رگرسیون غیر پارامتری در R

چکیده

در مدل های رگرسیون پارامتری سنتی، شکل عملکرد مدل قبل از تناسب مدل با دادهَ ها مشخص شده است و هدف برآوردکردن پارامترهای مدل می باشد. در مقابل، رگرسیون غیر پارامتری، هدف برآورد عملکرد رگرسیون بطورمستقیم بدون مشخص کردن شکل آن به روش صریح می باشد. فاکس و وایزبرگ (2011) در ضمیمه مقاله، ما توصیف می کنیم چگونه چند نوع مدل رگرسیون غیر پارامتری در R متناسب شود، شامل صاف کننده طرح مجزا، که یک پیشگویی واحد وجود دارد؛ مدل های رگرسیون چندگانه؛ مدل های رگرسیون افزایشی؛ و مدل های غیر پارامتر-رگرسیون کلی که مشابه مدل های خطی تعمیم یافته می باشد.

1 مدل های رگرسیون غیر پارامتری

مدل رگرسیون غیر خطی سنتی (در ضمیمه در رگرسیون غیرخطی توصیف شد) متناسب با مدل ذیل می باشد

$$y = m(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon$$

که ø یک بردار پارامترهای برآورد شده و x یک بردار پیش بینی کننده است؛ اشتباهات به طور عادی و به طور مستقل با میانگین 0 و واریانس ثابت σ فرض و توزیع می شود. تابع m(x,ø) مربوط به مقدار میانگین پاسخ y به پیش بینی کننده ها می باشد، که از قبل مشخص شده است، همانطور که در مدل رگرسیون خطی است.

مدل رگرسیون غیر پارامتری کلی به شیوه ای مشابه نوشته می شود، اما تابع نامشخص است:

$$\begin{aligned} y &= m(\mathbf{x}) + \varepsilon \\ &= m(x_1, x_2, \ldots, x_p) + \varepsilon \end{aligned}$$

برای پیشگو کننده x = (x1; x2; : : : ; xp) . علاوه بر این، هدف از رگرسیون غیر پارامتری برآورد تابع رگرسیون m (x) به طور مستقیم، به جای برآورد پارامترها می باشد. اکثر روش های رگرسیون غیر پارامتر به

طور ضمنی فرض می کنند که m یک عملکرد صاف و پیوسته است[1]. همانطور که در رگرسیون غیر خطی، استاندارد است فرض می شود که $\varepsilon_i \sim \mathrm{NID}(0, \sigma^2)$.

مورد خاص مهم از مدل عمومی، رگرسیون ساده غیر پارامتری است که تنها یک پیش بینی کننده وجود دارد:

$$y = m(x) + \varepsilon$$

رگرسیون ساده غیر پارامتر اغلب به نام نرم کننده طرح مجزا نامیده می شود زیرا کاربرد مهم، رسیدن به یک منحنی صاف از طریق یک صفحه پراکنده Y به نسبت x می باشد. ما اغلب از رگراسیون غیر پارامتری بدین شکل در بدنه متن استفاده می کنیم.

زیرا سخت است تا مدل های رگراسیون غیرپارامتری کلی مناسب باشد، زمانیکه پیشگو کننده های بسیاری وجود دارد و زیرا سخت است تا مدل های مناسب را نمایش دهد زمانی که بیش از دو یا سه پیش بینی کننده وجود دارد، مدل های محدود کننده تر توسعه یافته اند. یکی از این مدل ها

$$y = \beta_0 + m_1(x_1) + m_2(x_2) + \cdots + m_p(x_p) + \varepsilon$$

مدل رگرسیون افزودنی است که توابع جزئی رگرسیون mj (xj) فرض می شود صاف باشد و از داده ها تخمین زده می شود. این مدل بسیار محدودتر از مدل رگرسیون غیر پارامتری عمومی است، اما کمتر محدود کننده از مدل رگرسیون خطی است، فرض می شود که تمام بخش های جزئی توابع رگرسیون خطی هستند.

تغییرات در مدل رگرسیون افزایشی شامل مدل های نیمه پارامتری است که در آن تعدادی از پیش بینی ها به صورت خطی وارد می شوند، به عنوان مثال

$$y = \beta_0 + \beta_1 x_1 + m_2(x_2) + \cdots + m_p(x_p) + \varepsilon$$

(به ویژه مفید است زمانی که بعضی از پیش بینی کننده ها، عوامل هستند)، و مدل هایی که برخی از پیش بینی کننده ها وارد تعاملات می شوند، که به عنوان مثال به عنوان اصطلاحات با ابعاد بزرگ در مدل ظاهر می شود.

---

[1] به استثنای فرض ضمنی صافی بودن، رگراسیون موجک است، که در این ضمیمه بحث نشده است که در R اجرا می شود، به عنوان مثال، در بسته موج؛ به ناسون و سیلورمن (1994، 2000)؛ ناسون (2008) مراجعه کنید.

همه این مدل ها به طور مستقیم به رگرسیون غیر پارامتری عمومی، به طور گسترده ای مدل های خطی برای مدل های خطی تعمیم یافته (که در فصل 5 مورد بحث قرار گرفت) گسترش می یابد. اجزای لینک و تصادفی در مدل های خطی تعمیم یافته هستند، اما پیشگویی کننده خطی ازGLM

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

به عنوان مثال، با یک عملکرد صاف غیرمستقیم پیش بینی کننده

$$\eta = m(x_1, x_2, \ldots, x_p)$$

برای عمومی ترین مورد، و یا با مجموع توابع رگرسیون سهمی صاف

$$\eta = \beta_0 + m_1(x_1) + m_2(x_2) + \cdots + m_p(x_p)$$

در مدل افزایشی تعمیم یافته جایگزین می شود.

