رگرسیون ناپارامتری

حمزه ترابی[۱] ، مژگان بقایی‌پور[۲]

چکیده:

در این مقاله، پس از یادآوری رگرسیون پارامتری به بررسی معایب آن می‌پردازیم. سپس معرفی اجمالی بر رگرسیون ناپارامتری خواهیم داشت و در انتها دو روش از روش‌های رگرسیونی ساده‌ی ناپارامتری به نام‌های رگرسیون نادارایا - واتسون و رگرسیون چندجمله‌ای موضعی را معرفی خواهیم کرد.

واژه‌های کلیدی: اعتبارسنجی متقابل، رگرسیون جمعی، رگرسیون چندجمله‌ای موضعی، رگرسیون ساده ناپارامتری.

۱ مقدمه

تحلیل رگرسیون یکی از شاخه‌های مهم علم آمار است که هدف آن، بررسی اثر متغیرهای مستقل (پیش‌بین) روی متغیر وابسته (پاسخ) است.

رگرسیون را از لحاظ نوع تابعی که رابطه‌ی بین متغیرهای مستقل و وابسته را تعیین می‌کند، می‌توانیم به دو دسته‌ی رگرسیون پارامتری و ناپارامتری تقسیم کنیم.

در رگرسیون پارامتری، رابطه‌ی بین متغیرهای مستقل و وابسته از طریق تابع پارامتری f به صورت زیر معین می‌شود:

$$Y_i = f(\underline{\beta}', \mathbf{x}_i') + \varepsilon_i,$$

که در آن $\varepsilon_i \sim (۰, \Sigma^۲)$، $\mathbf{x}_i' = (x_{i۱}, ..., x_{ip})$ و $\underline{\beta}' = (\beta_۱, \ldots, \beta p)$ برداری از پارامترهای مجهول است. در صورت برآورد $\underline{\beta}$، تابع پارامتری f و بنابراین مدل به صورت $\hat{y} = f(\hat{\underline{\beta}}', \mathbf{x}')$ برآورد می‌شود.

اما در رگرسیون ناپارامتری، رابطه‌ی بین متغیرهای مستقل و وابسته از طریق تابع ناپارامتری $m(.)$ به صورت زیر معین می‌شود:

$$Y_i = m(\mathbf{x}_i') + \varepsilon_i,$$

که در آن $m(.)$ تابع نامعلومی است و در اکثر روش‌های ناپارامتری، تابعی پیوسته و هموار (مشتق‌پذیر) در نظر گرفته می‌شود.

در این مقاله به معرفی رگرسیون ناپارامتری پرداخته می‌شود. در ابتدا، برای فهم بهتر مطالب مربوط به رگرسیون ناپارامتری، به یادآوری مختصری از آن چه در دوره‌ی کارشناسی به عنوان رگرسیون پارامتری (سنتی) با آن آشنا شده‌ایم و بررسی معایب این روش می‌پردازیم[۱،۲،۸]. در ادامه، پس از معرفی کلی رگرسیون ناپارامتری به بحث و بررسی بیشتر درباره‌ی دو مدل از مدل‌های رگرسیونی ساده‌ی ناپارامتری به نام‌های رگرسیون نادارایا - واتسون و رگرسیون چندجمله‌ای موضعی پرداخته می‌شود. در این قسمت با ارائه‌ی

[۱] عضو هیأت علمی دانشگاه یزد
[۲] فارغ التحصیل کارشناسی ارشد از دانشگاه یزد

مثال‌های مختلف سعی در فهم بهتر مفاهیم و مطالب بیان شده داریم.

۲ یادآوری رگرسیون پارامتری

رگرسیون پارامتری را می‌توان به دو دسته رگرسیون پارامتری تک‌متغیره و چندمتغیره تقسیم‌بندی کرد.

۱) تک‌متغیره:

$$Y_i = f(x_i) + \varepsilon_i,$$

مانند رگرسیون چندجمله‌ای که در آن

$$f(x_i) = \beta_0 + \beta_1 x_i + \ldots + \beta_q x_i^q.$$

۲) چندمتغیره:

$$Y_i = f(x_{i1}, \ldots, x_{ip}) + \varepsilon_i,$$

مانند رگرسیون خطی چندگانه که در آن

$$f(x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

در این قسمت برای سادگی کار تنها مدل رگرسیونی تک‌متغیره چندجمله‌ای، یعنی مدلی با یک متغیر رگرسیونی x که با یک متغیر پاسخ Y دارای ارتباط است را مورد بررسی قرار می‌دهیم. داده‌ها به صورت (x_i, y_i)، $1 \le i \le n$ هستند. مدل

$$y_i = \sum_{j=0}^{q} \beta_j x_i^j + \varepsilon_i.$$

را در نظر بگیرید. در واقع هدف، برآورد پارامترهای نامعلوم $\{\beta_0, \ldots, \beta_q\}$ به روش کمترین مربعات خطا از طریق مینیمم کردن رابطه‌ی

$$Q = \sum_{i=1}^{n} (\sum_{j=0}^{q} \beta_j x_i^j - y_i)^2.$$

است. با به کارگیری نمادهای ماتریسی که در ادامه معرفی می‌شوند، مدل رگرسیونی چندجمله‌ای به صورت زیر بازنویسی می‌شود:

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon},$$

که در آن نمادهای ماتریسی مذکور عبارت اند از:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^q \\ 1 & x_2 & x_2^2 & \ldots & x_2^q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^q \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_q \end{pmatrix}, \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

با مشتق‌گیری از Q برحسب β_j ها و به کارگیری نمادهای ماتریسی، معادلات نرمال به صورت زیر به‌دست می‌آیند:

$$\mathbf{X}'\mathbf{X}\underline{\beta} = \mathbf{X}'\mathbf{Y}.$$

بنابراین برآورد پارامترهای نامعلوم مدل به صورت یکتا، از رابطه‌ی زیر حاصل می‌شود:

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

و در آخر مدل به صورت زیر برآورد می‌شود:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\underline{\hat{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}, \end{aligned}$$

به طوری که $\hat{\mathbf{Y}} = (\hat{y}_1 \quad \ldots \quad \hat{y}_n)$. ماتریس $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$، $n \times n$، که در واقع ماتریس تصویر، روی فضای تولید شده توسط ستون‌های

ماتریس $\mathbf{X}$ است، ماتریس هَت نام دارد که معمولاً ماتریس برازش نیز نامیده می‌شود. اگر ارتباط بین مقادیر مشاهدات y_j و مقادیر برازش شده‌ی $\hat{y}_i$ را از طریق رابطه‌ی $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$، مورد توجه قرار دهیم، در واقع می‌توانیم h_{ij} را به عنوان ij امین عضو ماتریس $\mathbf{H}$ در نظر بگیریم. ماتریس هت و ویژگی‌ها آن نقشی مهم در تحلیل رگرسیونی بازی می‌کنند.

