم بالا می‌رود. برای برآورد میزان نیتروژن، نمونه‌ای از خاک برای انجام آزمایش و تحلیل‌های آزمایشگاهی انتخاب می‌شود. میزان نیتروژن مشاهده شده برآوردی از X_i می‌باشد که با ξ_i نشان می‌دهیم و $\xi_i = X_i + u_i$ که u_i خطای اندازه‌گیری به وسیله نمونه‌گیری و تحلیل‌های آزمایشگاهی می‌باشد.

در این مدل X_iها ممکن است متغیرهای تصادفی و یا مقادیر ثابت باشند. اگر X_iها متغیرهای تصادفی باشند. مدل را ساختاری و اگر مقادیر ثابت باشند، مدل را تابعی می‌نامیم (برای اطلاعات بیش‌تر به Fuller در سال ۱۹۸۷ مراجعه شود).

در مدل رگرسیون خطی معمولی برآورد گر پارامتر β_1 به صورت زیر به دست می‌آید.

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{-1} \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

که $\hat{\beta}_1$ برآوردگری ناریب و دارای کمترین واریانس در بین برآوردگرهای خطی ناریب می‌باشد.

و در مدل رگرسیون خطی با خطای اندازه‌گیری برآوردگر پارامتر β_1 به صورت زیر معرفی می‌شود:

$$\hat{\beta}_1^* = \left[\sum_{i=1}^{n}(\xi_i - \bar{\xi})^2\right]^{-1} \sum_{i=1}^{n}(\xi_i - \bar{\xi})(Y_i - \bar{Y})$$

که این برآوردگر، برآوردگری اریب برای β_1 می‌باشد.

$$E(\hat{\beta}_1^*) = \sigma_{\xi\xi}^{-1}\sigma_{\xi Y} = \beta_1(\sigma_X + \sigma_{uu})^{-1}\sigma_{XX}$$

فرض کنید ما اثر خطای اندازه‌گیری را بر ضرائب حداقل مربعات خطا در مدل رگرسیونی ساده و مدل خطا تحت فرضیاتی که x_i متغیرهای تصادفی با $\sigma_{xx} > 0$ هستند.ما فرض می کنیم:

$$(x_i, e_i, u_i)^T = NI[(\mu_i, 0,0)^T, diag(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})]$$

بطوریکه NI نشان دهنده توزیع نرمال مستقل است و $diag(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})$ ماتریس قطری می‌باشد.

فرض کنید مدل با خطای اندازه گیری معرفی شده دارای توزیع نرمال دو متغیره با میانگین و ماتریس واریانس کواریانس زیر است:

$$E[(Y, X)] = (\beta_0 + \beta_1\mu_x, \mu_x)$$
$$\begin{bmatrix}\sigma_{YY} & \sigma_{XY}\\ \sigma_{XY} & \sigma_{XY}\end{bmatrix} = \begin{bmatrix}\beta_1^2\sigma_{xx} + \sigma_{ee} & \beta_1\sigma_{xx}\\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu}\end{bmatrix}.$$

پس نتیجه می‌گیریم که برای این مدل ضرایب رگرسیونی حداقل مربعات اریب هستند. ضریب اریبی برآوردگر β_1 برابر است با: $K_{XX} = \sigma_{\xi\xi}^{-1}\sigma_{\xi Y}$، و این ضریب اریبی را نسبت قابلیت اعتماد معرفی می‌کنند.(برای اطلاعات بیش‌تر به Fuller در سال ۱۹۸۷ مراجعه شود).

مشخص است که مدل رگرسیون ساده که در آن متغیر مستقل با خطا اندازه‌گیری شده است، بدون اطلاعات اضافی قابل شناسایی نمی‌باشد. به عنوان مثال چنین اطلاعات اضافی می‌توانند عبارت باشند از: معلوم بودن واریانس خطاهای متغیر مستقل، معلوم بودن نسبت واریانس خطای متغیر مستقل و وابسته و یا معلوم بودن عرض از مبدأ β_0(برای اطلاعات بیش‌تر به Fuller در سال ۱۹۸۷ مراجعه شود).

قابل شناسایی بودن مدل

یکی از مهم ترین تفاوت های مدل های خطای اندازه گیری خطی و مدل های رگرسیونی معمولی مربوط به قابل شناسایی بودن مدل است.