2. برآورد

روش های متعددی برای تخمین مدل های رگرسیون غیر پارامتری وجود دارد که ما دو مورد توصیف خواهیم کرد: رگرسیون چندجمله ای محلی و اسپیلین های صاف. با توجه به پیاده سازی این روش ها درR ، شرمندگی زیادی به همراه خواهد داشت:

- رگرسیون چندجمله ای محلی با استفاده از تابع استاندارد لس R انجام می شود (به صورت محلی با صاف کننده طرح مجزا وزنی، برای پرونده ساده رگرسیون) و لس (رگرسیون محلی، بصورت کلی تر)
- برآورد رگراسیون-ساده نوار-صاف توسط تابع استاندارد R نوار-صاف انجام می شود
- رگراسیون کلی غیرپارامتریک با برآورد احتمالی محلی (که رگراسیون محلی مورد خاصی برای مدلهایی با خطای عادی هستند) که در بسته لاک فیت (تناسب محلی) (لورد 1999) اجرا می شود که برآورد چگالی را انجام می دهد.
- مدل های افزایشی عمومی ممکن است متناسب با عملکرد گروهی هستی و تیبشیرانی((1990) در بسته گروهی باشد)، که از صاف کننده اسپلین یا محلول رگرسیون محلی استفاده می کند. عملکردgam در بسته بندی وود mgcv (2000، 2001، 2006) بخشی از توزیع استاندارد R است، همچنین این کلاس از مدل ها با استفاده از صاف کننده اسپلین و ویژگی های انتخاب اتوماتیک پارامترهای صاف کننده نیز استفاده می

شود (نام بسته روش مورد استفاده برای انتخاب پارامترهای هموار حاصل می شود: چندین اعتبارسنجی متقابل تعمیم یافته).

- چندین بسته R دیگر برای رگرسیون غیر پارامتری وجود دارد شامل بومان وآزالینی (1997) بسته sm نرم کننده که رگرسیون محلی و برآورد احتمال محلی را انجام می دهند و همچنین شامل امکانات برای تخمین چگالی غیر پارامتری می باشد؛ و گو (2000) بسته sm (اسپلاین صاف کردن عمومی) که متناسب با مدل های رگرسیون تعمیم یافته و رگراسیون اسپلاین صاف کننده مختلف می باشد. این لیست جامع نیست!

2.1 رگرسیون چندجمله ای محلی

2.1.1رگرسیون ساده

در اینجا ما به دنبال تناسب با این مدل هستیم

$$y = m(x) + \varepsilon$$

اجازه بدهید ما بر روی ارزیابی عملکرد رگرسیون در یک مقدار خاص x- x0 تمرکز کنیم. در نهایت، ما مدل را در یک طیف نماینده از مقادیر x یا صرفا در مشاهداتn ، .xi متناسب خواهیم بود. ما ادامه می دهیم تا یک رگرسیون چندجمله ای با وزن حداقل-مربع مرتبه p از y درx انجام دهیم،

$$y = b_0 + b_1(x - x_0) + b_2(x - x_0)^2 + \cdots + b_p(x - x_0)^p + e$$

ارزیابی مشاهدات در رابطه با نزدیکی آنها به ارزش کانونیx0 ؛ وزن معمولی تابع از تابع tricube استفاده می کند:

در زمینه کنونی، ,z = (x - x0)=hکه h نیم عرض یک پنجره محصور شده مشاهدات برای رگرسیون محلی است. مقدار متناسب درx0 ، یعنی ارتفاع پیش بینی شده از منحنی رگرسیون، صرفا توسط b0 0 = است، به راحتی با داشتن مرکز پیش بینی x در مرکز تولید شده است.

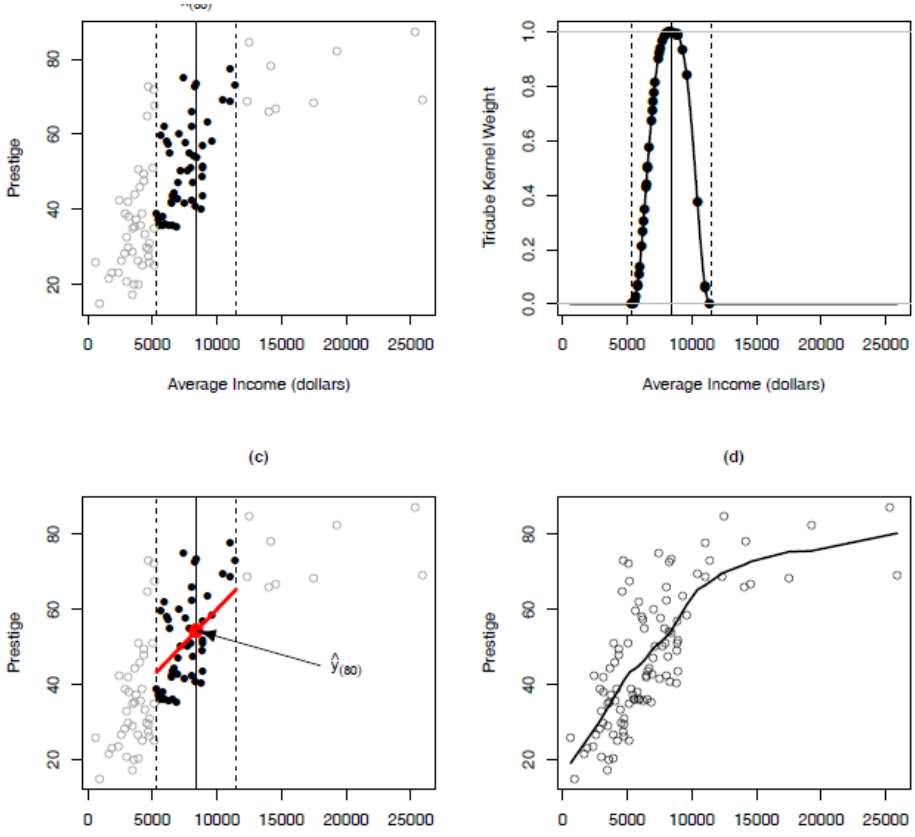
روال است که h تنظیم می شود تا اینکه هر رگرسیون محلی شامل یک مقدار ثابت s داده ها می باشد. سپس، Sدامنه صاف کننده رگرسیون محلی نامیده می شود. طول بزرگتر، نتیجه نرمتر در مقابل، ترتیب بزرگتر رگرسیون های محلی می باشد، لذا دامنه و مرتبه رگرسیون های محلی راحت تر است تا به صورت یک طرفه به فروش برسد.

روند تناسب رگرسیون محلی در شکل 1 نشان داده شده است، با استفاده اطلاعات معتبر شغلی کانادایی در متن فصل 2 بیان شده است. ما رگرسیون اعتبار را در درآمد، در ابتدا تمرکز بر مشاهده با 80 مقدار بزرگ درآمد، x (80)، در شکل 1 توسط خط عمودی جامد نشان داده شده است[۲].