به اختلاف بین مقادیر مشاهده شده‌ی y_i و مقادیر برازش شده‌ی متناظر $\hat{y}_i$، یعنی $e_i = y_i - \hat{y}_i$، باقیمانده می‌گوییم. مناسب است که n باقیمانده‌ی مذکور به صورت نماد ماتریسی زیر نوشته شود:

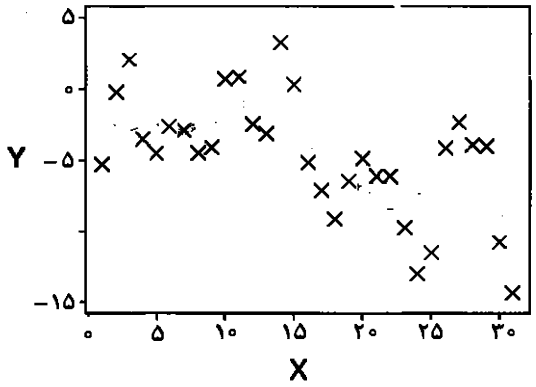
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}},$$

هم‌چنین مجموع مربعات مانده‌ها عبارت است از

$$\begin{aligned} SSE &= \mathbf{e}'\mathbf{e} \\ &= \mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}. \end{aligned} \qquad (۱)$$

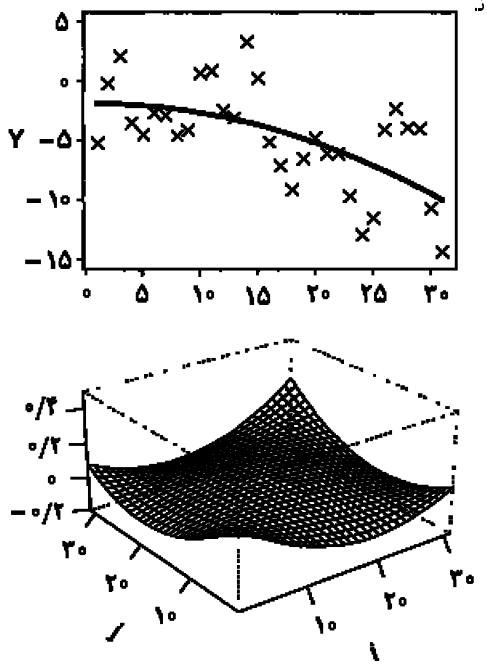
توجه کنید که در رگرسیون کاربردی، معادله‌ی رگرسیونی برازش شده‌ای مفید و مورد توجه است که دارای کمترین مقدار SSE باشد. همان طور که از رابطه‌ی ((۱)) مشخص است، کاهش SSE در رابطه‌ی مستقیم با ماتریس هت است به این صورت که اگر درایه‌های روی قطر اصلی ماتریس هت به یک و درایه‌های غیر قطر اصلی به صفر نزدیک باشند، میزان SSE کاهش می‌یابد. از این جاست که اهمیت ماتریس هت در مباحث رگرسیونی مشخص می‌شود. مثال زیر این نکته را روشن‌تر می‌کند.

مثال ۲–۱ میانگین روزانه‌ی درجه حرارت هوا در شهر واکانای[۳] (Y) برحسب روزهای ماه ژانویه‌ی ۱۹۹۶ ثبت شده است. روشن است که مقادیر متغیر مستقل، اعداد ۱، ۲، ...، ۳۰ هستند. نمودار پراکنش حاصل در شکل زیر نمایش داده شده است:



شکل ۱. نمودار پراکنش.

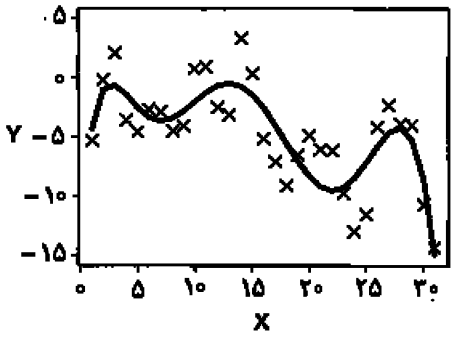
یک معادله‌ی چندجمله‌ای درجه‌ی ۲ در شکل ۲ و مقادیر درایه‌های ماتریس هت متناظر با برازش در شکل ۳ ن



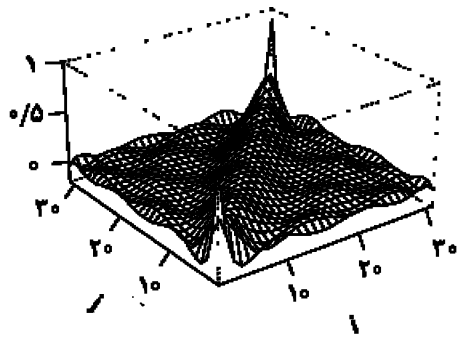
[۳] Wakkanai

شکل ۳. درایه‌های ماتریس هت متناظر با شکل ۲.

همان طور که دیده می‌شود نمودارهای برازش داده شده از متغیرهای Y_i اختلاف زیادی دارند و این از معایب برازش معادله‌ی چندجمله‌ای محسوب می‌شود؛ برای نزدیک کردن مقادیر $\hat{y}_i$ به متغیرهای Y_i از چندجمله‌ای‌هایی با درجه‌ی بالاتر بهره می‌گیریم. در ادامه یک معادله‌ی درجه‌ی ۷ در شکل ۴ به داده‌ها برازش می‌دهیم.



شکل ۴. برازش معادله‌ی درجه‌ی ۷.



شکل ۵. درایه‌های ماتریس هت متناظر با شکل ٦.

اگرچه در این نمودار مقادیر $\hat{y}_i$ به داده‌های y_i نزدیک‌تر شده است اما باز هم نمودار برازش داده شده نمودار مناسبی به نظر نمی‌رسد. اولین مسأله‌ای که جلب توجه می‌کند در نقطه $x_i = ۲۲$ رخ داده است. در این نقطه داده‌ها یک ماکسیمم نسبی را نشان می‌دهند در حالی که نمودار برازش داده شده حاکی از یک مینیمم نسبی است. بنابراین، این نمودار برازش داده شده، به‌ویژه در مواقعی که هدف پیش‌بینی است، مناسب نخواهد بود. این نکته با توجه به نمودار ماتریس هت متناظر در شکل ۷ نیز نمایان است. این نمودار از نظر شکل ظاهری نوسانات زیادی دارد؛ یعنی ماتریس هت علاوه بر گرفتن مقادیر منفی، دارای مقادیر کوچکی از نظر قدرمطلق هم نیست و این مطلب با توجه به رابطه ((۱)) باعث افزایش چشم‌گیر SSE می‌شود. بنابراین همان طور که دیدید برازش یک معادله‌ی درجه ۷ نیز چندان مناسب به نظر نمی‌رسد.

در واقع این نقص جزء ویژگی‌های ذاتی ماتریس هت در روش رگرسیون پارامتری چندجمله‌ای است که با افزایش درجه‌ی چندجمله‌ای مقادیر درایه‌های ماتریس هت به شدت افزایش پیدا می‌کنند.

اگرچه ماتریس هت یک آماره‌ی مهم را برای تشخیص رابطه‌ی بین داده‌ها و برآوردها فراهم می‌کند ولی این ماتریس همه‌ی ویژگی‌های رگرسیونی را پوشش نمی‌دهد. بنابراین برازش رگرسیون چندجمله‌ای که توسط ماتریس هت ارائه می‌شود، غالباً منجر به برآوردهایی دور از واقعیت می‌گردد و به کارگیری آن توصیه نمی‌شود.