تعریف: فرض کنید پارامترهایی که در مدل مورد توجه هستند، به صورت بردار $\boldsymbol{\theta}\epsilon\Theta$ باشند که Θ فضای مقادیر ممکن $\boldsymbol{\theta}$ است. همچنین Z یک بردار تصادفی است که توزیع آن از خانواده توابع توزیع $\Phi = \{F_{\boldsymbol{\theta}}\ ;\ \boldsymbol{\theta}\epsilon\Theta\}$ باشد. آنگاه می گوییم پارامتر θ_i (i-امین مولفه بردار $\boldsymbol{\theta}$) قابل شناسایی است، اگر و تنها اگر هیچ دو مقداری از مقادیر $\boldsymbol{\theta}\epsilon\Theta$ که i-امین مولفه متفاوت دارند، وجود نداشته باشند که منجر به توزیع یکسان برای Z شوند. همچنین بردار $\boldsymbol{\theta}$ قابل شناسایی است، اگر و تنها اگر همه مولفه های $\boldsymbol{\theta}$ قابل شناسایی باشند. به عبارت دیگر $\boldsymbol{\theta}$ قابل شناسایی است، اگر به ازای هر $\boldsymbol{\theta}_1\epsilon\Theta$ و $\boldsymbol{\theta}_2\epsilon\Theta$ که $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$، بتوان نتیجه گرفت که:

$F_{\boldsymbol{\theta}_1}(a) \neq F_{\boldsymbol{\theta}_2}(a)$; a برای یک مقدار

در نهایت مدل قابل شناسایی است اگر و تنها اگر $\boldsymbol{\theta}$ قابل شناسایی باشد.

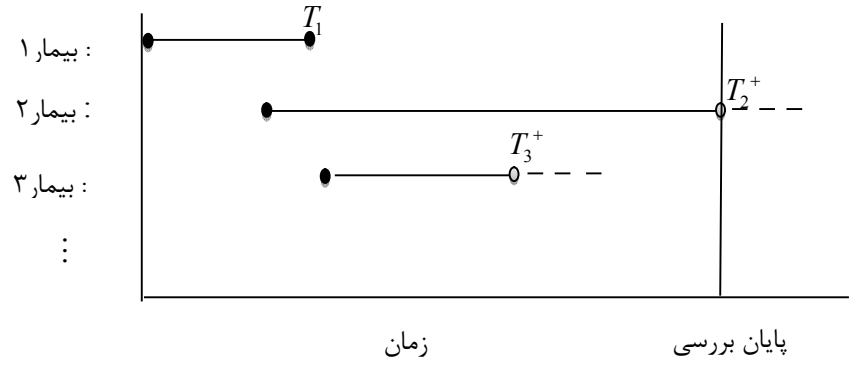
تذکر. در مدل‌های رگرسیون معمولی برآورد پارامتر یکتاست، اما در مدل‌های رگرسیونی با خطای اندازه‌گیری به دلیل وجود ضریب قابلیت اعتماد برآورد یکتا نیست. البته با در نظر گرفتن ۶ فرضیه زیر این مشکل برطرف می‌شود.

۲-قطع معالجه:

ممکن است به علت عوارض بد جانبی، درمان آن را متوقف کنیم. یا این‌که، ممکن است بیمار هنوز در تماس باشد ولی از ادامه معالجه خودداری کند.

۳-اتمام بررسی

نمودار زیر یک بررسی ممکن را تشریح می‌کند:



در این‌جا، بیمار۱ در زمان $t=0$ مورد بررسی قرار گرفته و در زمان T_1 فوت شده است. در نتیجه، یک مشاهده سانسور نشده بدست آمده است. بیمار۲ مورد بررسی قرار گرفته و در پایان بررسی هنوز زنده است. در نتیجه یک مشاهده سانسور نشده T_2^+ بدست آمده است. در نهایت، بیمار۳ مورد بررسی قرار گرفته و قبل از پایان بررسی، معالجه را قطع نموده است. در نتیجه، یک مشاهده سانسور شده T_3^+ بدست آمد است.

تذکر: در سانسور تصادفی متغیرهای C_i و T_i از هم مستقل هستند.

۴- انواع دیگر سانسور

انواع دیگر سانسور در منابع مختلف مورد بررسی قرار گرفته اند. انواع قبلی سانسور به سانسور راست و چپ تقسیم می‌شوند. اگر متغیر مورد مطالعه بسیار بزرگ باشد و نخواهیم آن را به طور کامل مشاهده کنیم، آن را "راست سانسور" نامند. به طور مشابه چپ سانسور قابل تعریف است.

فرض کنید که نمونه تصادفی $\{Y_i, \boldsymbol{\xi}_i, i = 1, \dots, n\}$از مدل (۱,۱,۳) که بصورت زیر حاصل شده است:

$$\begin{cases} Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \\ \boldsymbol{\xi}_i = \boldsymbol{X}_i + \boldsymbol{u}_i. \end{cases} \qquad (1.2.3)$$

در اینجا متغیر پاسخ Y ممکن است سانسور تصادفی از راست شده باشد، که متغیر سانسور با C نشان داده شده و بنابراین نمی‌تواند بصورت کامل مشاهده شود. فقط $\{Z_i, \delta_i\}$ مشاهده می‌شود، که بصورت زیر تعریف می‌گردد:

$$Z_i = min(Y_i, C_i), \quad \delta_i = I[Y_i \leq C_i], \quad i = 1,2, \dots n.$$

که $\{C_i\}$ یک نمونه مستقل و هم توزیع از تابع توزیع G می‌باشد. فرض کنید که برای.i=1,…n، به شرط داشتن X_i، C_i از Y_i مستقل است.

