مجله علوم آماری، بهار و تابستان ۱۳۹۵
جلد ۱۰، شماره ۱، ص ۹۵-۱۱۲
DOI: 10.7508/jss.2016.01.006

برآورد تفاضلی استوار مدل‌های خطی جزئی

مهدی روزبه، جلال چاچی
گروه آمار، دانشگاه سمنان

تاریخ دریافت: ۱۳۹۳/۵/۴ تاریخ آخرین بازنگری: ۱۳۹۵/۳/۳۰

چکیده : رگرسیون خطی استوار یکی از متداول‌ترین رویکردها در روش‌های آماری استوار است. پارامترهای این روش اغلب از طریق کمترین توان‌های دوم پیراسته برآورد می‌شوند که در آن تابع هدف به‌گونه‌ای صورت‌بندی می‌شود که مجموع kتا از کوچکترین توان دوم باقیمانده‌ها (خطاها) کمینه شود. لذا این روش در مقایسه با روش متداول کمترین توان دوم خطا از محاسبات پیچیده‌تری برخوردار است. هدف اصلی این مقاله ارائه یک روش جدیدِ برآورد مدل‌های خطی جزئی با رویکرد تشخیص داده‌های پرت و معرفی برآوردگرهای استوار بر مبنای کمترین توان‌های دوم پیراسته است. در این راستا ابتدا روش تفاضلی در برآورد پارامترهای مدل خطی جزئی بیان می‌شود. سپس روش به‌دست آوردن برآوردگرهای تفاضلی استوار در مدل‌های خطی جزئی بر اساس یک مسئله بهینه‌سازی مبتنی بر کمینه‌سازی مجموع kتا از کوچکترین توان دوم باقیمانده‌ها معرفی می‌شود. این رویکرد توانایی تشخیص داده‌های پرت را دارد. نتایج عددی مطالعه شبیه‌سازی و مطالعه کاربردی با داده‌های واقعی نشان‌دهنده دقت بسیار زیاد برآوردگرهای تفاضلی استوار معرفی شده در این مقاله در مقایسه با برآوردگرهای کلاسیک و متداول مدل‌های خطی

آدرس الکترونیک مسئول مقاله: جلال چاچی، jchachi@profs.semnan.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۲۰، ۶۲G۳۵، ۶۲G۰۸

جزئی هستند.

واژه‌های کلیدی : برآوردگر تفاضلی استوار، کمترین توان‌های دوم پیراسته، مدل خطی جزئی استوار، داده‌های پرت.

۱ مقدمه

مدل‌های خطی جزئی انعطاف‌پذیرتر از مدل‌های خطی متداول هستند، که از دو قسمت خطی (بخش پارامتری مدل) و غیر خطی (بخش ناپارامتری مدل) تشکیل شده‌اند. در واقع استفاده از این مدل زمانی مفید است که متغیر پاسخ y به طور خطی با متغیر توضیحی x و به طور غیر خطی با متغیر توضیحی t (از طریق تابع $f(\cdot)$) رابطه داشته باشد. این‌گونه مدل‌ها نخستین بار توسط انگل و همکاران (۱۹۸۶) معرفی شدند و تا کنون و در طول دهه‌های اخیر بسیار مورد توجه محققان قرار گرفته‌اند و کاربردهای فراوانی در زمینه‌های علمی مختلف پیدا کرده‌اند، از قبیل اقتصاد (ویلیس، ۱۹۸۶؛ بلانچ‌فلاور و اسوالد، ۱۹۹۴)، مصرف سوخت خانوارها (اشمالنسی و استوکر، ۱۹۹۹). همچنین در مطالعات اخیر می‌توان به کاربردهای زیر اشاره نمود. انگل و همکاران (۱۹۸۶) برای بررسی رابطهٔ بین مصرف برق ماهیانه خانوارها (متغیر وابسته y) با متغیرهای مستقل درآمد ماهیانه (X_1)، قیمت برق ماهیانه (X_2) و دمای هوا (t) از مدل خطی جزئی استفاده کردند. آنها حدس زدند که متغیر وابسته y رابطه خطی با متغیرهای X_1 و X_2 و رابطه غیر خطی با متغیر t دارد. یو و چن (۲۰۰۷) مدل‌های خطی جزئی را برای بررسی رابطهٔ بین مصرف مشروبات الکلی (y) با متغیرهای درآمد (X_1)، قیمت مشروبات الکلی (X_2) و سال (t) به‌کار بردند. یو و همکاران (۲۰۰۷) از این مدل‌ها برای بررسی رابطهٔ بین دستمزد ساعتی افراد شاغل (y) با متغیرهای مدت زمان اشتغال فرد در یک سال (X_1)، توانایی شغلی افراد (X_2)، آموزش‌های اضافی (X_3)، شهری و روستایی بودن افراد (X_4) و تعداد سال‌های تحصیل (t) استفاده نمودند. آکدنیز و همکاران (۲۰۱۲) مدل‌های خطی جزئی را برای به‌دست آوردن رابطهٔ بین مصرف ماهیانه برق (y) با متوسط درآمد افراد خانوار (X_1)، نسبت نرخ برق به نرخ گاز (X_2) و متوسط دمای هوا (t) به‌کار بردند.

کاربردهای گسترده اینگونه مدل‌ها در شرایط مختلف باعث ارائه روش‌ها و رویکردهای متنوع و متعددی در برآورد مدل‌های خطی جزئی شده است، از جمله: روش تفاضلی (یاتچو، ۱۹۹۷)، روش کمترین مربعات جریمه‌ای (انگل و همکاران، ۱۹۸۶)، روش مانده‌های جزیی (کوزیک، ۱۹۹۲)، روش درست‌نمایی نیم‌رخ (کارول و همکاران، ۱۹۹۷) و همچنین

روش‌هایی که در دو مرحله به برآورد قسمت خطی مدل (بخش پارامتری) و قسمت غیرخطی مدل (بخش ناپارامتری) می‌پرداختند (فن و همکاران، ۱۹۹۸) (برای مطالعه بیشتر در این زمینه به روزبه (۱۳۹۰) و روزبه همکاران (۲۰۱۱) مراجعه شود).