بنابراین باید از ماتریس هتی استفاده کنیم که یک برازش طبیعی را به داده‌ها ارائه دهد و این یکی از دلایلی است که از روش‌های بنا شده بر پایه‌ی رگرسیون ناپارامتری برای بررسی رابطه‌ی بین متغیرهای مستقل و پیش‌بین بهره می‌گیریم.

۳ معرفی رگرسیون ناپارامتری

هدف از رگرسیون ناپارامتری، برآورد تابع رگرسیون به صورت مستقیم است. رگرسیون ناپارامتری اثر یک یا چند متغیر مستقل را روی یک متغیر وابسته بررسی می‌کند، بدون این که از قبل تابع ویژه‌ای را برای ارتباط بین متغیر مستقل و وابسته در نظر گرفته باشد. رگرسیون ناپارامتری را می‌توان از لحاظ تعداد متغیرهای مستقل به دو گروه زیر تقسیم‌بندی کرد:

۱) تک‌متغیره (رگرسیون ساده‌ی ناپارامتری):

$$Y_i = m(x_i) + \varepsilon_i.$$

۲) چندمتغیره (رگرسیون جمعی ناپارامتری):

$$Y_i = m_0 + m_1(x_{i1}) + m_2(x_{i2}) + \ldots + m_k(x_{ik}) + \varepsilon_i, \quad (2)$$

به طوری که به $m_j(.)$ ها، $j = 1, \ldots, k$، تابع‌های رگرسیونی جزئی گفته می‌شود و $\varepsilon_i \sim (0, \Sigma^2)$.

رگرسیون ساده‌ی ناپارامتری یکی از ساده‌ترین مدل‌های رگرسیونی ناپارامتری است که در آن متغیر مستقل به صورت تک‌متغیره در نظر گرفته می‌شود. در این مدل، داده‌ها به صورت (x_i, y_i)، $i = 1, 2, \ldots, n$، هستند که غالباً x_i ها به صورت صعودی مرتب شده‌اند و داریم:

$$m(x_i) = E(Y_i | X_i = x_i).$$

با توجه به این که در این روش، نمودار پراکنش y_i برحسب x_i کاربرد زیادی دارد، در برخی از مراجع نام هموارسازی پراکنش نیز برای این روش در نظر گرفته شده است. توجه کنید که در انواع روش‌های برازش یک نمودار به داده‌ها، منظور از هموارسازی، برازش یک نمودار مشتق‌پذیر به داده‌هاست.

$m(x)$ در معادله‌ی رگرسیون ساده‌ی ناپارامتری تابعی است که رفتار ذاتی داده‌ها را ارائه می‌دهد و به تصادف یا شانس وابسته نیست. مقدار برآورد شده‌ی $m(x)$ را با $\hat{m}(x)$ نمایش می‌دهیم که تابعی هموار از x است.

همان طور که در رگرسیون پارامتری گفتیم همیشه برازش چندجمله‌ای به روش کمترین مربعات مناسب نیست زیرا در این روش تنها یک چندجمله‌ای به ناحیه‌ای که داده‌ها در درون آن قرار گرفته‌اند، برازش داده می‌شود. برازش تنها یک معادله به تعداد زیادی از داده‌ها با رفتار پیچیده گاهی به نتایج نامطلوبی منجر می‌شود.

این مشکلات به مفهوم افراز کردن داده‌ها به چند ناحیه برای برازش منحنی به داده‌ها در هر ناحیه و عطف آن‌ها به طوری که برآوردگرهای مناسب و مفیدی باشند، منتهی می‌شود. در واقع ساده‌ترین روش، برازش معادله‌ی یک منحنی به صورت پاره‌ای (تکه‌ای) است به این شرط که برآوردها در انتهای هر زیر ناحیه مقادیر یکسانی را اختیار کنند.

انواع روش‌هایی که در رگرسیون ساده ناپارامتری برای برآورد $m(x)$ وجود دارند، عبارتند از:

۱) رگرسیون نادارایا - واتسون (برآوردگر هسته‌ای)

۲) رگرسیون چندجمله‌ای موضعی

۳) اسپلاین هموارساز

۴) روش LOESS

اگرچه در نگاه اول ممکن است این نکته مشهود نباشد، اما تمام روش‌هایی که در این قسمت با آن‌ها آشنا می‌شویم در چارچوب کلی میانگین وزنی جای می‌گیرند.

اما در رگرسیون جمعی که توسط رابطه‌ی ((۲)) معرفی شد، داده‌های نمونه‌ی تصادفی به صورت $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ هستند، به طوری که $y_i \in \mathbb{R}$ و $\mathbf{x}_i \in \mathbb{R}^k$. در رابطه‌ی ((۲))، m_0 مقداری ثابت است که توسط $E(Y)$ برآورد می‌شود و تابع‌های رگرسیونی جزئی $j = 1, 2, \ldots, n$، که به وسیله‌ی داده‌ها برآورد می‌شوند، تابع‌هایی هموار هستند به طوری که برای هر $j > 0$،

$$E[m_j(X_j)] = 0.$$

این مدل را می‌توان در دو جهت مختلف گسترش داد:

۱) برای اضافه کردن اثر متقابل بین متغیرهای مستقل مورد نظر، مانند

$$Y_i = m_0 + m_1(x_{1i}) + m_{2,3}(x_{2i}, x_{3i}).$$

۲) برای اضافه کردن جملات خطی در مدل، مانند

$$Y_i = m_0 + \beta_1 x_{1i} + m_2(x_{2i}), \qquad (۳)$$

مدل ارائه شده در ((۳)) مدل نیمه‌پارامتری نامیده می‌شود.

از حدود پانزده سال قبل روش قدیمی پس‌برازشی که توسط بیوجا[۴] و همکاران در سال ۱۹۸۹ و هاستای[۵] و تیبشیرانی[۶] [۴] معرفی و گسترش داده شد، مشهورترین روش برای برآورد مدل جمعی بوده است. این روش در مطالعات شبیه‌سازی و در به کارگیری داده‌های واقعی روشی مناسب به حساب می‌آمد اما به علت طبیعت تکراری این روش، تحلیل ویژگی‌های آماری مدل با مشکلات زیادی روبه‌رو بود. به همین دلیل روش انتگرال‌گیری کناری توسط تی‌جوس‌تیم[۷] و اوستاد[۸] در سال ۱۹۹۴ و لین‌تون و نیل‌سن[۹] [۵] معرفی شد. برای آشنایی بیشتر با این دو روش می‌توانید به مراجع [۴]، [۵] و [۶] مراجعه کنید.

در ادامه تنها با دو روش اول از رگرسیون ساده‌ی ناپارامتری آشنا می‌شویم. برای آشنایی بیشتر با رگرسیون ناپارامتری می‌توانید به فاکس[۱۰] [۳]، سیمنو[۱۱] [۷] و تاکزاوا[۱۲] [۸] مراجعه کنید.

۱.۳ برآوردگر نادارایا - واتسون

در ادامه به معرفی روش رگرسیونی برآوردگر نادارایا - واتسون می‌پردازیم که از میانگین وزنی داده‌ها به طور مستقیم برای برآورد $m(x)$ بهره می‌گیرد.

