رگرسیون ناپارامتری

در این فصل میخواهیم در مورد رگراسیون ناپارامتری صحبت کنیم.جفت مشاهدات داده شده مانند $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ داریم در معادله ی زیر به Y متغیر پاسخ میگویند که واسته به متغیر x است.

$$Y_i = r(x_i) + \epsilon_i \quad (۱)$$

که به r تابع رگرسیون میگویند.ما می خواهیم عملکردی را تحت فرضیات ضعیف تخمین بزنیم.که برآوردگر $r(x)$ را به صورت $\hat{r}(x)$ نشان میدهیم و همچنین ترجیح میدهیم که برآورد ما هموار باشد.و با توجه به این فرض که $\mathbb{V}(\epsilon_i) = \sigma^2$ به x وابسته نیست.

ما فرمول بالا به گونه ایی رفتار میکردیم که که همبستگی متغیرهای x_i ثابت باشد،اگر به صورت تصادفی باشند داده ها به صورت $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
و $r(x)$ به صورت میانگین شرطی Y بر حسب $X = x$ تعریف میشود و خواهیم داشت.

$$r(x) = \mathbb{E}(Y|X = x) \quad (۲)$$

مثال ۱ داده های CMB فراخوانی کنید و دو تا نمودار را رسم کنید. [۱] محور افقی شکل لحظه ی چند قطبی فرکانس نوسانات در زمینه ی دما و محور عمودی قدرت نوسانات در هر فرکانس میباشد.

[۱] داده های تابش زمینه کیهانی که دارای متغیر های ۱)لحظه ی چند قطبی فرکانس نوسانات در زمینه ی دما (گشتاور چند قطبی اندازه ایی از بار) ۲)قدرت نوسانات در هر فرکانس ۳)موج صدا برآورد شده ۴)انحراف معیار برآورد شده میباشد

روشی که در این فصل ما در نظر گرفتیم

شیوه ی اول رگرسیون محلی:برای فیت کردن توابع و یا سطوح رگسیون در داده ها

۱ رگرسیون هسته

۲ رگرسیون چند جمله ایی محلی

شیوه ی دوم شیوه ی تنبیه کردن یا مجازات کردن:مدل خطی استاندارد(روش حداقل مربعات) در شرایطی ضعیف عمل میکند ،در جاهایی که تعداد متغیر ها بیشتر از تعداد نمونه هاست که در این شرایط رگرسیون مجازات شده مناسب تر است.باضافه کردن یک محدودیت در معادله متغیر هارا مجازات میکند و از تعداد آن ها میکاهد .نتیجه ی این عمل کاهش مقادیر ضریب به سمت صفر

۱ رگرسیون ریج

۲ رگرسیون لاسو

۳ رگرسیون شبکه‌ی الاستیک

۱ مروری بر رگسیون خطی و لجستیک

فرض کنید داده هایی مانند:
$(x_1, Y_1), \ldots, (x_n, Y_n)$ $Y_i \in \mathbb{R}$ $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ داریم که و مدل رگرسیون خطی آن برابر است با:

$$Y_i = r(x_i) + \epsilon_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \qquad i = 1, \ldots, n \tag{۳}$$

که در عبارت فوق $\mathbb{E}(\epsilon_i) = 0$ و $\mathbb{V}(\epsilon_i) = \sigma^2$ برقرار است.
ماتریس X به صورت زیر است:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

بردار $\mathcal{L}$ مجموعه ایی است از ترکیبات خطی ستون های ماتریس X ،با داشتن مقادیر $Y = (Y_1, \ldots, Y_n)^T$،$\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$،$\beta = (\beta_1, \ldots, \beta_p)^T$ عبارت (۳) به صورت زیر نوشته میشود

$$Y = X\beta + \epsilon \tag{۴}$$

برآورد حداقل مربعات $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ برداری است که مجموع باقیمانده مربعات خطارا مینیمم میکند

$$RSS = (Y - X\beta)^T(Y - X\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

پس ما به دنبال مینیمم کردن عبارت فوق هستیم.
فرض کنید که X^TX معکوس پذیر باشد،آنگاه برآورد حداقل مربعات به صورت زیر است

$$\hat{\beta} = (X^TX)^{-1}X^TY \quad (۵)$$

و برای برآورد $r(x)$ توسط $x = (x_1, \ldots, x_p)^T$ داریم

$$\widehat{r_n}(x) = \sum_{j=1}^{p} \hat{\beta}_j x_j = x^T\hat{\beta}$$

به این ترتیب مقادیر فیت شده $r = (\widehat{r_n}(x_1), \ldots, \widehat{r_n}(x_n))^T$ را میتوان به صورت زیر نوشت:

$$r = X\hat{\beta} = LY \quad (۶)$$

$$L = X(X^TX)^{-1}X^T \quad (۷)$$

که به ماتریس ایجاد شده ماتریس هت گفته میشود.اعضای بردار $\hat{\epsilon} = Y - r$ را باقیمانه مینامند.

