سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

تابع بقاء شرطی تعمیم یافته ناپارامتری وتعیین توزیع داده های سانسور فاصله ای

زهره سلیم جهانتیغ[۱]، محمد حسین دهقان [۲]
[۱]دانشگاه سیستان و بلوچستان،دانشکده ریاضی،گروه آمار
[۲]دانشگاه سیستان و بلوچستان،دانشکده ریاضی،گروه آمار

چکیده: در این مقاله پس از معرفی تابع بقاء شرطی ناپارامتری تعمیم یافته به شرط متغیر کمکی پیوسته برای داده های سانسور شده فاصله ای [۱] و بیان برتری آن نسبت به سایر برآوردگرهای تابع بقاء، به بررسی خواص این برآوردگر (GTE) می پردازیم. سپس با استفاده از این برآوردگر و بکارگیری روش مونت کارلو [۲] و شبیه سازی با استفاده از نرم افزار R ، به تولید متغیرهای تصادفی از جامعه ای که داده های سانسور شده فاصله ای موجود است می پردازیم. فواصل اطمینان و آزمون فرض ها را برای جامعه مفروض محاسبه و مورد بررسی و تجزیه و تحلیل قرار می دهیم و نیز نشان خواهیم داد که این روش معیاری برای تشخیص اینکه مدل کاکس بر داده ها مناسب است یا نه، می باشد.

واژه‌های کلیدی: سانسور فاصله ای، تابع بقاء شرطی،نمونه گیری مونت کارلو، تابع کرنل،مدل کاکس.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62N01، 62G15، 62G09، 62G05،60A10.

۱ مقدمه

داده های نادقیق که معمولاً برای مشاهدات طول عمر اتفاق می افتد به داده هایی گفته می شود که زمان دقیق رخ دادن آن ها معلوم نیستند که معمولاً به شکل سانسور از راست یا چپ یا فاصله ای ظاهر می شوند.اصولاً کار کردن با داده های نادقیق خیلی راحت نیستند و در بعضی موارد با روش های معمول برای داده های دقیق و پارامتری غیرممکن خواهند بود. در این مقاله ما از داده های سانسور فاصله ای استفاده می کنیم و نشان خواهیم داد این داده ها می توانند شامل داده های دقیق، سانسور از راست و سانسور از چپ باشند. تابع بقاء ناپارامتری برای داده های دقیق را با $S(t)$ نشان میدهیم و با احتمالی که یک فرد بیشتر از زمان t زنده بماند تعریف میشود، درواقع اگر $F(t) = p(T \leq t)$ تابع توزیع تجمعی باشد آنگاه تابع بقاء بصورت $S(t) = ۱ - F(t) = p(T > t)$ تعریف میشود. از تابع بقاء معمولا در مطالعات پزشکی و مسائل طول عمر استفاده می شود. از آنجاییکه بدست آوردن تابع بقاء برای داده های نادقیق

[۱] GTE
[۲] MonteCarlo

کار آسانی نیست افرادی در این زمینه به موفقیت هایی دست یافته اند مانند کاپلان و مایر [۳] (۱۹۵۸) که برآوردگر جدیدی برای برآورد تابع بقاء برای داده های سانسور از راست ارائه دادند که به برآوردگر کاپلان مایر یا برآوردگر حاصلضرب مشهور است و پس از آن بران [۴] (۱۹۸۱) با استفاده از تابع کرنل (هسته) به شرط متغیر کمکی برآوردگر کاپلان مایر را تعمیم داد ، همچنین ترن بول [۵] (۱۹۷۶) یک برآوردگر تابع بقاء برای داده های سانسور فاصله ای معرفی کرد. در این مقاله ما به بررسی برآوردگر ترن بول تعمیم یافته می پردازیم یعنی تعمیمی از تابع بقاء به شرط متغیر کمکی پیوسته برای داده های سانسور فاصله ای معرفی می کنیم و بیان می کنیم که روش ما بر روش های دیگر برتری خواهد داشت و در نهایت با استفاده از این برآوردگر و بکارگیری روش مونت کارلو و شبیه سازی برای داده های سانسور فاصله ای ، پس از برآورد تابع توزیع ناپارامتری شرطی به تولید داده ها ، برآورد پارامترهای توزیع و فاصله های اطمینان جهت تعیین پوشش $(coverage)$ می پردازیم وارتباط بین این روش و مدل کاکس را بیان می کنیم. مدل کاکس یک مدل پایه و نیمه پارامتری در تحلیل داده های تابع بقاء است که هدف آن بررسی همزمان اثر چند متغیر روی بقاء می باشد.