فرض کنید که متغیرهای مستقل و وابسته، هر دو متغیر تصادفی پیوسته باشند. می‌دانیم که هرگاه $f_X(x)$ تابع چگالی X باشد آنگاه برای هر Δx کوچک

$$P(x < X < x + \Delta x) \approx f_X(x)\Delta x, \qquad (۴)$$

[۴] Buja
[۵] Hastie
[۶] Tibshirani
[۷] Tjøstheim
[۸] Auestad
[۹] Linton and Nielsen
[۱۰] Fox
[۱۱] Simonoff
[۱۲] Takezawa

به همین ترتیب هرگاه $f(x,y)$ تابع چگالی توأم X و Y باشد آنگاه برای هر Δx و Δy کوچک

$$P(x<X<x+\Delta x, y<Y<y+\Delta y) \approx f(x,y)\Delta x\Delta y, \qquad (۵)$$

هم‌چنین هرگاه $f(y|x)$ تابع چگالی احتمال شرطی Y به شرط x باشد آنگاه

$$P(y<Y<y+\Delta y|x<X<x+\Delta x) \approx f(y|x)\Delta y,$$

با توجه به این نکته که مقدار تابع $m(x)$ برای مقدار مستقل x هم‌ارز با امید ریاضی متغیر Y به شرط $X=x$ است، به‌دست می آوریم که

$$\begin{aligned} m(x) &= E(Y|X=x) \\ &= \int_{-\infty}^{\infty} y f(y|x)dy \\ &= \frac{۱}{f_X(x)}\int_{-\infty}^{\infty} y f(x,y)dy. \end{aligned} \qquad (۶)$$

رابطه‌ی ((۶)) نشان می‌دهد که برای برآورد تابع $m(x)$ کافی است تابع $f(y|x)$ یا در واقع توابع $f_X(x)$ و $f(x,y)$ را برآورد کنیم. یکی از ساده‌ترین روش‌ها برای برآورد توابع $f_X(x)$ و $f(x,y)$ به کارگیری برآوردگر چگالی هسته‌ای است. با استفاده از ایده‌ی روش گشتاورها و روابط ((۴)) و ((۵))، برآوردگرهای توابع $f_X(x)$ و $f(x,y)$ به صورت زیر به‌دست می آیند:

$$\begin{aligned} \hat{f}(x,y) &= \frac{۱}{nh_xh_y}\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x})k_y(\frac{y-Y_i}{h_y}), \\ \hat{f}_X(x) &= \frac{۱}{nh_x}\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x}), \end{aligned}$$

به طوری که $k_x(.)$ و $k_y(.)$ تابع‌های هسته‌ای و h_x و h_y مقادیری ثابت و مثبت به نام پهنه هستند که در ادامه بیشتر با آن‌ها آشنا می‌شویم.

تابع‌های هسته‌ای معمولاً در شرایط زیر صدق می‌کنند:

$$\begin{aligned} \int_{-\infty}^{\infty} k_x(u)du &= \int_{-\infty}^{\infty} k_y(u)du = ۱, \qquad (۷) \\ \int_{-\infty}^{\infty} uk_x(u)du &= \int_{-\infty}^{\infty} uk_y(u)du = ۰, \qquad (۸) \end{aligned}$$

$$\int_{-\infty}^{\infty} u^۲k_x(u)du < \infty,$$
$$\int_{-\infty}^{\infty} u^۲k_y(u)du < \infty.$$

از معادلات بیان شده در روابط ((۷)) و ((۸)) نتیجه می‌گیریم که

$$\begin{aligned} \frac{۱}{h_x}\int_{-\infty}^{\infty} k_x(\frac{x-X_i}{h_x})dx &= \frac{۱}{h_y}\int_{-\infty}^{\infty} k_y(\frac{y-Y_i}{h_y})dx \\ &= ۱, \end{aligned}$$

$$\frac{۱}{h_x}\int_{-\infty}^{\infty} xk_x(\frac{x-X_i}{h_x})dx = X_i,$$

$$\frac{۱}{h_y}\int_{-\infty}^{\infty} yk_y(\frac{y-Y_i}{h_y})dy = Y_i. \qquad (۹)$$

به این ترتیب برای به‌دست آوردن $\hat{m}(x)$ کافیست برآوردهای $f_X(x)$ و $f(x,y)$ را در رابطه‌ی ((۶)) جایگزین کنیم بنابراین

$$\begin{aligned} \hat{m}(x) &= \int_{-\infty}^{\infty} y\frac{\hat{f}(x,y)}{\hat{f}_X(x)}dy \\ &= \int_{-\infty}^{\infty} y\frac{\frac{۱}{nh_xh_y}\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x})k_y(\frac{y-Y_i}{h_y})}{\frac{۱}{nh_x}\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x})}dy \\ &= \frac{\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x})\int_{-\infty}^{\infty} yk_y(\frac{y-Y_i}{h_y})dy}{h_y\sum_{i=۱}^{n} k_x(\frac{x-X_i}{h_x})}. \qquad (۱۰) \end{aligned}$$

با جایگزین کردن رابطه‌ی (۹) در (۱۰) و جایگزین کردن $k_x(.)$ و h_x با $k(.)$ و h به‌دست می آوریم:

$$\hat{m}(x) = \sum_{i=۱}^{n} w_i(x,h)Y_i, \qquad (۱۱)$$

به طوری که

$$w_i(x,h)=\frac{k(\frac{x-X_i}{h})}{\sum_{k=1}^{n}k(\frac{x-X_k}{h})}.$$

برآوردگری که برای برآورد $m(x)$، در رابطه‌ی (۱۱) معرفی شد، برآوردگر نادارایا - واتسون نام دارد.

تابع $w_i(x,h)$ که هسته هم‌ارز نامیده می‌شود، وزن‌هایی را معین می‌کند که برای ارزیابی چگونگی اهمیت x_i ها در رابطه با Y_i، مورد توجه قرار می‌گیرد. با توجه به رابطه‌ی (۱۱) دیده می‌شود که این برآوردگر، یک برآوردگر خطی است. نکته قابل توجه دیگر این که $w_i(x,h)$ در برآوردگر نادارایا - واتسون دارای ویژگی زیر است:

$$\sum_{i=1}^{n}w_i(x,h)=1.$$

یعنی برآوردگر نادارایا - واتسون میانگین وزنی Y_i ها است. روشن است که تابع $w_i(x,h)$ در رابطه‌ی مستقیم با مقدار h و تابع $k(.)$ است.

توجه کنید که شرط بیان شده در رابطه‌ی ((۷)) برای توابع هسته‌ای $k(.)$ قابل صرف نظر کردن است یعنی اگر تابع هسته‌ای در این شرط صدق نکند، در به‌دست آوردن $\hat{m}(x)$ تأثیری ندارد. یکی از ساده‌ترین تابع‌های هسته‌ای، تابع $\exp(-u^2)$ است که به صورت‌های زیر نیز می‌توان آن را به کار گرفت:

$$\exp(\frac{-u^2}{2}),\quad \frac{1}{\sqrt{2\pi}}\exp(\frac{-u^2}{2}).$$

بهره‌گیری از هر کدام از توابع بالا، برآورد تابع رگرسیونی $\hat{m}(x)$ را بدون تغییر باقی می‌گذارد.