- یک پنجره شامل نزدیکترین 50– همسایگان x) (80) (یعنی برای فاصله2 = 1 = 102 = 50 = S)

در شکل 1 aنشان داده شده است.

- وزن tricube برای مشاهدات در این همسایگی در شکل 1 bنشان داده شده است.



شکل 1: رگرسیون خطی محلی اعتبار بر درآمد برای داده های اعتبار شغلی کانادا: (الف) خطوط تجزیه شده 50 نزدیکترین همسایگان (80) xدر خط عمودی محکم) را تعیین حدود می کند. (ب) وزن های Tricube برای مشاهدات در محدوده.(80) x بود ج) رگرسیون خطی محلی وزنی در محله(80) x ؛ نقطه جامد قرمز بزرگ تر متناسب با مقدار (80)توسط (80) بالاتر از (80) x است(د) رگرسیون خطی محلی تکمیل شده، مقادیر متناسب در محدوده x متصل می کند.

[۲] Cf شکل 7.13 (صفحه 345) در متن، برای شکل مشابهی که بیانگر رگرسیون هسته ی نزدیکترین همسایه می باشد.

- شکل c1 خط رگرسیون وزن محلی متناسب با داده ها در محله 0x نشان می دهد (یعنی، یک رگرسیون چندجمله ای محلی مرتبه 1p =)؛ مقدار متناسب $\widehat{y}|x_{(80)}$ در این گراف به عنوان نقطه جامد بزرگتر نمایش داده می شود.

- در نهایت، در شکل d1، رگرسیون های محلی برای طیف وسیعی از مقادیر x تخمین زده و مقدار متناسب در یک منحنی رگرسیون غیر پارامتری متصل می شوند.

شکل d1 توسط دستورات R زیر تولید می شود، با استفاده از تابع لس:

```
> library(car) # for data sets
> plot(prestige ~ income, xlab="Average Income", ylab="Prestige", data=Prestige)
> with(Prestige, lines(lowess(income, prestige, f=0.5, iter=0), lwd=2))
```

استدلال f به لس دامنه ای از رگرسیون محلی نرم تر می دهد؛ iter = 0 مشخص می کند که رگرسیون های محلی نباید برای مشاهدات بیرونی از نظر کمبود وزن دوباره متناسب شود.[۳]

2.1.2 رگرسیون چندگانه

مدل رگرسیون چندگانه غیر پارامتری عبارتند از

$$\begin{aligned} y &= f(\mathbf{x}) + \varepsilon \\ &= f(x_1, x_2, \ldots, x_p) + \varepsilon \end{aligned}$$

در حال گسترش رویکرد محلی چندجمله ای به رگرسیون چندگانه، مفهومی ساده است، اما میتواند برای مشکلات عملی اجرا شود.

- اولین گام این است که محدوده چند متغیره در اطراف یک نقطه کانونی 0x را تعریف کنیم رویکرد پیش فرض در تابع لس، استفاده از مقادیر مسافت اقلیدسی است:

$$D(\mathbf{x}_i, \mathbf{x}_0) = \sqrt{\sum_{j=1}^{k} (z_{ij} - z_{0j})^2}$$

[۳] پیش فرض، لس تکرارپذیری قدرتمند iter=3، با استفاده از یک تابع وزن مجذور مربع را انجام می دهند. ایده مشاهدات وزن گذاری برای به دست آوردن برآوردگرهای رگرسیون قوی در ضمیمه در رگرسیون قوی شرح داده شده است. متناوبا، از سوی دیگر، یکی می تواند از تابع نرم لس برای دریافت مختصات برای یک چند ضلعی محلی صاف یا یا صاف پراکنده برای رسم نمودار استفاده کند.

که zj پیش بینی کننده های استاندارد شده است،

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}$$

در اینجا xi یک بردار پیش بینی برای مورد i است؛ xij مقدار پیش بینی کننده jth برای مورد ith؛ xj میانگین پیش بینی کننده j است؛ و sj انحراف استاندارد آن است.

- وزن ها با استفاده از فاصله های مقیاس پذیر تعریف می شوند:

$$w_i = W\left[\frac{D(\mathbf{x}_i, \mathbf{x}_0)}{h}\right]$$

که W (0) یک تابع وزن مناسب است، مانند تری کوب، که مورد h نیمه عرض محله (یعنی، شعاع) است. همانطور که در رگرسیون ساده محلی، h می تواند برای تعریف یک محله از جمله نزدیکترین همسایگان 0x ns[] تنظیم شود (که علامت مربع بر گرد کردن نزدیک ترین عدد صحیح دلالت می کند)

یک رگرسیون چندجمله ای وزن Y را روی x انجام دهید؛ برای مثال، یک خط محلی متناسب با فرم های زیر می باشد:

$$y = b_0 + b_1(x_1 - x_{01}) + b_2(x_2 - x_{02}) + \cdots + b_k(x_k - x_{0k}) + e$$

مقدار نصب شده در 0x صرفا $\widehat{y}_0 = b_0.$ می باشد.

- این روش برای ترکیب نمایشی مقادیر پیش بینی کننده برای ایجاد تصویری از سطح رگرسیون تکرار می شود.

گسترش تصویر بخش قبلی و استفاده از تابع لس، اعتبار ما را در هردو میزان درآمد فرض کنید و سطح تحصیلات شغلها را به حال خود رها کنیم

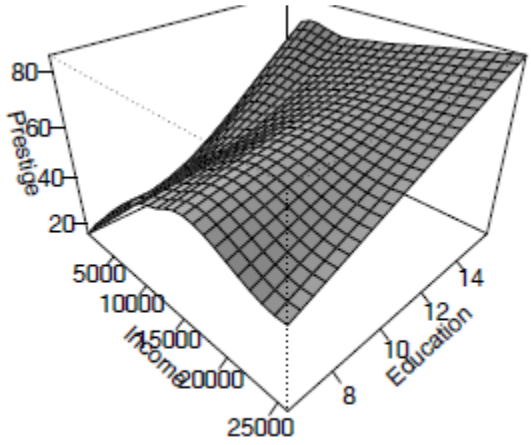
```
> mod.lo <- loess(prestige ~ income + education, span=.5, degree=1, data=Prestige)
> summary(mod.lo)

Call:
loess(formula = prestige ~ income + education, data = Prestige,
    span = 0.5, degree = 1)

Number of Observations: 102
Equivalent Number of Parameters: 8.03
Residual Standard Error: 6.91
Trace of smoother matrix: 10.5

Control settings:
  normalize:  TRUE
  span           :  0.5
  degree   :  1
  family   :  gaussian
  surface  :  interpolate         cell = 0.2
```