ماتریس هت متقارن است یعنی $L = L^T$

ماتریس هت خودتوان است یعنی $L^2 = L$

که طبق (۷) داریم و هرگاه ماتریس ما خودتوان باشد داریم که رتبه L برابراست با $tr(L)$ یعنی داریم

$$\begin{aligned} rank(X(X^TX)^{-1}X^T) &= tr(X(X^TX)^{-1}X^T) \\ &= tr(X^TX(X^TX)^{-1}) \\ &= tr(I_p) \\ &= p \end{aligned}$$

با توجه به ۱ خواهیم داشت:

$$\widehat{r_n}(x) = \ell(x)^TY = \sum_{i=1}^{n} \ell_i(x)Y_i \quad (۸)$$

که داریم

$$\ell(x)^T = x^T(X^TX)^{-1}X^T$$

و برای برآورد ناریب σ^2 خواهیم داشت

$$\hat{\sigma^2} = \frac{\sum_{i=1} n(Y_I - \widehat{r_n}(x_i))^2}{n-p} = \frac{||\hat{\epsilon}||^2}{n-p} \qquad (۹)$$

برای پیدا کردن یک فاصله اطیمنان برای $r(x)$ باید تابع های $a(x), b(x)$ به گونه ایی باشند که

$$\mathbb{P}(a(x) \leq r(x) \leq b(x)) \geq 1 - \alpha \qquad (۱۰)$$

با توجه به (۸) برای به دست آوردن واریانس $\widehat{r_n}(x)$ خواهیم داشت:

$$\mathbb{V}(\widehat{r_n}(x)) = \sigma^2 \sum_{i=1}^{n} \ell_i^2(x) = \sigma^2 ||\ell(x)||^2$$

که پیشنهاد میشود به فرم زیر مورد استفاده قرار بگیرد:

$$I(x) = (a(x), b(x) = (\widehat{r_n}(x) - c\hat{\sigma}||\ell(x)||, \widehat{r_n}(x) + c\hat{\sigma}||\ell(x)||) \qquad (۱۱)$$

که c مقداری ثابت است.اگر تابع توزیع $F_{p,n-p}$ داشته باشیم که F تابع توزیع برای p متغیر و با درجه آزادی $n-p$ میباشد،آنگاه با تعریف یک کران بالا مانند : $F_{\alpha;p,n-p}$ آنگاه خواهیم داشت:
$\mathbb{P}(F_{p,n-p} > F_{\alpha;p,n-p}) = \alpha$

قضیه ۱ در فاصله اطمینان در رابطه ی (۱۱) اگر به جای مقدار c رابطه ی $c = \sqrt{pF_{\alpha;p,n-p}}$ قرار بگیرد آنگاه در رابطه ی (۱۰) هم صدق میکند.

۲ رگرسیون لجستیک

یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی مانند بیماری یا سلامت، مرگ یا زندگی است. این مدل را می‌توان به عنوان مدل خطی تعمیم‌یافته‌ای که از تابع لوجیت (لگاریتم طبیعی نسبت به شانس) [۲] به عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به‌حساب‌آورد. منظور از دو سویی بودن، رخ داد یک واقعه تصادفی در دو موقعیت ممکنه است.تفاوت مهم مدل رگرسیون خطی و رگرسیون لجستیک در دو ویژگی رگرسیون لجستیک می‌تواند دیده شود. اول توزیع شرطی $y|\vec{x}$ یک توزیع برنولی به جای یک توزیع گوسی است چونکه متغیر وابسته دودویی است. دوم مقادیر پیش‌بینی احتمالاتی است و محدود بین بازه صفر و یک و به کمک تابع توزیع لجستیک بدست می‌آید رگرسیون لجستیک احتمال خروجی پیش‌بینی می‌کند.

---

[۲] $logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \qquad i = 1, \ldots, n$

تفاوت رگرسیون خطی با رگرسیون لجستیک

۱. برای مدل کردن متغیرهائی که مقادیر محدودی به خود میگیرند بهتر از رگراسیون خطی عمل میکند زیرا مدل خطی هر مقداری را در خروجی تولید میکند درحالی که برای چنین متغیرهائی مقادیر محدودی مورد نیاز است.

۲. در رگراسیون خطی مقدار متغیر مورد نظر از ترکیب خطی متغیرهای مستقل بدست می آید در حالیکه در لجستیک رگراسیون از ترکیب خطی تابع logit استفاده میشود.

۳. در رگراسیون خطی پارامترها به روش least squares بدست می آیند در حالیکه این روش برای لجستیک رگراسیون فاقد کارائی بوده و از روش $\text{maximum likelihood estimation}$ برای پیدا کردن پارامترها استفاده میشود.

وقتی Y_i ها پیوسته نباشد،ممکن است رگراسیون خطی معمولی مناسب نباشد. برای نمونه فرض کنید $Y_i \in \{0,1\}$ در این حالت مدلی که استفاده میشود مدل رگرسیون لجستیک است.

$$p_i = p_i(\beta) = \mathbb{P}(Y_i = 1) = \frac{e^{\sum_j \beta_j x_{ij}}}{1 + e^{\sum_j \beta_j xij}} \qquad (۱۲)$$

میتوانیم با جدا کردن $x_{i1} = 1$ برای همه iها، در این مدل Y_i دارای توزیع برنولی با میانگین p_i میباشد نسبت درست نمایی پارامتر $\beta = (\beta_1, \ldots, \beta_p)^T$ اگر تابع احتمال به صورت $f(y) = p^y(1-p)^{1-y}$ باشد. برابر خواهد بود با:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} \qquad (۱۳)$$

بیشترین نسبت درست نمایی برآوردگر $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ را از الگوریتم زیر میتوان به دست آورد. بدین صورت است که ابتدا با یک $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ شروع میکنیم و در گذاشتن در فرمول بالا p_i را محاسبه میکنیم سپس با انجام الگوریتم زیر و تکرار آن تا آن جایی که همگرا شود. الگریتم زیر از سه مرحله تشکیل شده