پهنه‌ی h در رابطه‌ی (۱۱) را می‌توان به صورت دنباله‌ای مثبت $\{h_n\}$ در نظر گرفت که با افزایش حجم نمونه به سمت صفر میل می‌کند. از h می‌توان به عنوان پارامتر هموارسازی نیز یاد کرد؛ به این مفهوم که هرچه مقدار h افزایش (کاهش) یابد، نمودار حاصل از برازش رگرسیون ناپارامتری هموارتر (ناهموارتر) می‌گردد. نکته‌ای که در این‌جا باید به آن توجه کنیم این است که افزایش h نمی‌تواند به صورت دلخواه ادامه یابد زیرا افزایش h در رابطه‌ی مستقیم با افزایش مربع انحرافات از میانگین [۱۳] است و همان طور که می‌دانیم برآورد یک تابع رگرسیونی با MSE زیاد مطلوب نیست؛ بنابراین h را باید به صورتی انتخاب کنیم که نمودار حاصل هموارترین نمودار با کوچک‌ترین MSE باشد. در ادامه دو معیار برای انتخاب h بهینه معرفی خواهیم کرد.

برآوردگر Y_i، یعنی $\hat{Y}_i$ را می‌توان به صورت زیر مشخص نمود:

$$\hat{Y}_i=\sum_{j=1}^{n}h_{ij}Y_j,$$

که در آن

$$\begin{aligned}h_{ij} &= w_j(x_i,h)\\ &= \frac{k(\frac{x_i-x_j}{h})}{\sum_{k=1}^{n}k(\frac{x_i-x_k}{h})},\end{aligned}\qquad (۱۲)$$

که h_{ij} در رابطه‌ی (۱۲)، ij امین عضو ماتریس هت H خواهد بود. اعضای قطری ماتریس هتی که در رابطه‌ی (۱۲) معرفی شد، کاربردهای مفیدی دارند که در ادامه به بررسی آن‌ها می‌پردازیم.

۲.۳ شاخصی سودمند برای انتخاب معادله‌ی رگرسیونی

همان طور که ملاحظه کردید، تابع $\hat{m}(x)$، از طریق تابع وزنی $w_i(x,h)$، در ارتباط با پهنه‌ی h است و با تغییر

[۱۳] MSE

مقدار h می‌توان هموارسازی‌های متفاوتی را از $\hat{m}(x)$ به‌دست آورد. تعیین مقدار پهنه به نحوی که نمودار رگرسیونی برازش داده شده نموداری مفید باشد، نکته‌ای قابل توجه است. بنابراین باید از بین نمودارهایی که می‌توان به داده‌ها برازش داد، بهترین را انتخاب نمود. اما رسم همه‌ی نمودارهای ممکن و سپس انتخاب یکی از بین آن‌ها کار عاقلانه‌ای به نظر نمی‌رسد و غیر ممکن است. به همین دلیل معرفی شاخص‌هایی برای انتخاب بهترین مدل از طریق انتخاب h الزامی به نظر می‌رسد.
یکی از شاخص‌های سودمند و کاربردی برای مقایسه‌ی نمودارهای رگرسیونی، شاخص اعتبارسنجی متقابل است.

تعریف ۳–۱ شاخص اعتبارسنجی متقابل $\hat{m}(x)$ شاخصی برای مقایسه‌ی پیش‌بینی داده‌های مربوط به آینده است و فرمول آن از رابطه‌ی زیر به‌دست می‌آید:

$$\mathrm{CV}[\hat{m}(x)] = \sum_{i=1}^{n} \frac{(Y_i - \hat{m}(X_i))^2}{n(1-h_{ii})^2}.$$

نکته‌ای که باید در بهره‌گیری از شاخص CV مورد توجه قرار دهیم این است که این رابطه برای معادله‌های رگرسیونی که در شرایط زیر صدق می‌کنند، مفید است:

۱) معادله‌ی رگرسیونی به روش کمترین مربعات به وسیله‌ی ماتریس هت به‌دست آمده باشد.

۲) معادله‌ی رگرسیونی مورد استفاده، یکتا باشد.

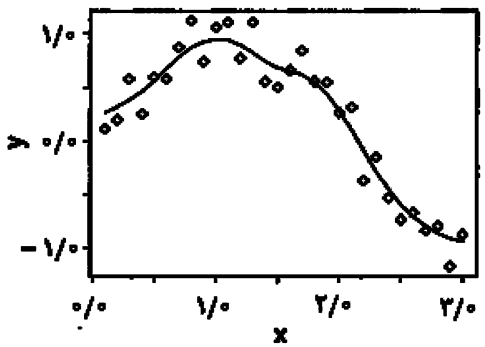
شاخص دیگری که برای مقایسه‌ی معادله‌های رگرسیونی به کار گرفته می‌شود، شاخص اعتباربخشی متقابل تعمیم یافته نام دارد.

تعریف ۳–۲ شاخص اعتباربخشی متقابل تعمیم یافته، شاخصی برای مقایسه‌ی بین نمودارهای رگرسیونی است که به صورت زیر تعریف می‌شود:

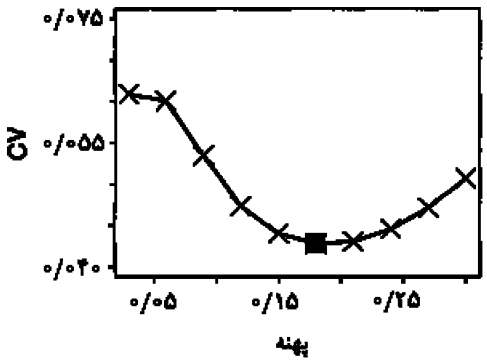
$$\mathrm{GCV}[\hat{m}(x)] = \frac{\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2}{n(1-\frac{\sum_{i=1}^{n} h_{ii}}{n})^2}.$$

روشن است که h بهینه مربوط به مینیمم CV و GCV است. در این‌جا برای آشنایی بیشتر با این روش مثالی ارائه می‌کنیم.

مثال ۳–۲ در شکل ٦ نموداری با استفاده از روش رگرسیونی برآوردگر نادارایا - واتسون به داده‌هایی که با علامت ◊ مشخص شده‌اند، برازش داده‌ایم. در این برازش h بهینه برابر با ۰/۱۸ در نظر گرفته شده است. این انتخاب با استفاده از معیار CV که در شکل ۷ نمایش داده شده‌است، صورت گرفته است.



شکل ٦. برازش نمودار به وسیله‌ی رگرسیون نادارایا - واتسون.



شکل ۷. نتایج محاسبه‌ی CV برای داده‌های شکل ٦.

۳.۳ اریبی کران

یک اشکال ذاتی در روش برآوردگر هسته‌ای وجود دارد. این اشکال وقتی رخ می‌دهد که x_i که می‌خواهیم $\hat{y}_i$ متناظرش را برآورد کنیم یکی از کران‌های داده‌ها یعنی مینیمم یا ماکسیمم داده‌ها باشد. این محدودیت از چپ یا راست x_i باعث می‌شود که $\hat{y}_i$ به شدت تحت تأثیر داده‌هایی قرار گیرد که x_i در آن سمت محدود نیست. در واقع در این موقعیت، پهنه‌ی h به طور متقارن گسترده نشده است و به مقادیری که در سمت کران x_i واقع شده‌اند، به ناچار وزن صفر داده می‌شود. این نکته باعث می‌شود که $\hat{y}_i$ برآورد شده به شدت تحت تأثیر داده‌هایی که در جهت دیگر قرار گرفته‌اند، واقع شود. به این اشکال در رگرسیون نادارایا - واتسون اریبی کران گفته می‌شود.