تعیین درجه = 1 متناسب با رگرسیون خطی محلی؛ به طور پیش فرض درجه = 2 (یعنی، رگرسیون های محلی درجه دوم است). برای درک کردن طیف وسیعی از استدلال برای تابع لس ، لس مشورت کنید؟ خروجی خلاصه شامل انحراف معیار از باقیمانده های مدل زیر و برآورد تعداد معادل پارامترها (یا درجه آزادی) است که توسط مدل | در این مورد، حدود هشت پارامتر استفاده می شود. در مقابل، مدل رگرسیون خطی استاندارد از سه پارامتر (ثابت و دو دامنه) استفاده خواهد کرد.

همانطور که در رگرسیون ساده غیر پارامتری، برآوردهای پارامتری وجود ندارد: برای درک نتیجه رگرسیون، ما باید سطح رگرسیون نصب شده را به صورت گرافیکی بررسی کنیم، همانطور که در شکل 2 توسط دستورات زیر R تولید می شود:[۴]

```
> inc <- with(Prestige, seq(min(income), max(income), len=25))
> ed <- with(Prestige, seq(min(education), max(education), len=25))
> newdata <- expand.grid(income=inc, education=ed)
> fit.prestige <- matrix(predict(mod.lo, newdata), 25, 25)
> persp(inc, ed, fit.prestige, theta=45, phi=30, ticktype="detailed",
+     xlab="Income", ylab="Education", zlab="Prestige", expand=2/3,
+     shade=0.5)
```

[۴] نمایندگی گرافیکی جایگزین سطح رگرسیون، مانند قطعه های کنتور و قطعه های مشترک وجود دارد. دومی می تواند اجرا شود زمانیکه بیش از دو پیش گو کننده وجود دارد.



شکل 2: سطح متناسب برای رگرسیون چند متغیره خطی اعتبار در درآمد و آموزش می باشد.

ما از تابع شبکه گسترش استفاده می کنیم تا کادر داده ای حاوی ترکیبی از مقادیر دو پیش بینی کننده، درآمد و آموزش می باشد؛ برای هر پیش بینی کننده، ما مقادیر 25 را به طور مساوی در امتداد دامنه متغیر تقسیم می کنیم. سپس، مقادیر متناسب مربوط به سطح رگرسیون، با پیش بینی محاسبه می شوند. این مقادیر پیش بینی شده به ماتریس 25 تا 25 تغییر یافته است که برای تابع persp همراه با مقادیر پیش بینی کننده (inc و ed) تصویب می شود که برای تولید سطح رگرسیون استفاده می شود. استدلالtheta و phi به persp جهت گیری طرح را کنترل می کند؛ کنترل طول نسبی محور z را گسترش می دهد و سایه ، سطح هاشور خورده نمودار را کنترل می کند برای جزئیات بیشتر به persp?مراجعه کنید.

رابطه اعتبار آموزش و درآمد به نظر می رسد، به ویژه در جهت درآمد غیر خطی باشد (به خطوط شبکه در سطح رگرسیون نگاه کنید). رگرسیون جزئی در جهت هر یک از پیش بینی ها به نظر می رسد بسیار تغییر نکند، همانطورکه پیش بینی کننده دیگر متفاوت است، پیشنهاد می کند که مدل افزایشی برای این داده ها ممکن است مناسب باشد. ما چنین مدل زیر را در نظر می گیریم

ما همچنین می توانیم اهمیت آماری هر پیش بینی کننده با حذف آن از مدل و انجام یک تست-F تقریبی افزایشی برای تغییر در مجموع باقیمانده مربعات بررسی کنیم. در تناسب مدل های جداگانه مذکور، دامنه رگرسیون های ساده محلی را به $0.7 \simeq \sqrt{0.5}$: تنظیم کردیم

```
> mod.lo.inc <- loess(prestige ~ income, span=.7, degree=1,
+     data=Prestige)  # omitting education
> mod.lo.ed <- loess(prestige ~ education, span=.7, degree=1,
+     data=Prestige)  # omitting income
```

```
> mod.lo.inc <- loess(prestige ~ income, span=.7, degree=1,
+     data=Prestige)  # omitting education
```

```
> mod.lo.ed <- loess(prestige ~ education, span=.7, degree=1,
+     data=Prestige)  # omitting income
> anova(mod.lo.inc, mod.lo)  # test for education

Model 1: loess(formula = prestige ~ income, data = Prestige,
    span = 0.7, degree = 1)
Model 2: loess(formula = prestige ~ income + education, data = Prestige,
    span = 0.5, degree = 1)

Analysis of Variance:   denominator df 90.66

      ENP   RSS F-value  Pr(>F)
[1,] 3.85 12006
[2,] 8.03  4246    20.8 4.8e-16

> anova(mod.lo.ed, mod.lo)  # test for income

Model 1: loess(formula = prestige ~ education, data = Prestige,
    span = 0.7, degree = 1)
Model 2: loess(formula = prestige ~ income + education, data = Prestige,
    span = 0.5, degree = 1)

Analysis of Variance:   denominator df 90.66

      ENP  RSS F-value  Pr(>F)
[1,] 2.97 7640
[2,] 8.03 4246    7.79 7.1e-08
```

بنابراین، هردو درآمد و آموزش، اثرات قابل توجهی از نظر آماری دارند.

2.2 اسپلاین های صاف

اسپلاین های صاف به عنوان راه حل برای مشکل ساده رگرسیون زیر مطرح می شوند: یافتن عملکرد bm (x)با دو مشتقهٔ پیوسته که مجموع مربعات جریمه را به حداقل می رساند

$$\mathrm{SS}^*(h) = \sum_{i=1}^{n} [y_i - m(x_i)]^2 + h \int_{x_{\min}}^{x_{\max}} [m''(x)]^2 \, dx \qquad (1)$$

که h یک پارامتر صاف است، مشابه با عرض محله از برآوردگر چند جمله ای محلی است.