اما باید در نظر داشت که اعتبار روش‌های برآوردیابی در آمار کلاسیک (به‌خصوص برآوردگرهای مدل‌های خطی از جمله رگرسیون کمترین توان دوم خطا) مبتنی بر فرضیه‌های زیربنایی هستند. یک برآوردگر یا یک روش برآوردیابی آماری استوار نامیده می‌شود هرگاه با به کار بردن آن به توان نتایج قابل اطمینانی حداقل در شرایط زیر گرفت:

۱) برخی فرضیه‌های زیربنایی روش برقرار نباشد؛

۲) در داده‌ها، مشاهدات پرت وجود داشته باشد.

تعیین داده‌های پرت و معرفی روش‌های آماری استوار در هر تجزیه و تحلیل آماری به خصوص برازش مدل‌های رگرسیون آماری بسیار مهم و قابل توجه است (حاجی باقری و همکاران، ۱۳۹۳). در رویکردهای مبتنی بر کمترین توان دوم خطا یک داده پرت تاثیر مخرب (بسیار) زیادی در شیوهٔ برآورد پارامترها و در مدل برازش شده به داده‌ها دارد. داده‌های پرت نباید صرفاً فقط به خاطر اینکه خارج از محدوده دیگر داده‌ها قرار می‌گیرند از مطالعه حذف شوند. بلکه بسیار مهم است که تاثیر این داده‌ها در برازش مدل شناخته شود و رویکردهایی در این زمینه معرفی شوند که با وجود داده‌های پرت در مشاهدات نمونه به توانند استنباط قابل قبولی ارائه دهند. در این راستا می‌توان به برازش مدل‌های رگرسیون استوار مانند برآورد پارامترهای مدل بر اساس کمینه‌سازی مجموع توان دوم خطاهای پیراسته که اثر مخرب داده‌های پرت در برازش مدل را تا حدود (بسیار) زیادی کاهش می‌دهد، اشاره نمود. لذا در این راستا و در مقالهٔ حاضر رویکردی پیشنهاد می‌شود که با تعیین و شناسایی مشاهدات پرت به برآورد استوار پارامترهای مدل رگرسیون خطی جزئی بر مبنای روش تفاضلی و بر اساس یک مسئله بهینه‌سازی مبتنی بر کمینه‌سازی مجموع kتا از کوچکترین توان دوم باقیمانده‌ها می‌پردازد. نتایج عددی مطالعه‌های شبیه‌سازی و کاربردی با داده‌های واقعی نشان‌دهنده دقت بسیار زیاد برآوردگرهای استوار معرفی شده در این مقاله در مقایسه با برآوردگرهای کلاسیک و متداول موجود هستند.

در بخش ۲ روش تفاضلی در برآورد مدل‌های خطی جزئی تشریح می‌شود و نحوه برآورد استوار پارامترهای مدل رگرسیون خطی جزئی بر مبنای روش تفاضلی در بخش ۳ بیان می‌شود. در بخش‌های ۴ و ۵ نتایج مثال‌های عددی (مطالعه شبیه‌سازی و تحلیل با داده‌های واقعی) آورده می‌شود و در انتها به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ روش تفاضلی برآورد مدل‌های رگرسیون خطی جزئی

مدل رگرسیون خطی جزئی[۱] به‌صورت

$$y = X\beta + f(t) + \epsilon, \qquad (۱)$$

تعریف می‌شود، که در آن $y = (y_۱, \ldots, y_n)'$ مشاهدات متغیر پاسخ، $X = (x_۱, \ldots, x_n)'$ ماتریس $n \times p$ از مشاهدات متغیر توضیحی، $f(\cdot)$ یک تابع حقیقی-مقدار (نامعلوم) از مقادیر $t_۱ \leq t_۲ \leq \ldots \leq t_n$، $\beta = (\beta_۱, \ldots, \beta_p)'$ بردار ضرایب مدل، و ϵ بردار nتایی خطاها از توزیعی با شرایط $E(\epsilon) = ۰$ و $E(\epsilon\epsilon') = \sigma^۲ V$ (V ماتریس واریانس-کواریانس است) می‌باشند. در مدل فوق متغیر وابسته y به طور خطی با متغیر توضیحی X رابطه دارد و به طور غیرخطی و از طریق تابع $f(\cdot)$ با متغیر t در ارتباط است (هاردل و همکاران، ۲۰۰۰).

در روش تفاضلی مرتبهٔ m برای برآورد پارامترهای مدل، ابتدا ماتریس تفاضلی $D_{(n-m)\times n}$ به‌صورت

$$D = \begin{bmatrix} d_۰ & d_۱ & \ldots & d_m & ۰ & ۰ & \ldots & ۰ \\ ۰ & d_۰ & d_۱ & \ldots & d_m & ۰ & \ldots & ۰ \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ ۰ & ۰ & \ldots & ۰ & d_۰ & d_۱ & \ldots & d_m \end{bmatrix},$$

تشکیل می‌شود، که در آن d_jها وزن‌های تفاضلی مدل با شرایط $\sum_{j=۰}^{m} d_j = ۰$ و $\sum_{j=۰}^{m} d_j^۲ = ۱$ هستند. به عنوان مثال وزن‌ها را می‌توان به‌صورت

$$d_j = \begin{cases} \sqrt{\frac{m}{m+۱}} & j = ۰, \\ -\sqrt{\frac{۱}{m(m+۱)}} & j = ۱, \ldots, m, \end{cases}$$

درنظر گرفت. هال و همکاران (۱۹۹۰) مقادیر بهینهٔ وزن‌های تفاضلی مرتبهٔ m را به‌صورت عددی تا مرتبهٔ $m = ۱۰$ مطابق جدول ۱ به‌دست آوردند.

با ضرب ماتریس تفاضلی D در طرفین معادله (۱) اثر تابع $f(\cdot)$ به‌صورت

$$\begin{aligned} Dy &= DX\beta + Df(t) + D\epsilon \\ &\cong DX\beta + D\epsilon, \qquad (Df(t) \cong ۰), \end{aligned}$$

حذف می‌شود. اکنون با تغییر متغیرهای $\tilde{y} = Dy$، $\tilde{X} = DX$ و $\tilde{\epsilon} = D\epsilon$، مدل (۱) به مدل خطی

$$\tilde{y}_{(n-m)\times ۱} \cong \tilde{X}_{(n-m)\times p}\beta_{p\times ۱} + \tilde{\epsilon}_{(n-m)\times ۱},$$