۲ مفاهیم اولیه

۱.۲ داده های سانسور شده

به داده هایی که میدانیم ممکن است اتفاق بیافتند اما ما آنها را دقیقا مشاهده نکرده ایم داده های نادقیق میگوییم. این داده ها معمولا برای مشاهدات طول عمر اتفاق می افتند و به سه نوع سانسور از راست ، فاصله ای و چپ تقسیم میشوند. در ادامه انواع سانسور را به اختصار شرح خواهیم داد.

۱.۱.۲ سانسور از راست

فرض کنید X $(X \geq ۰)$ یک متغیر تصادفی باشد، X را سانسور از راست گویند هرگاه تا لحظه سانسور مشاهده نشده باشد و پس از آن هر لحظه ممکن است مشاهده گردد. به عنوان یک مثال فرض کنید می خواهیم اثر یک دارو را بر روی تعدادی از بیماران سرطانی به مدت شش ماه بررسی کنیم، متغیر تصادفی در اینجا زمان مرگ بیمار است. ممکن است برخی از بیماران در مدت این شش ماه زنده بمانند و ما مطالعه را پایان دهیم اما می دانیم پس از اتمام مطالعه هر لحظه ممکن است بیمارانی که تا این لحظه زنده هستند، بمیرند، در این صورت سانسور از راست اتفاق می افتد.

۲.۱.۲ سانسور از چپ

فرض کنید X $(X \geq ۰)$ یک متغیر تصادفی باشد، X را سانسور از چپ گویند هرگاه بدانیم قبل از شروع مطالعه یا قبل از زمان مشخصی مشاهده شده باشد. به عنوان یک مثال فرض کنید می خواهیم در یک جامعه از افرادی که سیگار می کشند سن شروع سیگار کشیدن را بررسی کنیم، ممکن است بعضی از این افراد سن دقیق شروع سیگار کشیدن خود را به یاد نداشته باشند و بگویند که قبل از ۱۵ سالگی شروع به کشیدن سیگار کرده ایم ، در این صورت سانسور از چپ اتفاق می افتد.

---

[۳] Caplan–Meier
[۴] Beran
[۵] Turnbull

۳.۱.۲ سانسور فاصله ای

اگر $X_i, i = ۱, ۲, ..., n$ متغیرهای تصادفی مستقل طول عمر باشند و فقط بدانیم که $X_i \in [L_i, R_i]$ باشد آنگاه X_i را سانسور فاصله ای گویند. به عنوان مثال فرض کنید می خواهیم اثر یک دارو را بر بهبود تعدادی از بیماران در یک بازه زمانی بررسی کنیم، ما می دانیم که اگر این دارو مثلا در بازه زمانی $(L_۲, R_۲)$ مشاهده شده است اما زمان دقیق آن را نمی دانیم در این صورت سانسور فاصله ای اتفاق می افتد. نماد داده های سانسور فاصله ای (L, R) شامل تمام حالت های ممکن داده ها می باشد بطوریکه:

-اگر $R_i = \infty$ آنگاه سانسور از راست اتفاق می افتد.

- اگر $L_i = ۰$ سانسور از چپ اتفاق می افتد.

-اگر $L_i = R_i$ آنگاه داده دقیق بدست می آید و این یعنی سانسور اتفاق نمی افتد.

-اگر $L_i \neq R_i$آنگاه سانسور فاصله ای است.

۲.۲ برآوردگر تابع بقاء ترن بول

ترن بول در سال ۱۹۷۶ برای برآورد تابع بقاء وقتی که داده ها سانسور فاصله ای هستند یک برآوردگر ناپارامتری خودسازگار به روش درستنمایی ماکزیمم ارائه کرد. روش ترن بول شامل یک الگوریتم است که در ادامه شرح خواهیم داد.

اگر $S(t) = p(T > t)$ تابع بقاء باشد و $I_i = (L_i, R_i)$ ، i -امین سانسور فاصله ای مشاهده شده باشد و اگر $\{L_i, R_i, i = ۱, ۲, ..., n\}$ باشند،قرار می دهیم $۰ = \tau_۱ \prec \tau_۲ \prec ... \prec \tau_g \prec ۲n$ مقادیر مرتب شده از $B_j = (\tau_{j-۱}, \tau_j)$ و فرض میکنیم $\{A_\ell, \ell = ۱, ۲, ..., m\}$ مجموعه ای از داخلی ترین فواصل باشند یعنی

$$\{A_\ell, \ell = ۱, ۲, ..., m\} = \{B_j :, i' : \tau_{j-۱} = L_i, \tau_j = R_{i'}, j = ۱, ۲, ..., g\}.$$