- اولین اصطلاح در معادله 1، مجموع باقیمانده مربع ها است.

- اصطلاح دوم یک مجازات شدیدی است که بیشتر می شود زمانیکه مشتقات ثانویه یکپارچه از تابع رگرسیون m00 (x) بزرگ است / یعنی زمانی که' m (x) خشن' است (به سرعت شیب در حال تغییر است). انتهای انتگرال داده ها را محصور می کند.
- در حالت افقی، زمانی که ثابت صاف برای h = 0 تنظیم شده است (و اگر همه مقادیر-x متمایز باشند) ، bm (x) صرفا اطلاعات را درهم می زند؛ این شبیه به برآورد محلی رگرسیون با فاصله = n1 = است.
- در حالت افقی دیگر، اگر h بسیار بزرگ باشد، سپس bmانتخاب خواهد شد تا bm00 (x) در همه جا 0 باشد، به این معنی است که حداقل مربعات خطی در سطح جهانی متناسب با داده است (رگرسیون محلی با محله های بسیار گسترده معادل است).

تابع bm (x) که معادله 1 را به حداقل می رساند یک اسپلاین مکعبی طبیعی با گره در مقدار متمایز مشاهده شده x5 می باشد[5]. اگرچه این نتیجه به نظر می رسد که پارامتر n مورد نیاز است (هنگامی که تمام مقادیر x متفاوت هستند)، شدت مجازات محدودیت های بیشتری بر روی راه حل ها تحمیل می کند، که به طور معمول، تعداد معادل پارامترهای اسپلاین صاف را به طور قابل توجهی کاهش می دهد، و از bm (x) در تغیر داده ها جلوگیری می کند.

در واقع، انتخاب پارامتر صاف h به طور غیرمستقیم با تنظیم تعداد معادل پارامترها برای صاف کننده رایج است. زیرا یک تابع هدف صریح برای بهینه سازی وجود دارد، اسپلاین صاف بطور ریاضی از رگرسیون محلی ظریف تر است. به هر حال، کلی کردن اسپلاین های صاف به رگرسیون چندگانه، 6 سخت تر است [6] و اسپلاین-صاف و رگرسیون-محلی با تعداد مشابه معادل پارامترها معمولا بسیار شبیه متناسب هستند.

یک تصویر در شکل 3 ظاهر می شود، مقایسه یک اسپیلین صاف با یک خط محلی متناسب با اجرای تعداد مشابه پارامترهای معادل (درجه آزادی) می باشد. ما از تابع نرم –اسپلاین همراه با یک مدل قبلی لس برای

---

[5] اسپلاین توابع چند جمله ای قطعی هستند که با هم (در گره)؛ برای اسپلاین مکعبی متناسب هستند، مشتقات اول و دوم نیز در گره مداوم هستند. اسپلاین های طبیعی دو گره اضافی را در انتهای داده قرار می دهند و تابع را فراتر از این نقاط خطی مذکور محدود کنید.

[6] پیچیده ترین انواع، مانند اسپلاین های نازک صفحه، به راحتی به رگرسیون چندگانه به طور کلی ساده تر می شوند. به عنوان مثال به گو (2000) مراجعه کنید.

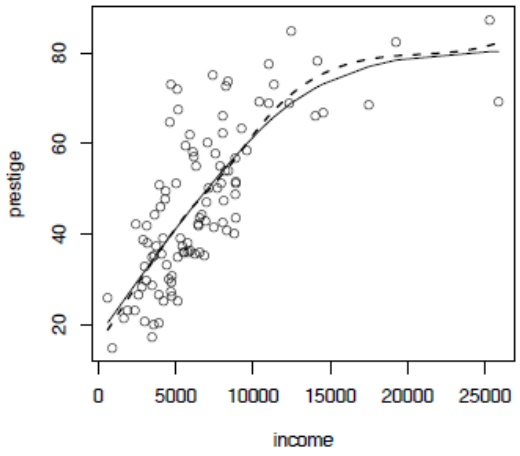
نشان دادن تناسب های جایگزین (هر کدام با 3.85 معادل پارامترها) به رابطه اعتبار با درآمد استفاده می کنیم:

```
> mod.lo.inc  # previously fit loess model

Call:
loess(formula = prestige ~ income, data = Prestige, span = 0.7,
    degree = 1)

Number of Observations: 102
Equivalent Number of Parameters: 3.85
Residual Standard Error: 11.1

> plot(prestige ~ income, data=Prestige)
> inc.100 <- with(Prestige, seq(min(income), max(income), len=100))  # 100 x-values
> pres <- predict(mod.lo.inc, data.frame(income=inc.100))  # fitted values
> lines(inc.100, pres, lty=2, lwd=2)  # loess curve
> lines(with(Prestige, smooth.spline(income, prestige, df=3.85),
+     lwd=2))  # smoothing spline
```

ما رگرسیون خطی محلی را با استفاده از پیش بینی برای محاسبه مقادیر 100 متناسب در دامنه درامد، محاسبه می کنیم. دو مسطح بسیار شبیه هستند: خط تجزیه تناسب خطی محلی است؛ خط جامد اسپلاین صاف است.

2.3 انتخاب پارامتر صاف

هر دو رگرسیون محلی-چند جمله ای و اسپلاین های صاف دارای پارامتر صاف قابل تنظیم هستند. این پارامتر ممکن است با آزمایش و خطای بصری انتخاب شود، با انتخاب یک مقدار که صافی را در مقابل وفاداری به داده ها متعادل می کند.



شکل 3: رگرسیون محلی (خط تجزیه) و اسپلاین-صاف (خط جامد) برای رگرسیون اعتبار درآمد متناسب است.

هر دو مدل از پارامترهای معادل 3.85 استفاده می کند.

روش های رسمی تر انتخاب پارامترهای صاف معمولا برای به حداقل رساندن خطای مجذور مربع تناسب، یا با استفاده از یک فرمول تقریبی از خطای مجذور مربع (مثلا، برآورد به اصطلاح افزونه) یا برخی از فرم های اعتبارسنجی متقابل تست می شود.