---

۱ Partial linear model

جدول ۱: ضرایب تفاضلی بهینه تا مرتبهٔ $m = ۱۰$

m	۱	۲	۳	۴	۵
$d_۰$	۰/۷۰۷۱	۰/۸۰۹۰	۰/۸۵۸۲	۰/۸۸۷۳	۰/۹۰۶۴
$d_۱$	−۰/۷۰۷۱	−۰/۵۰۰۰	−۰/۳۸۳۲	−۰/۳۰۹۹	−۰/۲۶۰۰
$d_۲$		−۰/۳۰۹۰	−۰/۲۸۰۹	−۰/۲۴۶۴	−۰/۲۱۹۷
$d_۳$			−۰/۱۹۴۲	−۰/۱۹۰۱	−۰/۱۷۷۴
$d_۴$				−۰/۱۴۰۹	−۰/۱۴۲۰
$d_۵$					−۰/۱۱۰۳

m	۶	۷	۸	۹	۱۰
$d_۰$	۰/۹۲۰۰	۰/۹۳۰۲	۰/۹۳۸۰	۰/۹۴۴۳	۰/۹۴۹۴
$d_۱$	−۰/۲۲۳۸	−۰/۱۹۶۵	−۰/۱۷۵۱	−۰/۱۵۷۸	−۰/۱۴۳۷
$d_۲$	−۰/۱۹۲۵	−۰/۱۷۲۸	−۰/۱۵۶۵	−۰/۱۴۲۹	−۰/۱۳۱۴
$d_۳$	−۰/۱۶۳۵	−۰/۱۵۰۶	−۰/۱۳۸۹	−۰/۱۲۸۷	−۰/۱۱۹۷
$d_۴$	−۰/۱۳۶۹	−۰/۱۲۹۹	−۰/۱۲۲۴	−۰/۱۱۵۲	−۰/۱۰۸۵
$d_۵$	−۰/۱۱۲۶	−۰/۱۱۰۷	−۰/۱۰۶۹	−۰/۱۰۲۵	−۰/۰۹۷۸
$d_۶$	−۰/۰۹۰۶	−۰/۰۹۳۰	−۰/۰۹۲۵	−۰/۰۹۰۵	−۰/۰۸۷۷
$d_۷$		−۰/۰۷۶۸	−۰/۰۷۹۱	−۰/۰۷۹۲	−۰/۰۷۸۲
$d_۸$			−۰/۰۶۶۶	−۰/۰۶۸۷	−۰/۰۶۹۱
$d_۹$				−۰/۰۵۸۸	−۰/۰۶۰۶
$d_{۱۰}$					−۰/۰۵۲۷

تبدیل می‌شود، که در آن $\tilde{\epsilon}_{(n-m)\times ۱} \in \mathbb{R}^{n-m}$ بردار خطاها از توزیعی با شرایط $E(\tilde{\epsilon}) = ۰$ و ماتریس واریانس-کواریانس $E(\tilde{\epsilon}\tilde{\epsilon}') = \sigma^۲ V_D$ است، به‌طوری که $V_D = DVD' \neq I_{n-m}$ و I_{n-m} ماتریس همانی از مرتبه $n-m$ است. مدل جدید یک مدل رگرسیون خطی متداول است، لذا پارامتر β را می‌توان با روش کمترین توان‌های دوم خطا برآورد کرد (بلسلی و همکاران، ۱۹۸۰). در روش کمترین توان‌های دوم تعمیم یافته، β با کمینه کردن مجموع توان‌های دوم وزنی خطاها به‌صورت

$$\begin{aligned} \widehat{\beta} &= \arg\min_{\beta} (\tilde{y} - \tilde{X}\beta)' {V_D}^{-۱} (\tilde{y} - \tilde{X}\beta) \qquad (۲) \\ &= (\tilde{X}' {V_D}^{-۱} \tilde{X})^{-۱} \tilde{X}' {V_D}^{-۱} \tilde{y}, \end{aligned}$$

برآورد می‌شود. $\widehat{\beta}$ برآوردگر کلاسیک مدل رگرسیون خطی جزئی است. با جایگذاری برآورد β در مدل اصلی، تابع f با روش هموارسازی رگرسیون خطی موضعی در مدل

$$f(t) = y - X\widehat{\beta} - \epsilon,$$

تخمین زده می‌شود (تاکزاوا، ۲۰۰۶؛ واسرمن، ۲۰۰۵).

۳ برآورد استوار مدل‌های رگرسیون خطی جزئی

در این بخش رویکردی پیشنهاد می‌شود که پس از تعیین مشاهدات پرت به برآورد استوار پارامترهای مدل رگرسیون خطی جزیی می‌پردازد. پارامترهای این روش از طریق کمینه کردن مجموع kتا از کوچکترین توان‌های دوم باقیمانده‌ها (خطاها) برآورد می‌شوند. در این مقاله فرض می‌شود که نسبت مشاهداتی که پرت نیستند، $\frac{k}{n}$، معلوم است (روسو و لروی، ۱۹۸۷؛ گوین و والش، ۲۰۱۰).

اکنون به منظور تبدیل مسئله بهینه‌سازی کمترین توان‌های دوم خطاها (۲) به مسئله بهینه‌سازی کمترین توان‌های دوم خطاهای پیراسته، شاخص z_i به‌صورت زیر تعریف می‌شود. اگر برای مشاهده iام $z_i = ۰$ باشد، این مشاهده پرت است و اگر $z_i = ۱$ باشد، این مشاهده پرت نیست. اکنون با اعمال این شاخص در مسئله بهینه‌سازی کمترین توان‌های دوم خطاها (۲)، مسئله بهینه‌سازی کمترین توان‌های دوم پیراسته به‌صورت

$$\min_{\beta,Z}\{(\tilde{y}-\tilde{X}\beta)'V_D^{-\frac{۱}{۲}}ZV_D^{-\frac{۱}{۲}}(\tilde{y}-\tilde{X}\beta)\};\ \ \sum_{i=۱}^{n} z_i = k,\ z_i \in \{۰,۱\}$$