رگرسیون چندجمله‌ای موضعی، یکی از انواع روش‌های رگرسیون ساده‌ی ناپارامتری است که این اشکال را به صورت چشم‌گیری تعدیل می‌کند و در واقع می‌توان روش رگرسیون چندجمله‌ای موضعی را به عنوان تعمیمی از رگرسیون نادارایا - واتسون در نظر گرفت.

۴.۳ رگرسیون چندجمله‌ای موضعی

رگرسیون چندجمله‌ای موضعی یکی از انواع روش‌های رگرسیونی ساده‌ی ناپارامتری است. در این روش نیز مانند اکثر روش‌های رگرسیون ناپارامتری ناحیه‌ای را که داده‌ها در آن قرار گرفته‌اند به زیرناحیه‌هایی تقسیم می‌کنیم و سپس معادله‌ی خطی را به هر زیرناحیه برازش می‌دهیم به طوری که نمودارها در انتهای هر زیرناحیه پیوسته باشند.

همان طور که گفته شد رگرسیون چندجمله‌ای موضعی، تعمیمی از روش برآوردگر نادارایا - واتسون است. معادله‌ی چندجمله‌ای موضعی وقتی متغیر مستقل x به x^* نزدیک است به صورت زیر تعریف می‌شود:

$$m(x,x^*)=a_0(x^*)+\sum_{j=1}^{p}a_j(x^*)(x-x^*)^j. \quad (۱۳)$$

ساده‌ترین و عمومی‌ترین حالت در این روش، حالتی است که $p=1$ در نظر گرفته شود؛ یعنی از روش رگرسیون خطی موضعی بهره بگیریم.

مقادیر $\{a_0(x^*),a_1(x^*),\ldots,a_p(x^*)\}$ از مینیمم کردن رابطه‌ی زیر به‌دست می‌آیند:

$$E_{local}(x^*)= \sum_{i=1}^{n}\left(w(\frac{x_i-x^*}{h})\,(m(x_i,x^*)-y_i)^2\right) \quad (۱۴)$$

که در آن

$$m(x_i,x^*)=a_0(x^*)+\sum_{j=1}^{p}a_j(x^*)(x_i-x^*)^j.$$

$w(\frac{x_i-x^*}{h})$ در رابطه‌ی (۱۴) همان تابع هسته‌ای است که درجه اهمیت نقطه‌ی (x_i,y_i)، وقتی معادله‌ی رگرسیونی مربوط به x^* مدنظر است را معین می‌کند. این تابع وزنی نامنفی، بیشترین مقدار خود را در حالت $x_i=x^*$

می‌گیرد. انواع مختلف توابع وزنی که در این روش می‌توانیم از آن‌ها بهره بگیریم عبارتند از:

$$w(\tfrac{x_i-x^*}{h}) = \exp\left(-\tfrac{1}{2}(\tfrac{x_i-x^*}{h})^2\right),$$

$$= \begin{cases} \left(1-(\tfrac{x_i-x^*}{h})^2\right)^2 & \text{if } (\tfrac{x_i-x^*}{h})^2 \le 1 \\ 0 & \text{if } (\tfrac{x_i-x^*}{h})^2 > 1, \end{cases}$$

$$= \begin{cases} \left(1-(\tfrac{x_i-x^*}{h})^2\right)^3 & \text{if } (\tfrac{x_i-x^*}{h})^2 \le 1 \\ 0 & \text{if } (\tfrac{x_i-x^*}{h})^2 > 1, \end{cases}$$

$$= \begin{cases} \left(1-(\tfrac{|x_i-x^*|}{h})^3\right)^3 & \text{if } (\tfrac{|x_i-x^*|}{h})^3 \le 1 \\ 0 & \text{if } (\tfrac{|x_i-x^*|}{h})^3 > 1. \end{cases}$$

با بهره‌گیری از نمادهای ماتریسی، $E_{local}(x^*)$ را به صورت ساده‌ترِ زیر می‌نویسیم:

$$E_{local}(x^*) = (\mathbf{Xa}-\mathbf{Y})'\mathbf{W}(\mathbf{Xa}-\mathbf{Y}), \quad (۱۵)$$

به کمک رابطه (۱۵) برآوردگر a از رابطه‌ی

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y},$$

به دست می‌آید که در آن ماتریس $\mathbf{X}$، ماتریس وزنی $\mathbf{W}$ و بردارهای $\mathbf{y}$ و $\mathbf{a}$ به صورت زیر تعریف می‌شوند:

$$\mathbf{X} = \begin{pmatrix} 1 & (x_1-x^*) & \dots & (x_1-x^*)^p \\ 1 & (x_2-x^*) & \dots & (x_2-x^*)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n-x^*) & \dots & (x_n-x^*)^p \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} w(\frac{x_1-x^*}{h}) & 0 & \dots & 0 \\ 0 & w(\frac{x_2-x^*}{h}) & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w(\frac{x_n-x^*}{h}) \end{pmatrix},$$

$$\mathbf{a} = \begin{pmatrix} a_0(x^*) \\ \vdots \\ a_p(x^*) \end{pmatrix}.$$

نکته‌ای که باید در این‌جا به آن توجه کنیم این است که ماتریس $\mathbf{X}$ که در این روش به کار می‌بریم، تابعی از x^* است. در واقع در رگرسیون چندجمله‌ای، معادله‌ی رگرسیونی به کل داده‌هایی که در مدل وجود دارند، برازش داده می‌شود؛ در حالی که در رگرسیون چندجمله‌ای موضعی، معادله‌ی رگرسیونی را به داده‌هایی که در همسایگی x^* وجود دارند، برازش می‌دهیم.

با توجه به رابطه‌ی (۱۳) مقدار $\hat{m}(x^*,x^*)$ برابر با $\hat{a}_0(x^*)$ است که با به کارگیری فرم ماتریسی می‌توانیم آن را به صورت زیر بنویسیم:

$$\hat{m}(x^*,x^*) = \mathbf{e}_1'\left((\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}\right),$$

که در آن $\mathbf{e}_1$ بردار ستونی $(1,0,0,\dots,0)'$ با $p+1$ عضو است. فرم دیگر نمایش $\hat{m}(x^*,x^*)$ به صورت زیر است:

$$\hat{m}(x^*,x^*) = \mathbf{q}(x^*)'\mathbf{Y} = \sum_{j=1}^{n} q_j(x^*)y_j, \quad (۱۶)$$

که در آن $\mathbf{q}(x^*)$ از رابطه‌ی زیر پیروی می‌کند:

$$\mathbf{q}(x^*)' = \mathbf{e}_1'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}. \quad (۱۷)$$

حال دو طرف رابطه‌ی (۱۷) را از سمت راست بردار در $\mathbf{X}$ ضرب می‌کنیم، داریم:

$$\mathbf{q}(x^*)'\mathbf{X} = \mathbf{e}_1'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{e}_1'.$$