مرحله ی اول مجموعه ی زیر را تعریف میکنیم:

$$Z_i = \log\left(\frac{p_i}{1-p_i}\right) + \frac{Y_i - p_i}{p_i(1-p_i)} \qquad i = 1, \ldots, n$$

مرحله ی دوم برای برآوردβ داریم

$$\hat{\beta} = (X^T W X)^{-1} X^T W Z$$

که در عبارت بالا W ماتریس قطری است که عناصر روی قطر اصلی آن برابر است با: $p_i(1-p_i)$.این بدان معناست که انجام دادن رگرسیون خطی Z روی X

مرحله ی سوم حساب کردن p_i با استفاده از برآوردگر $\hat{\beta}$

رگرسیون خطی و رگرسیون لجستیک مورد های خاص کلاسی به نام مدل های خطی تعمیم یافته هستند.

۳ هموار کننده ی خطی

همه ی برآورد های ناپارامتری در این فصل هموار کننده ی خطی هستند.

تعریف ۱ برآورگر $\widehat{r_n}$ از r یک هموار کننده ی خطی است هرگاه :

$$\forall x, \quad \exists \ell(x) = (\ell_1(x), \dots, \ell_n(x))^T, \qquad \widehat{r_n}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i \tag{۱۴}$$

که برای فیت کردن r با توجه به $Y = (Y_1 \dots, Y_n)^T$ خواهیم داشت:

$$\boldsymbol{r} = (\widehat{r_n}(x_1), \dots, \widehat{r_n}(x_n))^T \tag{۱۵}$$
$$\boldsymbol{r} = LY \tag{۱۶}$$

که در عبارت بالا L یک ماتریسی مربعی $n \times n$ میباشد که سطر i ام توسط $\ell(x_i)^T$ به جود می‌آید یعنی هر درآیه ی آن به صورت: $L_{ij} = \ell_j(x_i)$ می‌باشد.سطر i ام نشان‌دهنده ی وزن هایی است که به هر Y_i برای برآورد $\widehat{r_n}(x_i)$ میدهد.

تعریف ۲ ماتریس L ماتریس هموارساز یا هت ماتریس گفته می شود. i امین سطر از L کرنل موثر براي برآورد کردن $r(x_i)$ گفته می شود. مانند (۱) درجه‌ی آزادی موثر را به صورت زیر تعریف میکنیم:

$$\upsilon = tr(L) \tag{۱۷}$$

نکته:وزن هاي موجود در همه هموارسازها که ما از آنها استفاده می کنیم یک خصوصیت دارند که

$$\forall x, \sum_{i=1}^{n} \ell_i(x) = 1$$

این نشان می دهد که هموارساز، منحنی را ثابت نگه می دارد. پس اگر $Y_i = c$ برای همه ی i ها آنگاه : $\widehat{r_n}(x) = c$

مثال ۲ (رگرسوگرام).فرض کنید که $a \le x_i \le b \qquad i = 1, \dots, n$ که در بازه ی (a, b) است را به m بلوک مساوی تقسیم میکنیم که هر بلوک را با $\beta_1, \beta_2, \dots, \beta_m$ نمایش میدهیم. $\widehat{r_n}(x)$ را به صورت زیر تعرف میکنیم.

$$\widehat{r_n}(x) = \frac{1}{k_j} \sum_{i: x_i \in \beta_j} Y_i, \qquad \forall x_i \in \beta_j \tag{۱۸}$$

که k_j ها تعداد نقاط موجود در β_j ها میباشد. به عبارت دیگر برآورد $\widehat{r_n}$ تابعی است که توسط میانگین گیري کردن Y_i ها در هر بلوك به دست آمده است. این برآورد، رگرسوگرام گفته می شود.
حال تابع $\ell(x)$ را به صورت زیر تعریف میکنیم:

$$\ell_i(x) = \begin{cases} \frac{1}{k_j} & x_i \in \beta_j \\ 0 & o.w \end{cases}$$

که خواهیم داشت

$$\ell(x)^T = (0, 0, \ldots, 0, \frac{1}{k_j}, \ldots, \frac{1}{k_j}, 0, \ldots, 0)$$

براي دیدن شکل ماتریس هموارساز L فرض کنید که $n = 9, m = 3, k_1 = k_2 = k_3 = 3$ سپس برای ماتریس هت خواهیم داشت:

$$L = \frac{1}{3} \times \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

ماتریس فوق ماتریس مربعی است که $\upsilon = tr(L) = m$ درجه ی آزادی موثر میباشد.پهناي هر بلوك به صورت $h = \frac{(b-a)}{m}$ می باشد که نشان می دهد، برآورد به چه میزان هموار است. برای بدست آوردن هت ماتریس ،خود صورت سوال گفته که بازه را به سه قسمت مساوی تقسیم کنید و در هر بازه سه نقطه وجود دارد و در کل تعداد نقاط برابر است با ۹ تا پس یعنی سطر اول هت ماتریس سه نقطه ی اول در β_1 وجود دارد پس در سطر اول سه درآیه اول آن یک سوم میشود و همین طور تا آخر.