نوشته می‌شود، که در آن Z ماتریس قطری با مقادیر $z = [z_۱,\ldots,z_n]'$ در قطر اصلی است. مسئله بهینه‌سازی بالا با در نظر گرفتن $Z = I_n$ و $k = n$، به مسئله بهینه‌سازی کمترین توان‌های دوم خطای تعمیم‌یافته (۲) تبدیل می‌شود. در ادامه ساختار مسئله بهینه‌سازی فوق بیشتر مورد بررسی قرار می‌گیرد. برای این منظور فرض کنید

$$g(z,\beta) = (\tilde{y}-\tilde{X}\beta)'D(z)(\tilde{y}-\tilde{X}\beta),$$

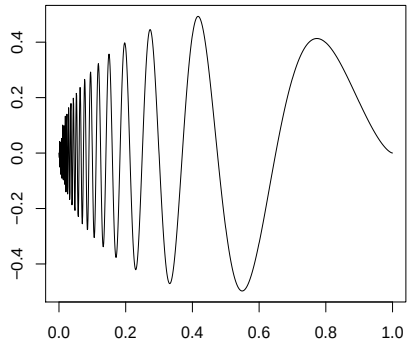
که در آن $D(z) = V_D^{-\frac{۱}{۲}}ZV_D^{-\frac{۱}{۲}}$. مسئله کمینه‌سازی $\min_{\beta,z} g(z,\beta)$ را می‌توان به‌صورت

$$\min_{\beta,z} g(z,\beta) = \min_{z}(\min_{\beta} g(z,\beta)),$$

در نظر گرفت. اکنون می‌توان مسئله بهینه‌سازی $\min_{\beta} g(z,\beta)$ را برای هر z ثابتی حل کرد و سپس جواب بهینه به‌دست آمده را روی کلیه مقادیر z کمینه کرد. با فرض ثابت بودن z بهترین جواب مسئله بهینه‌سازی $\min_{\beta} g(z,\beta)$ عبارت است از

$$\widehat{\beta}(z) = (\tilde{X}D(z)\tilde{X})^{-۱}\tilde{X}'D(z)\tilde{y}.$$

با در نظر گرفتن $\widehat{\beta}(z)$، مجموع مانده‌های توان دوم پیراسته (یا به عبارتی $g(z,\widehat{\beta}(z))$) فقط تابعی از z است که به‌صورت



شکل ۱: نمودار موجی شکل و هموار تابع داپلر

$$\begin{aligned} h(z) &= (\tilde{y} - \tilde{X}\widehat{\beta}(z))' D(z) (\tilde{y} - \tilde{X}\widehat{\beta}(z)) \\ &= \tilde{y}' D(z) \tilde{y} - \tilde{y}' D(z) \tilde{X} (\tilde{X}' D(z) \tilde{X})^{-1} \tilde{X}' D(z) \tilde{y} \end{aligned}$$

نوشته می‌شود. در نتیجه مسئله بهینه‌سازی در برآورد بردار z عبارت است از

$$\min_{z} h(z); \quad \sum_{i=1}^{n} z_i = k, \ z_i \in \{0, 1\}.$$

اکنون با حل مسئله بهینه‌سازی بالا و پس از به‌دست آوردن مقدار $\widehat{z}$ برآوردگر استوار پارامتر β در مدل خطی جزئی (۱) به‌صورت

$$\begin{aligned} \widehat{\beta}_R &= \widehat{\beta}(\widehat{z}) \\ &= (\tilde{X} D(\widehat{z}) \tilde{X})^{-1} \tilde{X}' D(\widehat{z}) \tilde{y} \end{aligned}$$

به‌دست می‌آید. با جایگذاری برآورد β در مدل اصلی، تابع f با روش هموارسازی رگرسیون خطی موضعی در مدل

$$f(t) = y - X\widehat{\beta}_R - \epsilon,$$

تخمین زده می‌شود.

۴ مطالعه شبیه‌سازی

در این بخش، در یک مطالعه شبیه‌سازی به بررسی و مقایسه برآوردگر استوار مدل‌های خطی جزئی و برآوردگر کلاسیکِ متداول مدل‌های خطی جزئی پرداخته می‌شود. در این مطالعه متغیر پاسخ از مدل

$$y_i = \beta_\circ + \beta_1 x_i + f(t_i) + \epsilon_i, \quad i = 1, \ldots, 200,$$

به حجم $n = 200$ شبیه‌سازی می‌شود، که در آن

$$\begin{aligned}
\beta &= [\beta_\circ, \beta_1]' = [1/5, -3]', \\
X_i &\sim N(1, 1), \\
f(t_i) &= \sqrt{t_i(1-t_i)}\sin(\frac{2/1\pi}{t_i + 0/05}), \quad t_i = \frac{i - 0/5}{200}, \\
\epsilon &= [\epsilon_1, \epsilon_2, \ldots, \epsilon_{150}, \epsilon_{151}, \epsilon_{152}, \ldots, \epsilon_{200}], \\
\epsilon_i &\sim \begin{cases} N(0, 0/01) & i = 1, \ldots, 150, \\ t_{2,8} & i = 151, \ldots, 200, \end{cases}
\end{aligned}$$

که در آن‌ها $f(\cdot)$ تابع معروف داپلر نام دارد و در شکل ۱ نمودار آن رسم شده است (واسرمن، ۲۰۰۵). باید توجه داشت که توابع سینوسی و موجی شکل هموار انتخاب‌های مناسبی برای تابع $f(\cdot)$ به منظور آزمون کارایی برآوردگرهای معرفی شده می‌باشند. اکنون برای اینکه داده‌ها شامل مشاهدات پرت نیز باشند، مقادیر بردار خطا از دو توزیع متفاوت شبیه‌سازی می‌شوند. در اینجا ۱۵۰تا از مقادیر خطاها به طور مستقل از توزیع نرمال با میانگین صفر و واریانس ۰/۰۱ تولید و شبیه‌سازی می‌شوند (مشاهدات خوب) و بقیه آنها، یعنی ۵۰تای باقیمانده به طور مستقل از توزیع t–استیودنت غیرمرکزی با ۲ درجه آزادی و پارامتر غیرمرکزی ۸ تولید و شبیه‌سازی می‌شوند (این تعداد داده نقش مشاهدات پرت را دارند). در این شبیه‌سازی از آنجا که حجم نمونه بزرگ است ($n = 200$) و دیگر اینکه همگرایی برآوردگرها مورد توجه نمی‌باشد، تعداد تکرارها برابر یک در نظر گرفته شده است (واسرمن، ۲۰۰۵).

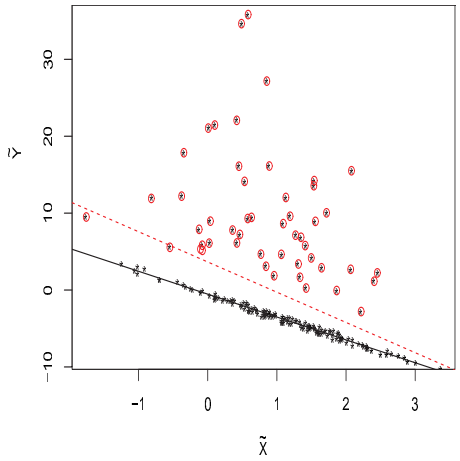
در ادامه در دو مرحله پارامتر β و تابع $f(\cdot)$ برآورد می‌شوند. ابتدا با روش تفاضلی مرتبه $m = 3$ پارامتر β برآورد می‌شود و سپس با جایگذاری $\widehat{\beta}$ در مدل اصلی، تابع f با روش هموارسازی رگرسیون خطی موضعی تخمین زده می‌شود.