در اعتبارسنجی متقابل، داده ها به زیر مجموعه ها تقسیم می شوند (احتمالا شامل مشاهدات فردی)؛ این مدل به طور پیوسته متناسب هر زیر مجموعه را به نوبه خود حذف می کند؛ و سپس مدل متناسب برای پیش بینی پاسخ برای زیر مجموعه چپ استفاده می شود. تلاش برای این روش برای مقادیر مختلف پارامتر صاف مقداری را پیشنهاد خواهد داد که برآورد متقابل اعتبار سنجی خطای مجذور مربع را به حداقل می رساند . زیرا اعتبار سنجی متقابل بسیار محاسباتی است، تقریب ها و تعمیم ها اغلب استفاده می شود (به عنوان مثال، به وود ، 2000، 2004 مراجعه کنید).

2.4 رگرسیون غیر پارامتری افزودنی

مدل رگرسیون غیر پارامتری افزودنی است

$$y = \beta_0 + m_1(x_1) + m_2(x_2) + \cdots + m_k(x_k) + \varepsilon$$

که توابع رگرسیون-سهمی m_j با استفاده از یک رگرسیون ساده صاف تر مانند رگرسیون چند جمله ای محلی یا اسپلاین های صاف متناسب می شوند. ما رگرسیون اعتبار در درآمد و آموزش، با استفاده از تابع gam در بسته mgcv را توضیح دادیم (وود، 2000، 2001، 2004، 2006):

```
> library(mgcv)
> mod.gam <- gam(prestige ~ s(income) + s(education), data=Prestige)
> summary(mod.gam)

Family: gaussian
Link function: identity
```

```
Formula:
prestige ~ s(income) + s(education)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.833      0.689      68   <2e-16

Approximate significance of smooth terms:
             edf Ref.df    F p-value
s(income)    3.12   3.88 15.3 1.7e-09
s(education) 3.18   3.95 38.8 < 2e-16

R-sq.(adj) =  0.836   Deviance explained = 84.7%
GCV score = 52.143  Scale est. = 48.414    n = 102
```

تابع s که در تعیین فرمول مدل استفاده می شود، نشان می دهد که هر شرایط متناسب با اسپلاین صاف است.

درجه آزادی برای هر شرایط توسط اعتبار سنجی متقابل کلی یافت می شود:[۷]

در این مورد، معادل 3: 118 پارامتر برای مدت درآمد استفاده می شود، و 3: 177 برای دوره آموزشی؛ درجه آزادی برای مدل مجموع این اضافه 1 برای رگرسیون ثابت است.

سطح رگرسیون افزودنی در شکل 4 ترسیم شده است:

```
> fit.prestige <- matrix(predict(mod.gam, newdata), 25, 25)
> persp(inc, ed, fit.prestige, theta=45, phi=30, ticktype="detailed",
+     xlab="Income", ylab="Education", zlab="Prestige", expand=2/3,
+     shade=0.5)
```

داده های قبلی، داده های جدید را قاب بندی می کنند، مقادیر پیش بینی شده بر روی سطح رگرسیون جهت یافتن استفاده می شود، زودتر برای طراحی شکل 2 (صفحه 7) برای چندین مدل رگرسیون غیر پارامتری کلی متناسب با داده های مذکور محاسبه می شود. دو تناسب کاملا مشابه هستند. علاوه بر این، به دلیل اینکه بخشهای سطح رگرسیون افزودنی در جهت پیش بینی کننده (نگه داشتن ثابت پیش بینی کننده دیگر) موازی

---

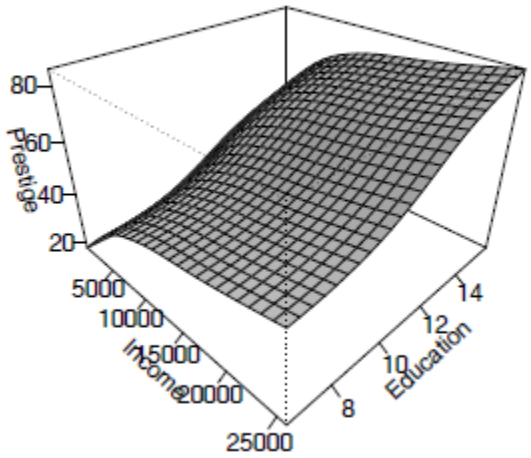
[7] پارامترهای صاف همراه با بقیه مدل برآورد می شود، تعمیم معیار اعتبار سنجی تقابلی را کاهش می دهد که $\widehat{\sigma}^2$ که واریانس خطای تخمینی است و df_{mod} معادل درجه آزادی برای مدل شامل هر دو شرایط پارامتریک و صاف می باشد. در مدل افزایشی تعمیم یافته (در زیر در نظر گرفته شده است) پراکندگی تخمینی ϕ واریانس خطا برآورد شده را جایگزین می کند.

```
> plot(mod.gam)
Press return for next page....
```

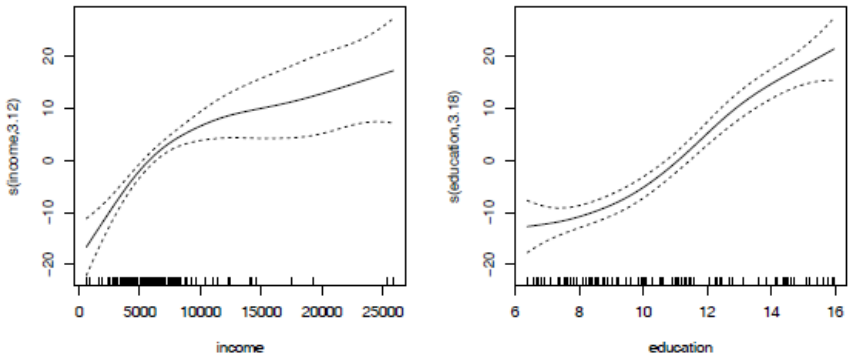
$$\frac{n\widehat{\sigma}^2}{n - df_{mod}}$$

است، آن برای رسم هر تابع جزئی رگرسیون به طور جداگانه کافی است. این مدل فضیلت عملی افزودنی-رگرسیون است: این مسئله رگرسیون چند بعدی (در این مورد، فقط سه بعدی) برای یک سری از دو بعدی گراف های رگرسیون جزئی را کاهش می دهد. روش نمونه برداری برای اجزای gam این نمودار ها را تولید می کند، پاکت اطمینان 95 درصد اطمینان در اطراف تناسب نشان می دهد (شکل 5):

عملکرد gam به طور قابل توجهی عمیق تر از این مثال بیان شده است:



شکل 4: سطح تناسب برای رگرسیون غیر پارامتری افزودنی اعتبار در درآمد و آموزش.