مثال ۳ میانگین محلی. $h > 0$ را در نظر بگیرید و مجموعه ایی مانند زیر داریم: $\beta_x = i : |x_i - x| \leq h$. و n_x تعداد نقاطی است که در β_x وجود دارد.برای هر x که $n_x > 0$ باشدبه صورت زیر تعریف میشود.

$$\widehat{r_n}(x) = \frac{1}{n_x} \sum_{i \in \beta_x} Y_i.$$

که یک برآورد میانگین محلی از $r(x)$ است.در این مورد $\widehat{r_n}(x) = \sum_{i=1}^{n} Y_i \ell_i(x)$ که تابع $\ell(x)$ به صورت زیر تعریف میشود.

$$\ell_i(x) = \begin{cases} \frac{1}{n_x} & |x_i - x| \leq h \\ 0 & o.w \end{cases}$$

حالا در یک حالت ساده تصور کنید که $n = 9, n_i = \frac{i}{9}, h = \frac{1}{9}$. که با توجه به موارد بالا ماتریس هت به صورت زیر خواهد شد.

$$L = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

سطر اول ماتریس مربوط به مشاهده اول است یعنی وزن y_1 را تعیین میکند. سطر دوم ماتریس مربوط به مشاهده دوم و وزن y_2 را تعیین میکندو همین طور تا آخر.
در نهایت با میانگین گیری وزنی برآورد تابع ناپارامتری تعیین میشود،یعنی

$$\widehat{r_n}(x) = \frac{1}{n_x} \sum_{i=1}^{n} y_i \ell_i(x)$$

برای حل مثال داریم ابتدا مقدار $i = 1$ را قرار میدهیم و خواهیم داشت.

$$\beta_1 = \{1, |x_1 - x| \leq \frac{1}{9}\}$$

$$x_1 = \frac{1}{9} \Rightarrow |\frac{1}{9} - x| \leq \frac{1}{9}$$

$$0 \leq x \leq \frac{2}{9}$$

$$x_2 = \frac{2}{9} \Rightarrow |\frac{2}{9} - x| \leq \frac{1}{9}$$

$$\frac{1}{9} \leq x \leq \frac{3}{9}$$

$$\vdots$$

$$\beta_1 = [0, \frac{2}{9}] \Rightarrow \{1/9, 2/9\} \rightarrow n_x = 2$$

$$\ell_1(x) = \begin{cases} \frac{1}{2} & i = 1, 2 \\ 0 & o.w \end{cases}$$

$$\beta_2 = [\frac{1}{9}, \frac{3}{9}] \Rightarrow \{1/9, 2/9, 3/9\} \rightarrow n_x = 3$$

$$\ell_2(x) = \begin{cases} \frac{1}{3} & i = 1, 2, 3 \\ 0 & o.w \end{cases}$$

$$\vdots$$

و همین طور تا آخر،که ماتریس L ساخته میشود.

۴ انتخاب پارامتر هموار

در برآورد $r(x)$ به مشاهده ی X_i وزن بیشتری و به نقاط دور تر وزن کمتری بدهد،معیاری که بتواند تعیین کند کدام مشاهده دور تر است و چه وزنی باید بدهد و کدام مشاهده نزدیک است و چه وزنی باید بگیرد را پهنای باند یا پارامتر هموارساز میگویند. پس ما نیاز داریم راهی برای انتخاب مناسب ترین h انتخاب کنیم. میانگین مربعات خطارا به صورت زیر تعریف میکنیم:

$$R(h) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{r_n}(x_i) - r(x_i))^2\right) \quad (۱۹)$$

ما باید h را به گونه ایی انتخاب کنیم تا $R(h)$ مینیمم شود،اما $R(h)$ وابسته است به تابع ناشناخته ایی به اسم $r(x)$ ،در عوض ما برآورد تابع ریسک را مینیمم میکنیم.
در حدس اول : اولین برآوردي که به ذهن می رسد، میانگین توان دوم خطاي مانده ها می باشد، به صورت زیر است:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}((Y_i - \widehat{r_n}(x_i))^2 \quad (۲۰)$$

این برآورد $R(h)$ را کمتر از مقدار واقعی برآورد می کند، چون از یک مجموعه داده ها دوبار استفاده میشود یعنی یک بار برای براورد $r_n(x_i)$ و بعد از برآورد آن دوباره از همون داده ها برای برآورد $R(h)$ به خاطر همین خطارا افزایش میدهد و برآورد را کمتر از مقدار واقعی نشان میدهد. در واقع مینیمم سازي عبارت $\sum_{i=1}^{n}(Y_i - \widehat{r_n}(x_i))^2$ براي تعیین h باعث می شود که $\hat{r}(x)$ طوري تخمین زده شود که $\hat{R}(h)$ مینیمم شود و این امر باعث کم کردن $R(h)$ می شود. به خاطر این مشکلی که به وجود میآید باید از اعتبارسنجی متقابل استفاده کرد.

قضیه ۲ ترک یک داده نمره اعتبار سنجی متقابل را به صورت زیر تعریف میکنیم:

$$CV = \widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}_{(-i)}(x_i))^2 \tag{۲۱}$$

که $\widehat{r}_{(-i)}$ برآوردی است که با حذف جفت i ام یعنی دوتایی (x_i, Y_i).