مرحله ۱. برآورد پارامتر β

با توجه به ضرایب تفاضلی بهینه مرتبه $m = ۳$ (جدول ۱) ماتریس تفاضلی

$$D = \begin{pmatrix} d_۰ & d_۱ & d_۲ & d_۳ & ۰ & ۰ & \dots & ۰ \\ ۰ & d_۰ & d_۱ & d_۲ & d_۳ & ۰ & \dots & ۰ \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ ۰ & ۰ & \dots & ۰ & d_۰ & d_۱ & d_۲ & d_۳ \end{pmatrix}_{۱۹۷\times ۲۰۰}$$

به‌دست می‌آید، که در آن $d_۰ = ۰/۸۵۸۲$، $d_۱ = -۰/۳۸۳۲$، $d_۲ = -۰/۲۸۰۹$ و $d_۳ = -۰/۱۹۴۲$. اکنون با توجه به روش معرفی شده در بخش‌های ۲ و ۳ پارامترهای مدل خطی جزئی به روش استوار (شناسایی داده‌های پرت و بی‌تاثیر ساختن آنها در برآورد پارامترهای مدل) و روش کلاسیک یا متداول برآورد می‌شوند. در شکل ۲ نمودار پراکنش داده‌ها به همراه نمودار برآوردهای تفاضلی تعمیم‌یافته استوار و کلاسیک داده‌ها رسم شده است. در این شکل نقاط پرت شناسایی شده توسط مدل استوار با دایره نشان داده شده‌اند.



شکل ۲: برآوردهای خطی استوار و غیراستوار

بدیهی و مشهود است که نقاط پرت شناسایی شده تاثیر زیادی در برآوردگر کلاسیک گذاشته و باعث انحراف نمودار به سمت این نقاط گردیده است. در جدول ۲ برآوردگر تفاضلی تعمیم‌یافته استوار ($\widehat{\beta}_R$) و برآوردگر تعمیم یافته تفاضلی کلاسیک یا غیر استوار ($\widehat{\beta}$) به همراه معیارهایی برای مقایسه آنها نشان داده شده است که در ادامه بیان می‌شوند:

۱) برآورد استوار $\widehat{\beta}_R = [۱/۳۸, -۲/۹۶]'$ در مقایسه با برآورد کلاسیک

$\widehat{\beta} = [۵/۸۹, -۳/۸۰]'$ بسیار نزدیکتر به مقدار واقعی پارامتر $\beta = [۱/۵, -۳]'$ است؛
۲) ضریب تعیین $R^۲$ برای مدل برآورد شده به شیوهٔ استوار برابر ۰/۹۰ است که این مقدار بسیار بیشتر از ۰/۱۶ ضریب تعیین مدلی است که پارامترهای آن به شیوه کلاسیک برآورد شده است. لازم به ذکر است که ضریب $R^۲$ در اینجا برابر است با تغییرات کل بیان شده توسط مدل خطی جزئی به تغییرات کل، یعنی

$$R^۲ = \frac{\sum_{i=۱}^{۲۰۰}(\widehat{y}_i - \bar{y})^۲}{\sum_{i=۱}^{۲۰۰}(y_i - \bar{y})^۲},$$

که در آن

$$\widehat{y}_i = \widehat{\beta}_۰ + \widehat{\beta}_۱ x_i + \widehat{f}(t_i), \quad i = ۱, \ldots, ۲۰۰.$$

۳) $s^۲$ خطای کل در برآورد مدل خطی جزئی به شیوهٔ استوار برابر ۰/۰۹ است که این مقدار بسیار کمتر از ۷/۹۵ خطای کل مدلی است که پارامترهای آن به شیوه کلاسیک برآورد شده است. $s^۲$ همچنین برآوردگر مناسبی برای $\sigma^۲$ است. لذا طبق نتایج فوق $s^۲$ برآورد دقیق‌تر و نزدیکتری از مقدار $\sigma^۲ = ۰/۰۱$ در مدل برآورد شده به شیوه استوار ارائه می‌دهد؛
۴) خطاهای استاندارد (SE) برآوردگر $\widehat{\beta}_R$ کمتر از خطاهای استاندارد برآورد کلاسیک $\widehat{\beta}$ است.

محاسبات این قسمت در نرم‌افزار R صورت گرفته است (فوکس و ویزبرگ، ۲۰۱۱).

جدول ۲: مقایسه برآوردگرهای استوار و غیر استوار در مطالعه شبیه‌سازی

غیر استوار		استوار		
$\widehat{\beta}$	SE	$\widehat{\beta}_R$	SE_R	
۵/۸۹	۰/۸۶	۱/۳۸	۰/۱۱	$\widehat{\beta}_۰$
-۳/۸۰	۰/۶۲	-۲/۹۶	۰/۰۸	$\widehat{\beta}_۱$
۰/۱۶		۰/۹۰		$R^۲$
۷/۹۵		۰/۰۹		$s^۲$

مرحله ۲. برآورد تابع $f(\cdot)$

در این مرحله پس از جایگذاری برآورد پارامتر β (به عنوان مثال برآورد استوار $\widehat{\beta}_R = [۱/۳۸, -۲/۹۶]'$) در مدل اصلی، یعنی

$$f(t_i) = y_i - (۱/۳۸ - ۲/۹۶ x_i) - \epsilon_i, \qquad i = ۱, \ldots, ۲۰۰,$$

جدول ۳: مقایسه کارایی نسبی برآوردگرهای استوار و غیر استوار در مطالعه شبیه‌سازی با حجم‌های نمونه‌ای مختلف و درصدهای متغیری از داده‌های پرت