شکل 5: توابع رگرسیون جزئی برای رگرسیون افزایشی اعتبار در درآمد و آموزش. خطوط تجزیه شده حاوی 95 درصد اطمینان در اطراف تناسب هستند.

- این مدل می تواند شامل شرایط صاف (تعامل) در دو یا چند پیش بینی کننده، به عنوان مثال، فرم s (درآمد، آموزش) باشد

- این مدل می تواند نیمه پارامتری باشد، شامل اصطلاح خطی | به عنوان مثال، اعتبار ~ s (درآمد) + آموزش.
- بعضی از گزینه های فنی خاص مانند انواع اسپلاین ممکن است توسط کاربر انتخاب شوند و کاربر می تواند درجه آزادی برای شرایط صاف را ثابت کند.
- همانطور که از نام آن مشخص است (مدل افزایشی تعمیم یافته = GAM)، عملکرد gam به مدل های با خطاهای نرمال و یک لینک هویت محدود نمی شود (به زیر نگاه کنید).

هستی و تیبشیرانی (1990) عملکرد gam در بسته gam را شکار می کنند و عملکرد gam در بسته mgcv متفاوت است. اول، احتمال دارد تا توابع رگراسیون سهمی را می توان از طریق رگرسیون چندجمله محلی متناسب کرد، با استفاده از تابع در فرمول مدل، و همچنین با اسپلاین صاف استفاده از s دوم، پارامتر صاف برای یک اصطلاح (دامنه رگرسیون محلی یا درجه آزادی برای یک اسپلاین صاف) مستقیما به جای تعیین توسط اعتبارسنجی متقابل تعمیم یافته مشخص می شود. همانطور که در بستهmgcv ، عملکرد gam در بسته gam همچنین می تواند به طور کلی مدل های افزودنی تعمیم را متناسب کنند

3. رگرسیون غیر پارامتر کلی

ما رگرسیون غیر پارامتر کلی را با تناسب مدل رگرسیون افزودنی نیمه پارامتریکی لجستیک، به داده های مشارکت نیروی کار مروز بیان خواهیم کرد (در فصل 5 شرح داده شده و در آن بسته ماشین گنجانده شده است). به یاد بیاورید که متغیر پاسخ در این مجموعه داده ها، LFP، یک عامل است که برای زنان در نیروی کار، بله و نه برای کسانی که نیستند، کد گذاری شده است. پیشگویی ها شامل تعداد کودکان 5 ساله یا کمتر (k5)می باشد؛ تعداد فرزندان 6 تا 18 ساله(k618) ؛ سن زن، در سال؛ عوامل نشان می دهد که آیا زن (wc) و شوهرش (hc) در کالج حضور داشتند، بله یا خیر؛ و درآمد خانواده (inc)، به استثنای درآمد همسر و در 1000 دلار داده شده است. ما نادیده گرفتیم متغیر باقیمانده در مجموعه داده، ورودی نرخ دستمزد مورد انتظار همسر، lwg؛ همانطور که در متن توضیح داده شد، تعریف خاص از lwg باعث استفاده از آن مشکل می شود. از آنجا که k5 و k618 گسسته هستند، با مقادیر نسبتا کمی متفاوت هستند، ما با این پیشگویی کننده مذکور به عنوان عوامل معامله خواهیم کرد، مدل سازی آنها به صورت پارامتری، همراه با عوامل wc وhc ؛ همچنین، به

خاطر تنها سه نفر با سه فرزند زیر 5 سال و فقط سه نفر با بیش از5 کودک بین 6 تا 18 ساله، ما از تابع ضبط در بسته بندی خودرو برای ضبط مقادیر غیرمعمول استفاده خواهیم کرد:

```
> remove(list=objects())  # clean up everything
> Mroz$k5f <- factor(Mroz$k5)
> Mroz$k618f <- factor(Mroz$k618)
> Mroz$k5f <- recode(Mroz$k5f, "3 = 2")
> Mroz$k618f <- recode(Mroz$k618f, "6:8 = 5")
> mod.1 <- gam(lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc,
+     family=binomial, data=Mroz)
> summary(mod.1)

Family: binomial
Link function: logit

    Formula:
    lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc

    Parametric coefficients:
                Estimate Std. Error z value Pr(>|z|)
    (Intercept)    0.542      0.180    3.01   0.0026
    k5f1          -1.521      0.251   -6.07  1.3e-09
    k5f2          -2.820      0.500   -5.64  1.7e-08
    k618f1        -0.342      0.227   -1.51   0.1316
    k618f2        -0.279      0.248   -1.12   0.2608
    k618f3        -0.333      0.284   -1.17   0.2415
    k618f4        -0.531      0.440   -1.21   0.2269
    k618f5        -0.491      0.609   -0.81   0.4206
    wcyes          0.980      0.223    4.39  1.1e-05
    hcyes          0.159      0.206    0.77   0.4422

    Approximate significance of smooth terms:
            edf Ref.df Chi.sq p-value
    s(age) 1.67   2.09   26.2 2.4e-06
    s(inc) 1.74   2.19   17.5 0.00020

    R-sq.(adj) =  0.128   Deviance explained = 10.9%
    UBRE score = 0.25363  Scale est. = 1         n = 753
```

برآوردهای ضریب و خطاهای استاندارد را برای پارامتریک بخش مدل نشان می دهد؛ gamخلاصه سازی اهداف) و یک آزمون مهم برای این شرایط؛ و چندین incدرجه آزادی برای هر شرایط صاف (به عنوان مثال برای سن و برای مدل استفاده می شود. [8] UBRE آمار خلاصه، از جمله نمره

تابع anova به یک شی تک گام آزمایش های Wald برای شرایط در این مدل گزارش می کند:

[8] برای یک مدل دو حالته، به صورت پیش فرض، گام UBRE معیار (برآوردگر ریسک بی طرفانه) را به حداقل می رساند (وهبا 1990) به جای معیار .S GCV (inc 1.74)

```
> anova(mod.1)

Family: binomial
Link function: logit

Formula:
lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc

Parametric Terms:
      df Chi.sq p-value
k5f    2  55.61 8.4e-13
k618f  5   3.28    0.66
wc     1  19.26 1.1e-05
hc     1   0.59    0.44
```

اهمیت تقریبی شرایط صاف:

شکل 6: شرایط صاف برای سن و inc در یک مدل افزایشی به طور کلی نیمه پارامتری برای داده های مشارکت نیروی کارMroz .

```
         edf Ref.df Chi.sq p-value
s(age) 1.67   2.09   26.2 2.4e-06
s(inc) 1.74   2.19   17.5 0.00020
```

روش طرح برای اشیای gam نمودار شرایط صاف در مدل، همراه با نقطه نظر پاکت نامه اعتماد 95 درصد (شکل 6):

```
> plot(mod.1)

Press return for next page....
```

خروج از خطی بودن عالی نیست. علاوه بر این، تابع رگراسیون برای inc خیلی زیاد است تقریبا به درستی تخمین زده می شود که در آن مقادیر داده ها کم است و ما احتمالا انتقال خوب inc با لگاریتم گرفتن قبل از تناسب مدل را انجام داده ایم.

یکی از مدل های رگرسیون افزودنی، شامل مدل های افزایشی عمومی، تست کردن غیرخطی بودن است: ما ممکن است برای مقابله با انحراف برای یک مدل ادامه دهیم که متناسب با شرایط غیر پارامتریک با انحراف

است در غیر این صورت مدل یکسان است که متناسب با شرایط خطی است. برای نشان دادن، ما شرایط صاف برای سن در مدل با اصطلاح خطی جایگزین کنید:

```
> mod.2 <- gam(lfp ~ age + s(inc) + k5f + k618f + wc + hc,
+     family=binomial, data=Mroz)
> anova(mod.2, mod.1, test="Chisq")

Analysis of Deviance Table

Model 1: lfp ~ age + s(inc) + k5f + k618f + wc + hc
Model 2: lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1       740        919
2       740        917 0.72     2.21      0.09
```

در غیر اینصورت، ما می توانیم برای غیرخطی بودن در مدل با شرایط خطی تست کنیم:

```
> mod.3 <- gam(lfp ~ s(age) + inc + k5f + k618f + wc + hc,
+     family=binomial, data=Mroz)
> anova(mod.3, mod.1, test="Chisq")

Analysis of Deviance Table

Model 1: lfp ~ s(age) + inc + k5f + k618f + wc + hc
Model 2: lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc
  Resid. Df Resid. Dev    Df Deviance P(>|Chi|)
1       740        920
2       740        917 0.783     2.55      0.08
```

تست از نظر آماری بسیار مهم نیست.[9]

بطور مشابه ما می توانیم اهمیت آماری یک شرایط را در مدل با حذف آن و با توجه به تغییر در انحراف تست کنیم. به عنوان مثال، برای آزمون شرایط سن:

```
> mod.4 <- update(mod.1, . ~ . - s(age))
> anova(mod.4, mod.1, test="Chisq")

Analysis of Deviance Table

Model 1: lfp ~ s(inc) + k5f + k618f + wc + hc
Model 2: lfp ~ s(age) + s(inc) + k5f + k618f + wc + hc
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1       741        945
2       740        917 1.48     27.5   4.2e-07
```

بنابراین، اثر سن از نظر آماری بسیار مهم است.[10] مقایسه کردن این با نتیجه مشابه است برای آزمون والد برای

[9] ما خواننده را برای انجام آزمایشات غیرخطی برای عوامل k5f و k618f دعوت می کنیم.

سن، در بالا ارائه شده است. ما آن را برای خوانده گذاشتیم تا آزمایشات مشابه ای برای پیش بینی های دیگر در مدل، شامل inc و شرایط پارامتریک انجام شود.

4. منابع و خواندن مکمل

رگرسیون غیر پارامتری در Fox ، (2008، فصل 18) بیان شده است.

تمام مدل های رگرسیون غیر پارامتری در این ضمیمه بحث شد(و برخی دیگر، مانند رگرسیون پیگیری-پروژه، رگرسیون، و طبقه بندی و درخت رگرسیون در فاکس (b2000، a) توصیف می شوند از آن نمونه هایی که در آپاندیس ظاهر می شوند، سازگار هستند.

کتاب های عالی و دامنه زیاد توسط هستی و تیبشیرانی (1990) و وود (2006) به ترتیب با بسته های gam و mgcv مربوط می شوند، دومی بخشی از توزیع استاندارد R است. یک درمان بی نقص از GAM و عملکرد gamدر بسته gam در مقاله توسط هاستی (1992) به نظر می رسد.

---

[10] برای مدل هایی اجرا می شود که در آن درجه صاف کردن توسط GCV یا UBRE انتخاب می شود، به جای آزمون های مذکور ثابت تمایل به افزایش معنادار آماری اصطلاحات در مدل دارند.

References

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach With S-Plus Illustrations*. Oxford University Press, Oxford.

Fox, J. (2000a). *Multiple and Generalized Nonparametric Regression*. Thousand Oaks, CA.

Fox, J. (2000b). *Nonparametric Simple Regression: Smoothing Scatterplots*. Thousand Oaks, CA.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Sage, Thousand Oaks, CA, second edition.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks, CA, second edition.

Gu, C. (2000). Multidimensional smoothing with smoothing splines. In Schmiek, M. G., editor, *Smoothing and Regression: Approaches, Computation, and Applications*. Wiley, New York.

Hastie, T. J. (1992). Generalized additive models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 421–454. Wadsworth, Pacific Grove, CA.

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

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. New York.

Nason, G. P. and Silverman, B. W. (1994). The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, 3:163–191.

Nason, G. P. and Silverman, B. W. (2000). Wavelets for regression and other statistical problems. In Schmiek, M. G., editor, *Smoothing and Regression: Approaches, Computation, and Applications*. Wiley, New York.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Socieity, Series B*, 62:413–428.

Wood, S. N. (2001). mgcv: GAMS and generalized ridge regression for R. *R News*, 1(2):20–25.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistics Association*, 99:673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall, Boca Raton, FL.