همان طور گفته شد تعریف قبلی در مورد $\widehat{r}_{(-i)}$ ناقص است پس خواهیم داشت:

$$\widehat{r}_{(-i)}(x) = \sum_{i=1}^{n} Y_i \ell_{j,(-i)}(x) \tag{۲۲}$$

$$\ell_{j,(-i)}(x) = \begin{cases} 0 & \text{if} j = i \\ \frac{\ell_j(x)}{\sum_{k\neq j}\ell_k(x)} & \text{if} j \neq i \end{cases} \tag{۲۳}$$

به معنای دیگر با قرار دادن وزن x_i به صفر و قرار دادن مجموع وزن های دیگر به یک.
برای درک اعتبار سنجی متقابل خواهیم داشت:

$$\begin{aligned}\mathbb{E}(Y_i - \widehat{r}_{(-i)}(x_i))^2 &= \mathbb{E}(Y_i - r(x_i) + r(x_i) - \widehat{r}_{(-i)}(x_i))^2 \\ &= \sigma^2 + \mathbb{E}(r(x_i) - \widehat{r}_{(-i)}(x_i))^2 \\ &\approx \sigma^2 + \mathbb{E}(r(x_i) - \widehat{r}_n(x_i))^2\end{aligned}$$

در رابطه ی بالا $Y_i = r(x_i) + \epsilon$.
از این رو خواهیم داشت

$$\text{پیش‌بینی خطر} = \mathbb{E}(\hat{R}) \approx R + \sigma^2 \tag{۲۴}$$

که در رابطه ی فوق σ^2 میانگین وزنی ϵ میباشد. در نمره اعتبار سنجی متقابل برآورد خطا تاثیر ندارد،نااریب از خطر است. به نظر می رسد که ممکن است ارزیابی $\hat{R}(h)$ وقت گیر باشد زیرا ظاهراً نیاز داریم که پس از رها کردن هر مشاهده ، برآوردگر را پس بگیریم. خوشبختانه یک فرمول میانبر برای محاسبات وجود دارد. براي ارزیابی (۲۱) وقت بسیاري نیاز است، چون نیاز به محاسبه مجدد برآوردگر پس ازحذف هر مشاهده می باشد.

قضیه ۳ اگر $\widehat{r_n}$ یک هموارکننده خطی باشد.آنگاه نمره اعتبار سنجی متقابل $\hat{R}(h)$ را میتوان به صورت زیر نوشت:

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{r_n}(x_i)}{1 - L_{ii}}\right)^2 \tag{۲۵}$$

که در آن $L_{ii} = \ell_i(x_i)$ عضو i ام قطری هت ماتریس میباشد.

پس پارامتر هموارساز h با مینیمم کردن $\hat{R}(h)$ قابل انتخاب است.

هشدار: شما نمی توانید فرض کنید $\hat{R}(h)$ همیشه از حداقل مشخصی برخوردار باشد. پس باید همیشه $\hat{R}(h)$ را به عنوان تابعی از h ترسیم کنید.

یک جایگزین به جای به حداقل رساندن نمره اعتبار سنجی متقابل به طور تقریبی اعتبار متقاطع کلی تعمیم یافته است که هر l_{ii} در معادله (۲۵) قرار میگیرد،با میانگین $n^{-1}\sum_{i=1}^{n} L_{ii} = \upsilon/n$ که $\upsilon = tr(L)$ درجه آزادی موثر میباشد.پس باید رابطه ی زیر را مینیمم کنیم:

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{r_n}(x_i)}{1-\upsilon/n}\right)^2 \qquad (۲۶)$$

پهنای باند میتواند نمره اعتبار سنجی متقابل عمومی را به حداقل برساند. با استفاده از $(1-x)^{-2} \approx 1+2x$ رابطه ی بالا به صور طقریبی برابر میشود با:

$$GCV(h) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r_n}(x_i))^2 + \frac{2\upsilon\hat{\sigma}^2}{n} \qquad (۲۷)$$

که در رابطه ی بالا $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{r_n}(x_i))$ میباشد.که معادله ی بالا معیاری برای انتخاب متغیر در رگرسیون خطی میباشد.به طور کلی بسیاری از معیار های انتخاب پهنای باند را به صورت زیر نوشت:

$$B(h) = \Xi(n,h) \times \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r_n}(x_i))^2 \qquad (۲۸)$$

در شرایط مناسب انتخاب h باید مینیمم کند $B(h)$ را حال اگر $\hat{h}_0$ مینیمم کند تابع زیان $L(\hat{h}) = n^{-1}\sum_{i=1}^{n}(\widehat{r_n}(x_i) - r(x_i))^2$ و h_0 ریسک را مینیمم کند. سپس همه ی $\hat{h}, \hat{h}_o, h_0$ تمایل دارند به صفر در نرخ $n^{-1/5}$ و همچنین برای برخی ثابت های مثبت مانند $C_1, C_2, \sigma_1, \sigma_2$

$$n^{3/10}(\hat{h} - \hat{h}_0) \rightsquigarrow N(0,\sigma_1^2) \qquad n(L(\hat{h}) - L(\hat{h}_0)) \rightsquigarrow C_1\chi_1^2$$
$$n^{3/10}(h_0 - \hat{h}_0) \rightsquigarrow N(0,\sigma_2^2) \qquad n(L(h_0) - L(\hat{h}_0)) \rightsquigarrow C_2\chi_1^2$$

بنابراین میزان نسبت همگرایی $\hat{h}$ برابر است با

$$\frac{\hat{h} - \hat{h}_0}{\hat{h}_0} = O_p\left(\frac{n^{3/10}}{n^{1/5}}\right) = O_p(n^{-1/10})$$

این سرعت آهسته نشان می دهد که تخمین پهنای باند دشوار است.این سرعت ذاتی است برای مشکل انتخاب پهنای باند از آنجا که این نیز صادق است.