فرض کنید F و G، به ترتیب توابع توزیع Y و C هستند، که به صورت زیر تعریف می شوند:

$$F(x) = P(Y \leq x) \text{ و } G(x) = P(C \leq x) .$$

برای سادگی، برای هر تابع توزیع $H(.)$، $\overline{H} = 1 - H(.)$ را تعریف می‌کنیم. همچنین فرض می شود که $Y_i \geq 0, C_i \geq 0, i = 1,2, \dots n$. در آنالیز بقا اگر Y_i داده‌های طول عمر باشند، آنگاه $Y_i \geq 0$ است. بنابراین فرضیه بالا معتبر می باشد.

۳-۲-۱- روش حداقل مربعات اصلاح‌شده

از آنجایی که روش‌های برآورد پارامتر $\boldsymbol{\beta}$ در زمان سانسور نمی‌توانند بصورت مستقیم مورد استفاده قرار گیرند، برای حل این مشکل نیاز به تعریف تبدیلی از داده‌ها داریم. زمانی که توزیع G معلوم باشد، تعریف می‌کنیم:

$$Y_{iG} = \frac{\delta_i Z_i}{1 - G(Z_i)}, \quad i = 1,2, \dots n \qquad (2.2.3)$$

به سادگی دیده می‌شود که $E(Y_{iG}|\boldsymbol{X}_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$. برآوردگر حداقل مربعات اصلاح شده برای پارامتر $\boldsymbol{\beta}$ بصورت زیر می‌تواند تعریف شود:

نسبت سانسور ۰,۳، همانطور که می بینید برابر با $\hat{\beta} = 1.1000$ می‌باشد یعنی با افزایش سانسور دقت $\hat{\beta}$ کاهش می‌یابد. و احتمال پوشش به ۰,۷۱۹ افزایش یافته است.

توجه داشته باشید که متغیر پاسخ Y سانسور شده است. و محاسبه نسبت لگاریتم درستنمایی تجربی ساده نیست. بنابراین برای بدست آوردن آن از بسته‌های نرم افزاری آماده در نرم افزار R استفاده می کنیم.

نتیجه گیری:

از جداول زیر می‌توانیم نتایج زیر را بدست آوریم. اول اینکه عملکرد (AEL) بهتر از (NA) است. چون میانگین طول فواصل اطمینان آن بطور یکنواخت کوتاهتر است و احتمال پوشش آن در مقایسه با NA بیشتر است. دوم اینکه، برای هر نسبت سانسور همه میانگین طول‌ها نزولی هستند و احتمالات پوشش صعودی‌اند. و سوم اینکه مشخص است که نسبت سانسور همچنین بر طول فاصله اطمینان و احتمال پوشش اثرگذار است. بطور کلی، برای هر اندازه نمونه ثابت، با افزایش نسبت سانسور طول فاصله اطمینان افزایش پیدا می‌کند و احتمال پوشش کاهش می‌یابد. در نهایت احتمالات پوشش به دست آمده توسط NA اغلب از مقدار اسمی $1-\alpha$ کمتر است که این مطلب همچنین سازگار با نتایج شبیه سازی نشان داده شده در مدل خطای سانسور شده توسط Qin و Jingدر سال ۲۰۰۱ می‌باشد.

جدول۱,۴

متوسط طول و احتمالات پوشش فواصل اطمینان بر اساس روش NA برای β تحت نرخ سانسورهای متفاوت و اندازه نمونه n، وقتی که سطح اطمینان ۹۵٪ است.

Δ	N	$\hat{\beta}_{NA}$	Bias NA	MSE	Length	Cov. Pro. NA
				NA	NA	

Central limit theorem	قضیه حد مرکزی

ک

Polynomial	کثیرالجمله
Least square error	کم‌ترین مربعات خطا
Ordinary least square	کم‌ترین مربعات معمولی
Least trimmed of square	کم‌ترین میانگین پیراسته مربعات
Least meadian of square	کم‌ترین میانه مربعات

گ

Moment	گشتاور

م

Censoring variable	متغیر سانسور
Independent	مستقل
Error in variable model	مدل خطا
Proportional hazard model	مدل خطرات متناسب
Linear regression	مدل خطی
Linear EV model	مدل خطی خطا
Regression model	مدل رگرسیونی
Accelerated failure time model	مدل شتاب‌دار زمان شکست