با توجه به این نکته که x_{i1} ها، $1 \le i \le n$ همگی برابر با یک هستند، رابطه‌ی زیر برقرار است:

$$\sum_{i=1}^{n} q_i(x^*) = 1. \qquad (18)$$

با توجه به روابط(۱۶) و (۱۸) نتیجه می‌گیریم که برآورد حاصل از روش رگرسیون چندجمله‌ای موضعی، همان میانگین وزنی Y_j ها است. اگر $x^* = x_k$ آنگاه

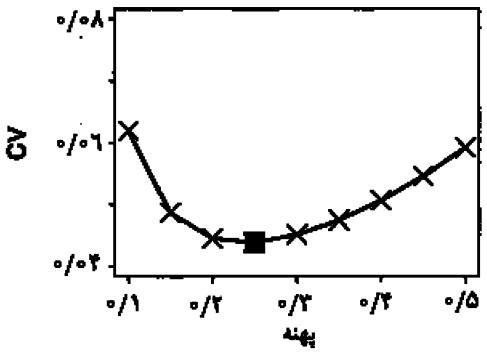
$$\begin{aligned} \hat{Y}_k &= \hat{m}(x_k, x_k) \\ &= \sum_{j=1}^{n} q_j(X_k) Y_j, \end{aligned}$$

که در آن $q_j(x_k)$، kj امین عضو ماتریس هت است. در این حالت k امین ستون ماتریس $\mathbf{X}'$ مانند بردار $\mathbf{e}'_1$ است و می‌توانیم k امین عضو روی قطر ماتریس هت را به‌صورت زیر تعریف کرد:

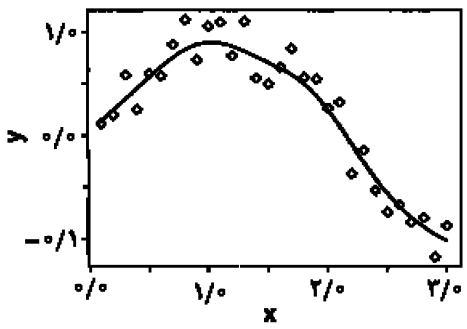
$$\begin{aligned} h_{kk} &= q_k(X_k) \\ &= [(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}]_{11}.w(0). \end{aligned}$$

با توجه به ماتریس هت روش رگرسیون چندجمله‌ای موضعی، مقادیر CV و GCV را برای تعیین پهنه‌ی h بهینه به کار می‌بریم. برای بررسی بیشتر جزئیات ذکر شده در این قسمت به [۸] صفحات ۱۵۱–۱۴۱ مراجعه کنید.

مثال ۳–۳ در این‌جا به داده‌های مثال قبلی، نموداری را به وسیله‌ی روش رگرسیون خطی موضعی برازش می‌دهیم. ابتدا با کمک نمودار مربوط به CV مقدار h بهینه را انتخاب می‌کنیم. این نمودارها در شکل زیر نمایش داده شده‌اند.



شکل ۸. نتایج محاسبه‌ی CV برای داده‌های شکل ۶ در روش رگرسیون خطی موضعی.



شکل ۹. برازش نمودار به وسیله‌ی روش رگرسیون خطی موضعی.

همان طور که می‌بینید، در شکل ۸ مقدار h بهینه با شاخص CV برابر با ۰/۲۵ تعیین شده است. با توجه به مقادیر h که در بالا معین شد و با بهره‌گیری از تابع وزنی چگالی نرمال، در شکل ۹ نموداری را به داده‌ها برازش می‌دهیم. برازش با خط پیوسته برای داده‌هایی که به وسیله‌ی ◇ مشخص شده‌اند، نمایش داده شده است.

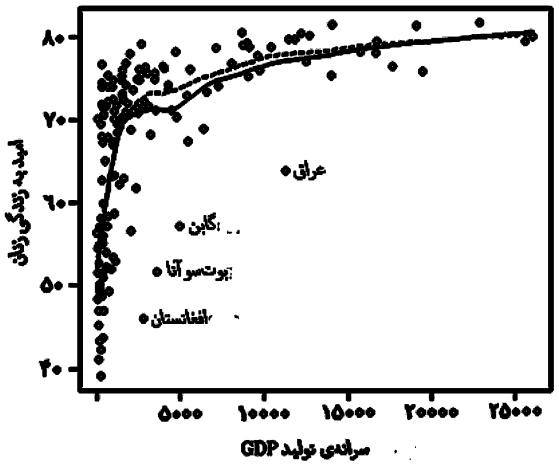
از مقایسه‌ی شکل ۹ با شکل ۶ متوجه می‌شویم که نمودار برازش داده شده در شکل ۹ کمتر تحت تأثیر نقاطی که داده‌ها در آن کراندارند، قرار گرفته است که این ویژگی از جمله امتیازات روش رگرسیون چندجمله‌ای موضعی

نسبت به رگرسیون نادارایا - واتسون است.

این مقاله را با ارائه‌ی مثال دیگری به پایان می‌رسانیم. در این مثال با بررسی گام‌به‌گام آن چه برای رسم یک نمودار در حالت رگرسیون ناپارامتری ساده انجام می‌دهیم، برآنیم تا درک درستی از این روش در ذهن مخاطبین ایجاد کنیم. البته رسم نمودار با روش‌های ناپارامتری به صورت دستی کار عاقلانه‌ای نیست و در عمل غیر ممکن است و در این موارد از برنامه‌های کامپیوتری کمک می‌گیریم[۸].

مثال ۳–۴ در شکل ۱۰ داده‌های مربوط به امید زندگی زنان و سرانه‌ی تولید ناخالص داخلی (GDP)[۱۴] نمایش داده شده‌اند. هدف، بررسی ارتباط بین امید به زندگی زنان با سرانه‌ی تولید GDP است. داده‌های مذکور از ۱۵۴ کشور جهان در سال ۱۹۹۸ توسط سازمان ملل جمع‌آوری شده است. در این شکل دو نمودار را به وسیله‌ی روش رگرسیون خطی موضعی به داده‌ها برازش داده‌ایم. نمودار پیوسته برازش را روی همه‌ی داده‌ها انجام داده است در حالی که نمودار نقطه‌چین چهار کشوری را که روی شکل مشخص شده‌اند، از قلم انداخته است.

واضح است که گرچه ارتباط بین امید به زندگی و GDP، مثبت است؛ این ارتباط کاملاً غیرخطی است.