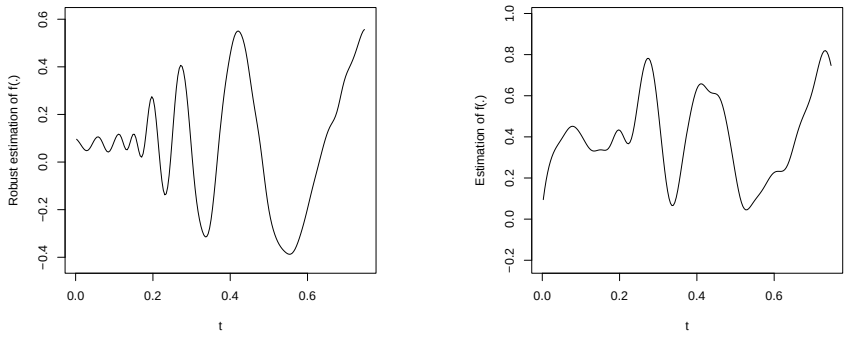
کارایی نسبی $\widehat{\beta}_R$ و $\widehat{\beta}$	حجم نمونه	در صد داده‌های پرت
۳/۹۲	۵۰	۲۵
۲/۲۰	۱۰۰	۲۵
۱/۹۸	۱۵۰	۲۵
۱/۶۹	۲۰۰	۲۵
۷/۴۰	۵۰	۳۳
۵/۰۲	۱۰۰	۳۳
۴/۹۵	۱۵۰	۳۳
۲/۲۵	۲۰۰	۳۳
۱۲/۶۰	۵۰	۵۰
۸/۴۶	۱۰۰	۵۰
۵/۵۳	۱۵۰	۵۰
۵/۶۴	۲۰۰	۵۰

به برآورد تابع $f(\cdot)$ با روش هموارسازی رگرسیون خطی موضعی پرداخته می‌شود. به طور مشابه با جایگذاری برآورد غیر استوار کلاسیک پارامتر β (یعنی $\widehat{\beta} = [\mathbf{۵/۸۹}, \mathbf{-۳/۸۰}]'$) در مدل اصلی تابع $f(\cdot)$ برآورد می‌شود. در شکل‌های ۱ و ۳ تابع f و برآوردهای آن با استفاده از روش‌های استوار $\widehat{\beta}_R$ و غیر استوار کلاسیک $\widehat{\beta}$ رسم شده‌اند. با توجه به مقایسهٔ تابعی این برآوردها با نمودار اصلی تابع $f(\cdot)$ بدیهی است که برآورد تابع $f(\cdot)$ با استفاده از برآورد استوار قسمت خطی مدل رگرسیون خطی جزئی (یعنی $\widehat{\beta}_R$) نزدیکتر به تابع اصلی است.

در انتها با محاسبه خطای کل مدل در برآورد قسمت‌های خطی و غیر خطی مدل، یعنی $SSE = \sum_{i=۱}^{۲۰۰}(y_i - \widehat{y}_i)^۲$، برای دو مدل برآورد شده به شیوه استوار و غیر استوار به مقایسه آنها پرداخته می‌شود. این مقدار برای مدلی که پارامترهای آن به شیوه استوار برآورد شده است برابر $SSE_R = \mathbf{۱۳/۵۵}$ است، که کمتر از $SSE = \mathbf{۱۵۸۹/۴۳}$، خطای کل مدلی است که پارامترهای آن به شیوه معمولی برآورد شده‌اند.

برای مقایسه کارایی نسبی برآوردگرهای استوار ($\widehat{\beta}_R$) و برآوردگرهای کمترین توان‌های دوم خطا ($\widehat{\beta}$) شبیه‌سازی به تعداد $M = ۱۰^۳$ بار تکرار می‌شود و مقادیر $\widehat{\beta}(m)$ و $\widehat{\beta}_R(m)$ برای $m = ۱, \ldots, M$ به‌دست آورده می‌شوند. کارایی نسبی این دو برآوردگر به‌صورت

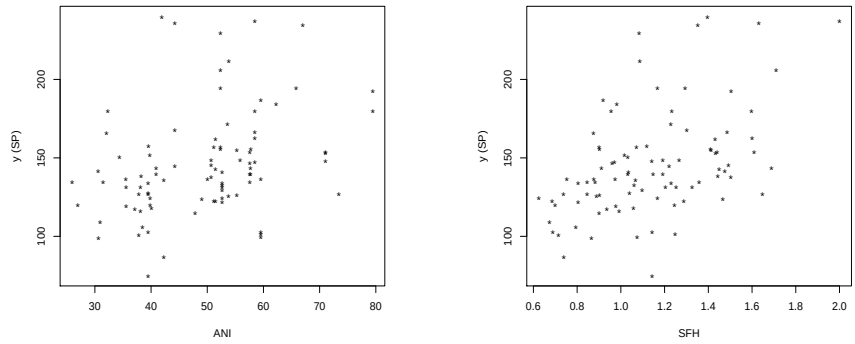
$$eff(\widehat{\beta}_R, \widehat{\beta}) = \frac{\frac{۱}{M}\sum_{m=۱}^{M} ||\widehat{\beta}(m) - \beta||_۲^۲}{\frac{۱}{M}\sum_{m=۱}^{M} ||\widehat{\beta}_R(m) - \beta||_۲^۲},$$



(ب) (الف)

شکل ۳: برآورد تابع داپلر به روش‌های الف: غیر استوار و ب: استوار

محاسبه می‌شود، که در آن $||x||_2^2 = x'x$ نرم اقلیدسی بردار x است. در جدول ۳ مقادیر کارایی نسبی این دو برآوردگر گزارش شده است. از طرفی برای این که نشان داده شود برآوردگرهای استوار معرفی شده در این مقاله برتری قابل ملاحظه‌ای بر برآوردگرهای کمترین توان‌های دوم خطا دارند، در این جدول کارایی نسبی برآوردگرها بر طبق حجم‌های نمونه‌ای مختلف و با درصدهای متغیری از داده‌های پرت محاسبه شده است. در کلیه حالت‌ها نتایج به‌دست آمده نشان‌دهنده این است که برآوردگرهای استوار کارایی بیشتری نسبت به برآوردگرهای کمترین توان‌های دوم خطا دارند. در انتها توجه کنید که روش کمترین توان‌های دوم پیراسته، روشی بسیار استوار در برآورد مدل‌های رگرسیون خطی است. نشان داده‌شده است که کارایی روش کمترین توان‌های دوم پیراسته وقتی تعداد داده‌های پرت از ۰ تا $\frac{n}{4}$ تغییر می‌کند در مقایسه با روش‌های کمترین توان‌های دوم خطا و کمترین میانه خطا بسیار قابل توجه است (روسو، ۱۹۸۴). از طرفی توجه کنید که وقتی در داده‌ها هیچ مشاهده پرتی وجود نداشته باشد این روش معادل روش کمترین توان‌های دوم خطا است. لذا می‌توان گفت که کارایی برآوردگرهای کمترین توان‌های دوم پیراسته حداقل به اندازه کارایی برآوردگرهای کمترین توان‌های دوم خطا است.