$$\frac{\hat{h}_0 - h_0}{h_0} = O_p\left(\frac{n^{3/10}}{n^{1/5}}\right) = O_p(n^{-1/10})$$

۵ رگرسیون محلی

جال مجدد به رگرسیون محلی ناپارامتری برمیگردیم.فرض کنید x_i ها اسکالر باشد،مدل رگراسیون (۱) را در نظر بگیرید.در این قسمت با در نظر گرفتن برآورد گر ها با توجه به میانگین وزن Y_i ها وزن بیشتری به نقاطی که نزدیک تر است میدهد.ما با برآوردگر رگرسیون هسته شروع می کنیم.

$$E(Y|X=x)=\int yf(y|x)dy=\int y\frac{f(x,y)}{f(x)}$$

با توجه به عبارت زیر است:

$$\hat{f}(x,y)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)K_h(y-y_i),\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x-x_i)$$

با جاگذاری در عبارت اول خواهیم دید که

$$\int y\frac{\sum_{i=1}^{n}K_h(x-x_i)K_h(y-y_i)}{\sum_{j=1}^{n}K_h(x-x_j)}dy$$

$$\int yK_h(y-y_i)dy=Y_i$$

تعریف ۳ اگر $h>0$ یک عدد مثبت باشد،که پهنای باند نامیده میشود،یک برآورد هسته نادارایا و واتسون است هر گاه:

$$\widehat{r_n}(x)=\sum_{i=1}^{n}\ell_i(x)Y_i \qquad (۲۹)$$

که $\ell_i(x)$ به صورت زیر میباشد.که در فرمول زیر K یک هسته میباشد.

$$\ell_i(x)=\frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{x-x_j}{h}\right)} \qquad (۳۰)$$

مثال ۴ با فراخوانی داده های CMB چهار رگرسیون مختلف هسته را نمایش میدهد،که بر اساس افزایش پهنای باند میباشد. هر چه پهنای باند کمتر باشد اتصالات خشن است و اریبی کوچک و واریانس بزرگ و هرچه پهنای باند بزرگ‌تر باشد تخمین ها همار تر اریبی کم و واریانس زیاد تر است.

انتخاب هسته K زیاد مهم نیست .برآوردهای به دست آمده با استفاده هسته های مختلف معمولاً از نظر عددی بسیار مشابه هستند.این مشاهدات توسط محاسبات نظری تأیید شده است که نشان می دهد خطر بسیار غیر حساس است.آنچه اهمیت بیشتری دارد انتخاب است پهنای باند h که میزان صاف کردن را کنترل می کند. ما به پهنای باند اجازه خواهیم داد که به اندازه نمونه بستگی داشته باشد .قضیه زیر چگونگی تأثیر پهنای باند بر برآوردگر را نشان می دهد.

قضیه ۴ ریسک در هسته برآودگر ناداریا واتسون به صورت زیر است.

$$R(\widehat{r_n}, r) = \overbrace{\frac{h_n^4}{4}\left(\int x^2 K(x)dx\right)^2 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx}^{\text{مربع اریبی}} + \overbrace{\frac{\sigma^2 \int K^2(x)dx}{nh_n}\int \frac{1}{f(x)}dx + o(nh_n^{-1}) + o(h)n^4}^{\text{واریانس}} \tag{۳۱}$$

به طوری که $h_n \to 0, nh_n \to \infty$

آنچه در عبارت بالا مهم است وجود عبارت زیر است:

$$2r'(x)\frac{f'(x)}{f(x)} \tag{۳۲}$$

که در قسمت اریبی وجود دارد.
عبارت بالا را نقشه اریبی مینامیم،زیرا بستگی به نقشه دارد،یعنی به توزیع x_i ها.این بدان معناست که اریبی حساس به موقعیت x_i ها است.می توان نشان داد که برآوردگرهای هسته همچنین در نزدیکی مرزها اریبی بالایی دارند.در ادامه خواهیم دید میتوان اریبی مرز هارا با استفاده از رگرسیون چند جمله ایی کاهش داد.
اگر عبارات (۳۱) را مشتق بگیریم و برابر صفر کنیم پهنای باند بهینه محاسبه میشود.

$$h_* = \left(\frac{1}{n}\right)^{1/5}\left(\frac{\sigma^2 \int K^2(x)dx \int \frac{dx}{f(x)}}{\int x^2K^2(x)dx)^2 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2}\right)^{1/5} \tag{۳۳}$$

در نتیجه داریم: $h_* = O(n^{-1/5})$ با قرار دادن h_* در معادله ی (۳۱) خواهیم دید که ریسک با سرعت $O(n^{-4/5})$ در حال کاهش است.در مدل های پارامتری ماکسیمم نسبت درستنمایی کاهش میدهد ریسک را تا صفر با سرعت $1/n$ پس مناسب است از روش های پارامتری استفاده شود .در عمل نمیتوان پهنای باند را از معادله ی (۳۳) بدست آورد زیرا وابسته به تابع ناشناخته ی rمیباشد در عوض از اعتبار سنجی متقابل استفاده میکنیم.

مثال ۵ نمره اعتبار سنجی متقابل را بر روی داده های CMB در ازای تاثیر درجه ی آزادی انجام دهید. بهترین پارامتر هموار آن است که این نمره را کاهش مینیمم کند.