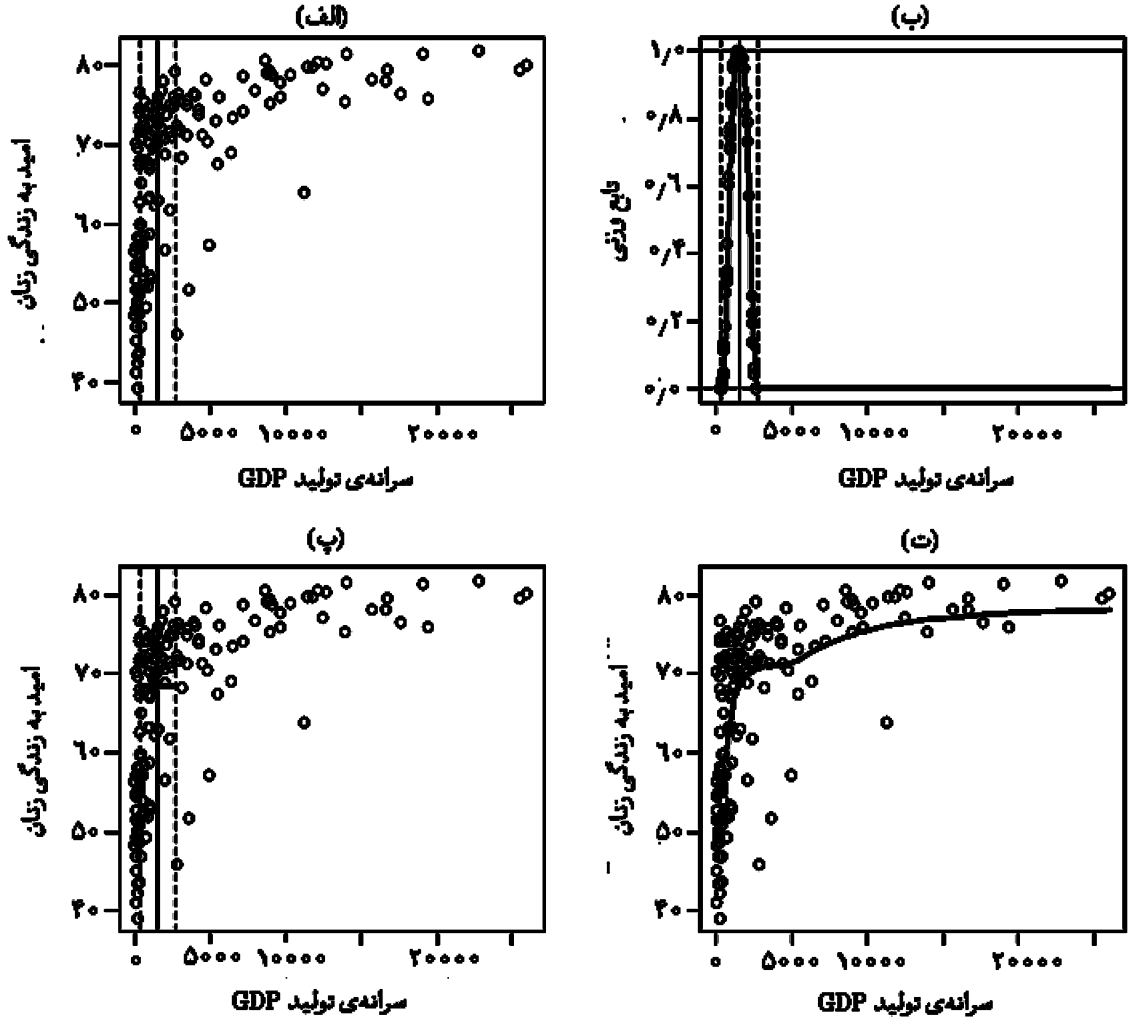
شکل ۱۰. برازش نمودار به وسیله‌ی روش رگرسیون خطی موضعی.

حال می‌خواهیم به وسیله‌ی رگرسیون نادارایا - واتسون به صورت گام‌به‌گام، نموداری را به این داده‌ها برازش دهیم.

۱) در گام اول یکی از مقادیر x_i ها را انتخاب می‌کنیم؛ برای مثال می‌توانید x_i ها را به صورت صعودی مرتب کرده و سپس یکی یکی آن‌ها را انتخاب کنیم. به x_i که در این مرحله انتخاب می‌کنیم، اصطلاحاً مقدار x کانونی می‌گوییم و آن را با x_0 (یا همان طور که در متن به کار گرفته شد x^*) نمایش می‌دهیم. در این‌جا برای مثال $x_{(80)}$ را انتخاب می‌کنیم.

حال نوبت به انتخاب پهنه‌ی h می‌رسد. برای انتخاب h راه‌های مختلفی وجود دارد؛ برای مثال می‌توانید m امین همسایگی نزدیک x_0 را در نظر بگیرید. در این‌جا h نصف طول بازه‌ی در برگیرنده‌ی m امین همسایگی نزدیک x_0 خواهد بود. این ناحیه در شکل ۱۱(الف) مشخص شده است. دقت کنید که هر چه h بزرگ‌تر باشد، نمودار حاصل هموارتر خواهد شد.

[۱۴] Gross Domestic Product



شکل ۱۱. برازش نمودار به وسیله‌ی روش برآوردگر نادارایا - واتسون.

۲) در گام دوم نوبت به انتخاب تابع وزنی می‌رسد. این تابع، باید یک تابع تک مدی متقارن باشد. انتخاب توابع وزنی متفاوت تأثیری در روند کلی کار ندارند و به دلخواه می‌تواند انتخاب شود. در این‌جا از تابع زیر به عنوان تابع وزنی بهره می‌گیریم. (به عنوان یک تابع وزنی دیگر می‌توانستیم تابع چگالی نرمال را به کار گیریم.)

$$w(\tfrac{x_i-x_\circ}{h}) = \begin{cases} \left(1-(\tfrac{|x_i-x_\circ|}{h})^3\right)^3 & \text{if } (\tfrac{|x_i-x_\circ|}{h})^3 \le 1 \\ 0 & \text{if } (\tfrac{|x_i-x_\circ|}{h})^3 > 1. \end{cases}$$

در شکل ۱۱(ب) منحنی تابع وزنی مورد استفاده ارائه شده است.

۳) در گام سوم با به کارگیری تابع وزنی که در گام دوم معرفی کردیم و با بهره‌گیری از داده‌های y_i، مقادیر $\hat{y}_i$ را با استفاده از رابطه‌ی زیر به‌دست می‌آوریم.

$$\hat{y}_\circ = \hat{m}(x_\circ)$$

$$= \frac{\sum w(x_i,h)y_i}{\sum w(x_i,h)}.$$

به شکل ۱۱ (پ) توجه کنید.

۴) در گام چهارم، مراحل سه‌گانه‌ی مذکور را برای تمام مقادیر x_i ها انجام می‌دهیم. نمودار شکل ۱۱ (ت) که برآورد تابع رگرسیونی مورد نظر است، حاصل می‌شود.

مراجع

[۱] رضوی پاریز، ا. (۱۳۸۲)، مقدمه‌ای بر تحلیل رگرسیون خطی، ترجمه، انتشارات دانشگاه شهید باهنر کرمان.

[۲] طالبی، ه. (۱۳۷۷)، رگرسیون کلاسیک و مدرن با کاربرد آن، ترجمه، انتشارات دانشگاه شهرکرد.

[3] Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots*, Sage, Thousand Oaks CA.

[4] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman and Hall, London.

[5] Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating nonparametric regression based on marginal integration, *Biometrika,* 82, 93-100.

[6] Nielsen, J.P. and Sperlich, S. (2005). Smooth backfitting in practice. *J. Roy. Statist. Soc.-Series B,* 67, 43- 61.

[7] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

[8] Takezawa, K. (2006). *Introduction to Nonparametric Regression*, Wiley, New York.