(ب) (الف)

شکل ۴: نمودار پراکنش متغیر وابسته قیمت مسکن در مقابل متغیرهای مستقل الف: مساحت مسکن و ب: متوسط درآمد خانوار

۵ تحلیل با داده‌های واقعی

در ادامه به بررسی و مقایسه برآوردگرهای استوار پیشنهاد شده در این مقاله با برآوردگرهای غیر استوار در مدل‌های خطی جزئی برای داده‌های قیمت مسکن در اوتاوا پرداخته می‌شود (هو، ۱۹۹۵). در این داده‌ها قیمت مسکن فروخته شده برحسب هزار دلار متغیر وابسته (SP) است و متغیرهای مستقل عبارتند از SFH مساحت خانه برحسب فوت مربع و ANI متوسط درآمد خانوار برحسب دلار. هو (۱۹۹۵) پس از بررسی متغیرهای موجود دریافت که قیمت مسکن رابطه‌ای خطی با مساحت آن و رابطه غیرخطی با متوسط درآمد خانوار دارد. در شکل ۴ متغیر قیمت مسکن در مقابل متغیرهای مستقل مساحت خانه برحسب فوت مربع (SFH) و متوسط درآمد خانوار برحسب دلار (ANI) رسم شده است. با توجه به رویکرد معرفی شده در بخش ۳، مدل خطی جزئی

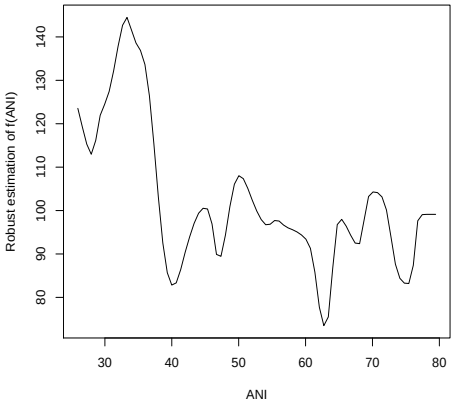
$$SP_i = \beta_0 + \beta_1 SFH_i + f(ANI_i) + \epsilon_i, \qquad i = 1, \ldots, 92,$$

بر اساس $n = 92$ مشاهده موجود درنظر گرفته می‌شود. برآوردگرهای استوار و غیر استوار قسمت خطی مدل در جدول ۴ به‌دست آمده‌اند و در شکل ۵ نشان داده شده‌اند. همچنین برآوردگرهای استوار و غیر استوار قسمت غیر خطی مدل (تابع f) در شکل ۶ نشان داده

شده‌اند.

جدول ۴: مقایسه برآوردگرهای استوار و غیر استوار برای داده‌های واقعی

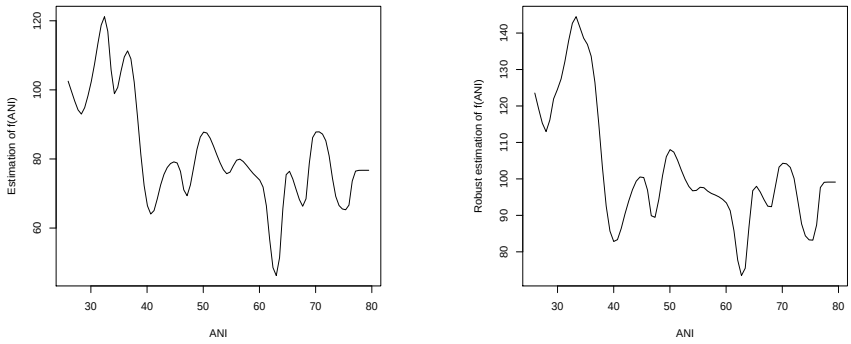
استوار		غیر استوار		
SE_R	$\widehat{\beta}_R$	SE	$\widehat{\beta}$	
۱/۹۲	−۱/۱۷	۲/۸۷	۰/۴۲	$\widehat{\beta}_۰$
۷/۲۷	۳۶/۹۷	۱۰/۳۷	۵۴/۹۵	$\widehat{\beta}_۱$
۰/۸۷۰۴		۰/۵۲۶۴		$R^۲$
۵۸/۸		۵۲۳/۵۵		$s^۲$



شکل ۵: برآوردهای استوار $\widehat{y} = -۱/۱۷ + ۳۵/۹۷x$ و غیراستوار $\widehat{y} = ۰/۴۲ + ۵۴/۹۵x$ برای قسمت خطی مدل

در جدول ۴ برآوردگر استوار ($\widehat{\beta}_R$) و برآوردگر غیر استوار ($\widehat{\beta}$) به همراه معیارهایی برای مقایسه آنها نشان داده شده‌اند. بر اساس معیارهای به‌دست آمده در این جدول برای مدل‌های خطی جزئی استوار و غیر استوار، به‌وضوح استدلال می‌شود (مشابه بحث و استدلال ارائه شده در مقایسه دو مدل در مطالعه شبیه‌سازی شده) که مدل استوار در برازش به داده‌ها بهتر عمل می‌کند. به عبارتی این معیارها را می‌توان به صورت زیر تفسیر نمود:

۱) ضریب تعیین $R^۲$ برای مدل برآورد شده به شیوهٔ استوار برابر ۰/۸۷۰۴ است که این مقدار بسیار بیشتر از ۰/۵۲۶۴ ضریب تعیین مدلی است که پارامترهای آن به شیوه کلاسیک



(الف) (ب)

شکل ۶: نمودار برآوردگر الف: استوار برای قسمت غیرخطی مدل $\widehat{f}_R$ و ب: غیراستوار برای قسمت غیرخطی مدل $\widehat{f}$

برآورد شده است.

۲) s^2 خطای کل در برآورد مدل خطی جزئی به شیوهٔ استوار برابر **۵۸/۸۱** است که این مقدار بسیار کمتر از **۵۲۳/۵۵** خطای کل مدلی است که پارامترهای آن به شیوه کلاسیک برآورد شده است.

۳) خطاهای استاندارد (SE) برآوردگر $\widehat{\beta}_R$ برابر **۲/۸۷** است که کمتر از خطاهای استاندارد برآورد کلاسیک $\widehat{\beta}$، یعنی **۱۰/۳۷**، است.