۶ رگسیون چند جمله ایی

برآوردگرهای هسته از اریبی مرزی رنج می برند. این مشکلات را می توان با استفاده از رگرسیون هسته به نام رگرسیون چند جمله ای محلی کاهش داد.برای شروع باید $a = \widehat{r_n}(x)$ را به گونه ایی انتخاب کرد که عبارت زیر را مینیمم کند. $\sum_{i=1}^{n}(Y_i - a)^2$ تابع ثابت $\widehat{r_n}(x) = \overline{Y}$ تخمین خوبی از $r(x)$ نیست.اکنون با معرفی تابع وزن : $w_i(x) = K(\frac{x_i - x}{h})$ حال باید $a = \widehat{r_n}(x)$ را به گونه ایی انتخاب کنیم،تامقدار وزنی مربعها را به حداقل برساند.

$$\sum_{i=1}^{n} w_i(x)(Y_i - a)^2 \qquad (۳۴)$$

با محاسابات خواهیم داشت:

$$\widehat{r_n}(x) = \frac{\sum_{i=1}^{n} w_i(x)Y_i}{\sum_{i=1}^{n} w_i(x)}$$

این یک برآوردگر ثابت محلی است ، از حداقل مربعات دارای وزن محلی بدست آمده .که نشان میدهدبه جای یک رگرسیون ثابت باید از یک رگرسیون چند جمله ایی استفاده کنیم .بگذارید x مقداری ثابت باشد که می خواهیم $r(x)$ را تخمین بزنیم. برای همسایگی x ارزشی به اسم u که به صورت چند جمله ایی به صورت زیر تعریف میشود.

$$P_x(u;a) = a_0 + a_1(u-x) + \frac{a_2}{2!}(u-x)^2 + \cdots + \frac{a_p}{p!}(u-x)^p \qquad (۳۵)$$

میتوان به صورت تقریبی برآورد $r(u)$ را در همسایگی x به صورت زیر نوشت

$$r(u) \approx P_x(u;a). \qquad (۳۶)$$

ما با برآورد $a = (a_0, \ldots, a_p)^T$ با انتخاب $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)^T$ برای مینیمم کردن مجموع مربعات وزن های محلی

$$\sum_{i=1}^{n} w_i(x)(Y_i - P_x(X_i;a))^2 \qquad (۳۷)$$

برآورد a چون به x وابسته است پس مناسب تر است به صورت : $\hat{a}(x) = (\hat{a}_0(x), \ldots, \hat{a}_p(x))^T$ نوشته شود.اگر بخواهیم این وابستگی را واضح تر بیان کنیم به صورت $\widehat{r_n}(u) = P_x(u;a)$ در حالت خاص اگر $u = x$ باشد آنگاه

$$\widehat{r_n}(x) = P_x(x;\hat{a}) = \hat{a}_0(x) \qquad (۳۸)$$

اگرچه که $\widehat{r_n}(x)$ فقط وابسته $\hat{a}_0(x)$ میباشد اما این بدین معنی نیست که به یک مقدار ثابت فیت میشود. با قرار دادن $p = 0$ برآوردگر هسته را باز میگرداند.مقدار مخصوص که $p = 1$ است رگرسیون محلی خطی نامیده میشود. برای پیدا کردن $\hat{a}(x)$ ابتدا داریم:

$$X_x = \begin{bmatrix} 1 & x_1 - x & \dots & \frac{(x_1-x)^p}{p!} \\ 1 & x_2 - x & \dots & \frac{(x_2-x)^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \dots & \frac{(x_n-x)^p}{p!} \end{bmatrix} \tag{۳۹}$$

ما در اینجا W_x را ماتریس قطری $n \times n$ مینامیم که درایه ی (i, i) آن از $w_i(x)$ تشکیل شده باشد.پس میتوانیم بنویسیم:

$$(Y - X_x a)^T W_x (Y - X_x a) \tag{۴۰}$$

حال با مینیمم کردن عبارت وزن حداقل مربعات بدست می‌آید.

$$\hat{a}(x) = (X_x^T W_x X_x)^{-1} X_x^T W_x Y \tag{۴۱}$$

به طور خاص اگر $\widehat{r_n}(x) = \hat{a}_0(x)$ آنگاه فقط سطر اول معادله ی (۴۰) باید استفاده شود.

قضیه ۵ برآورد رگرسیون چند جمله ایی محلی به صورت زیر میباشد:

$$\widehat{r_n}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i \tag{۴۲}$$

که داریم $\ell(x)^T = (\ell_1(x), \dots, \ell_n(x))$ آنگاه

$$\ell(x)^T = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x$$

که $e_1 = (1, 0, \dots, 0)^T$ میباشد.میانگین و واریانس برآورد هم به صورت زیر میباشد.

$$\mathbb{E}(\widehat{r_n}(x)) = \sum_{i=1}^{n} \ell_i(x) r(x_i)$$

$$\mathbb{V}(\widehat{r_n}(x)) = \sigma^2 \sum_{i=1}^{n} \ell_i(x)^2 = \sigma^2 ||\ell(x)||^2$$

ک بار دیگر ، تخمین ما یک سطح صاف خطی است و ما می توانیم با به حداقل رساندن فرمول اعتبار سنجی متقابل که در قضیه ۳ ارائه شده است ، پهنای باند را انتخاب کنیم.