با توجه به استدلال مطرح شده در بالا بر طبق نتایج عددی به دست آمده در جدول ۴، داده‌های پرت شناسایی شده در شکل ۵ (نقاط دایره‌ای شکل) تاثیر مخربی در برآورد پارامترهای مدل خطی جزئی به روش استوار وارد نمی‌کنند.

بحث و نتیجه‌گیری

در این مقاله روش جدیدی در برآورد مدل‌های خطی جزئی با رویکرد تشخیص داده‌های پرت و معرفی برآوردگرهای استوار ارائه شد. در این روش داده‌های پرت صرفاً فقط به خاطر اینکه خارج از محدوه دیگر داده‌ها قرار می‌گیرند از مطالعه حذف نمی‌شوند. بلکه رویکردی معرفی می‌شود که تاثیر مخرب داده‌های پرت را در برازش مدل بی‌اثر می‌کند. رویکرد استوار

معرفی شده در این مقاله با رویکرد غیر استوار متداول کمترین توان‌های دوم خطا در مدل‌های خطی جزئی مقایسه شد. نتایج عددی مطالعه‌های شبیه‌سازی و کاربردی با داده‌های واقعی نشان می‌دهد که تشخیص داده‌های پرت همچنان که برآوردگرهای دقیق‌تری برای قسمت خطی مدل رگرسیون خطی جزئی ($\mathbf{X}\beta$) ارائه می‌دهد، تاثیر فراوانی در به‌دست آوردن برآوردگری بهتر و با دقت بیشتر برای قسمت غیرخطی (یا تابع f) نیز دارد.
بررسی خصوصیات حدی، سازگاری و کارایی برآوردگرهای استوار مدل‌های خطی جزئی و یا همچنین کاربرد روش‌های بوت‌استرپ در تعیین خواص توزیعی این برآوردگرها (از قبیل میانگین، واریانس و ...) می‌توانند در تحقیقات آتی مورد توجه قرار گیرند (ویزک، ۲۰۰۶a؛ ۲۰۰۶b؛ ۲۰۰۶c).

تقدیر و تشکر

نویسندگان مقاله از داوران محترم که نظرات ارزشمند ایشان باعث بهبود مطالب ارائه شده در این مقاله گردید، کمال تشکر و قدردانی را دارند.

مراجع

حاجی باقری، ف.، راسخ، ع. و آخوند، م. (۱۳۹۳)، تشخیص نقاط پرت در مدل رگرسیونی لیو، مجله علوم آماری، ۸، ۳۶-۱۹.

روزبه، م. (۱۳۹۰)، برآورد در مدل‌های خطی جزیی، رساله دکترای آمار ریاضی، دانشگاه فردوسی مشهد.

Akdeniz Duran, E., Härdle, W. K. and Osipenko, M. (2012), Difference Based Ridge and Liu Type Estimators in Semiparametric Regression Models, *Journal of Multivariate Analysis*, **105**, 164-175.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980), *Regression Diagnostics*, John Wiley, New York.

Blanchfower, D. G. and Oswald, A. J. (1994), *The Wage Curve*, MIT Press Cambridge, MA.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), Generalized Partially Single-Index Models, *Journal of American Statistical Associations*, **92**, 477-489.

Cuzick, J. (1992), Semiparametric Additive Regression, *Journal of Royal Statistical Society*, Series B, **54**, 831-843.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986), Semiparametric Estimates of the Relation Between Weather and Electricity Sales, *Journal of American Statistical Associations*, **81**, 310-320.

Fan, J., Hardle, W. and Mammen, E. (1998), Direct Estimation of Additive and Linear Components for High-Dimensional Data, *Annals of Statistics*, **26**, 943-971.

Fox, J. and Weisberg, S. (2011), *An R Companion to Applied Regression*, 2nd Ed., Sage Publications, Thousand Oaks, CA.

Hall, P., Kay, J. W. and Titterington, D. M. (1990), On Estimation of Noise Variance in Two-Dimensional Signal Processing, *Advanced in Applied Probability*, **23**, 476-495.

Hardle, W., Liang, H. and Gao, J. (2000), *Partially Linear Models*, Physika Verlag, Heidelberg.

Ho, M. (1995), *Essay on the Housing Market*, Unpoblished Ph.D. Dissertation, University of Toronto.

Nguyen, T. D. and Welsch, R. (2010), Outlier Detection and Least Trimmed Squares Approximation Using Semi-definite Programming, *Computational Statistics and Data Analysis*, **54**, 3212-3226.

Roozbeh, M., Arashi, M. and Niroumand, H. A. (2011), Ridge Regression Methodology in Partial Linear Models with Correlated Errors, *Journal of Statistical Computation and Simulation*, **81**, 517-528.

Rousseeuw, P. J. (1984), Least Median of Squares Regression, *Journal of American Statistical Associations*, **79**, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley, New York.

Schmalensee, R. and Stoker, T. M. (1999), Household Gasoline Demand in the United States, *Econometrica*, **67**, 645-662.

Takezawa, K. (2006), *Introduction to Nonparametric Regression*, John Wiley, New Jersey.

Vısek, J. A. (2006a), The Least Trimmed Squares, Part I: Consistency, *Kybernetika*, **42**, 1-36.

Vısek, J. A. (2006b), The Least Trimmed Squares, Part II: $\sqrt{n}$-Consistency, *Kybernetika*, **42**, 181-202.

Vısek, J. A. (2006c), The Least Trimmed Squares, Part III: Asymptotic Normality, *Kybernetika*, **42**, 203-224.

Wasserman, L. (2005), *All of Nonparametric Statistics*, Springer, New York.

Willis, R. J. (1986), Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions, *in: Ashenfelter, O. and Layard, R. The Handbook of Labor Economics*, North Holland-Elsevier Science Publishers, Amsterdam, **1**, 525-602.

Yatchew, A. (1997), An Elementary Estimator of the Partial Linear Model, *Economic Letters*, **57**, 135-143.

You, J. and Chen, G. (2007), Semiparametric Generalized Least Squares Estimation in Partially Linear Regression Models with Correlated Errors, *Journal of Statistical Planning and Inference*, **137**, 117-132.

You, J., Chen, G. and Zhou Y. (2007), Statistical Inference of Partially Linear Regression Models with Heteroscedastic Errors, *Journal of Multivariate Analysis*, **98**, 1539-1557.