مثال ۶ با فراخوانی داده های LIDAR که در فصل چهار با آن ها روبرو شده بودیم. چهار نمودار به دست آمده که نمودار سمت چپ بالا داده ها میباشد با فیت کردن توسط رگرسیون خطی محلی با $h \approx 37$ که با اعتبار سنجی متقابل بدست آمده.نمودار بالا سمت راست مقدار باقی‌مانده ها هست،نمودار پایین سمت چپ برآورد $\sigma(x)$ میباشد و نمودار پایین سمت راست یک فاصله اطمینان ۹۵ درصدی میباشد.

قضیه ۶ هموارسازی خطی محلی:اگر $\widehat{r}_n(x) = \sum_{i=1}^n \ell_i(x) Y_i$ که $p = 1$

$$\ell_i(x) = \frac{b_i(x)}{\sum_{i=1}^n b_j(x)}$$

$$b_i(x) = K\left(\frac{x_i - x}{h}\right)(S_{n,2}(x) - (x_i - x)S_{n,1}(x)) \quad (۴۳)$$

که داریم:

$$S_{n,j}(x) = \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)(x_i - x)^j, j = 1, 2.$$

مثال ۷ با فراخوانی داده ی CMB برای $p = 0, p = 1$ رگرسیون محلی را رسم میکنیم. که نمودار پایینی زوم شده به کران سمت چپ توجه داشته باشید برای $p = 0$ فیت کردن در نزدیکی کران ها ضعیف است که از اریب بودن کران ها است.

مثال ۸ تابع دوپلر به صورت زیر تعریف میشود:

$$r(x) = \sqrt{x(1-x)}sin\left(\frac{2.1\pi}{x + .05}\right), 0 \leq x \leq 1 \quad (۴۴)$$

این تابع برای برآورد بسیار دشوار میباشد اما یک مورد مناسب برای تست روش رگرسیون ناپارامتری است.در فضایی ناهمگن میباشد یعنی هموار (مشتق دوم)آن نسبت به x متفاوت است.
نمودار های آن رسم شده است .که نمودار بالا سمت چپ نمودار خود تابع میباشد سمت راست تعداد ۱۰۰۰ داده شبیه‌سازی شده است از فرمول $Y_i = r(i/n) + \sigma\epsilon_i$ با توجه به مقادیر زیر $\sigma = .1, \epsilon_i \sim N(0, 1)$ نمودار پایین سمت چپ: نمره اعتبار متقابل سنجی اثر گذاری درجه ی آزادی که با توجه به این نمودار مشخص شد که مینیمم آن حدودا در نقطه ی ۱۶۶ اتفاق می‌افتد با پهنای باند 005. و نمودار پایین سمت راست که نمودار فیت شده ی آن می‌باشد.خط فیت شده چون دارای درجه آزادی موثر بالایی هست پس از هوش بالایی برخوردار میباشد،زیرا تلاش برآن است که در نوسانات سریع در نزدیکی صفر بهترین فیت را داشته باشیم اگر از هوار کننده بیشتر استفاده میکردیم قسمت راست فیت شده میتوانست مناسب تر باشد هرچند ساختار در نزدیکی صفر ازبین میرفت.این مشکلی است که در عملکرد های ناهمگن فضایی وجود دارد که در فصل نه بحث میشود.

درقضیه ی زیر رفتار نمونه بزرگ را در ریسک رگرسیون خطی محلی نشان میدهد و چرا رگرسیون خطی محلی از رگسیون هسته مناسب تر است.

قضیه ۷ اگر

$$Y_i = r(x_i) + \sigma(X_i)\epsilon_i \qquad fori = 1, \ldots, n \qquad a \leq X_i \leq b$$

را داشته باشیم و $X_1, \ldots, X_n$ یک نمونه باشد که دارای توزیع با تابع چگالی f باشد و شرط های زیر وجود داشته باشد

۱. $f(x) > 0$

۲. $f, r'', \sigma^2.$ در نزدیکی x پیوسته باشد.

۳. $h_n \to 0, nh_n \to \infty$

و اگر $x \in (a, b)$ باشد،و داده های $X_1, \ldots, X_n$ معلوم باشدآنگاه برآوردگر خطی محلی و برآوردگر هسته دارای واریانس

$$\frac{\sigma^2(x)}{f(x)nh_n} \int K^2(u)du + o_p\left(\frac{1}{nh_n}\right) \tag{۴۵}$$

و اریبی برآوردگر هسته نادارایا و واتسون به صورت

$$h_n^2 \left(\frac{1}{2}r''(x) + \frac{r'(x)f'(x)}{f(x)}\right) \int u^2 K(u)du + o_p(h^2) \tag{۴۶}$$

میباشد،در حالی که اریبی برآوردگر خطی محلی به صورت تقریبی برابر

$$h_n^2 \frac{1}{2} r''(x) \int u^2 K(u)du + o_p(h^2) \tag{۴۷}$$

میباشد.بنابراین برآوردگر خطی محلی بدون نقشه اریبی است. (۳۲) در نقاط مرزی a, b برآوردگر هسته نادارایا واتسون تقریبی دارای اریبی با نظم h_n است در حالی که برآوردگر خطی محلی دارای اریبی با نظم h_n^2 میباشد،این بدان معناست که برآوردگر خطی محلی اریبی مرز هارا از بین میبرد.

نتیجه ی بالا به طور کلی بیشتر برای محلی چند جمله ایی با درجه ی p میباشد.که با گرفتن p به طرز عجیب غریبی نقشه ی اریبی را در مرز هارا بدون افزایش واریانس کاهش میدهد