حال متغیر نشانگر زیر را تعریف میکنیم:

$$\alpha_{i,j} = \begin{cases} ۱, & B_j \subseteq I_i \quad , i = ۱, ۲, ..., n, j = ۱, ۲, ..., g \\ ۰, & Otherwise \end{cases}$$

فرض کنیم p_j جرم احتمال تعیین شده برای B_j در برآورد درستنمایی ماکزیمم ناپارامتری توسط ترن بول برای تابع $S(t) = p(T > t)$ باشد یعنی $p_j = \widehat{P}r\left[T \in B_j\right], j = ۱, ۲, ..., g$. بنابراین p_j با حل معادله خودسازگار زیر بدست می آید:

$$p_j = \sum_{i=۱}^{n} \frac{۱}{n} \frac{\alpha_{i,j} p_j}{\sum_{k=۱}^{g} \alpha_{i,k} p_k}$$

که در آن اگر B_j فاصله میانی نباشد $p_j = ۰$ و در غیر اینصورت $p_j > ۰$ می باشد.

اکنون با توجه به تعاریف و علامت گذاری بالا الگوریتم ترن بول را بصورت زیر بیان میکنیم:

الگوریتم ۱. این الگوریتم دارای گام‌های زیر است:

- گام اول: فرض کنیم $r = ۰$ و مقادیر اولیه $p_j^{(۰)}, j = ۱, ۲, ..., g$ را برای جرم های احتمال قرار می دهیم.
- گام دوم: معادله زیر را حساب می کنیم

$$p_j^{(r+۱)} = \sum_{i=۱}^{n} \frac{۱}{n} \frac{\alpha_{i,j} p_j^{(r)}}{\sum_{k=۱}^{g} \alpha_{i,k} p_k^{(r)}}$$

- گام سوم: اگر، $\|P^{(r)} - P^{(r-۱)}\| / \|P^{(r)}\| > \varepsilon$ ، به گام دوم باز میگردیم در غیر اینصورت الگوریتم را متوقف میکنیم. $P^{(r)} = (p_۱^{(r)}, ..., p_g^{(r)})'$ و ε یک عدد حقیقی کوچک مثبت است.

بنابراین برآوردگر ناپارامتری تابع بقاء ترن بول بصورت زیر معرفی می شود

$$\widehat{S}(t) = \begin{cases} ۱, & t < \tau_۱ \\ \sum_{j:\tau_j > t} \widehat{p}_j, & t \geq \tau_۱ \end{cases}$$

۳ تابع بقاء شرطی تعمیم یافته

برای بدست آوردن این تعمیم ما از برآوردگر تابع بقاء ترن بول برای داده های سانسور فاصله ای استفاده می کنیم، درواقع برآوردگر ارائه شده توسط ترن بول را برای داده های سانسور فاصله ای تعمیم میدهیم.

در این بخش برآوردگری را معرفی می کنیم که در آن متغیر تصادفی بطور فاصله ای سانسور شده و متغیر کمکی پیوسته است. بصورت زیر عمل می کنیم:

با شرطی کردن متغیر تصادفی طول عمر T روی یک یا چند متغیر تصادفی کمکی مانند Z و با استفاده از تابع وزنی کرنل (هسته) روی متغیر تصادفی کمکی به مرکزیت $z_۰$ و قرار دادن کرنل در برآوردگر ترن بول (۱۹۷۶) آن را بصورت یک برآوردگر ناپارامتری تابع بقاء شرطی [۶] برای داده های سانسور شده فاصله ای تعمیم می دهیم و آن را از این پس برآوردگر تعمیم یافته ترن بول می نامیم که بصورت زیر تعریف می گردد:

$$S(t|z_۰) = Pr\left[T > t | Z = z_۰\right]$$

از این پس داده ها بصورت سه تایی مرتب $(L_i, R_i, Z_i), i = ۱, ۲, ..., n$ نشان داده می شوند بطوریکه $T_i \in (L_i, R_i)$ است.

فرض کنید $\omega_i^h(z_۰)$ تابع وزن (کرنل) برای مشاهده i باشد که بصورت زیر تعریف می شود

$$\omega_i^h(z_۰) = \frac{K_h\left\{(Z_i - z_۰)\right\}}{\sum_{q=۱}^{n} K_h\left\{(Z_q - z_۰)\right\}}$$

که در آن $K(.)$ تابع کرنل و h پارامتر طول پنجره (پهنای باند) می باشد. $p_j(z_۰)$ را جرم احتمال برای B_j توسط برآوردگر $S(t|z_۰)$ در نظر می گیریم که با حل رابطه خودسازگار زیر بدست می آید

تعریف ۱.۳.

$$p_j(z_۰) = \sum_{i=۱}^{n} \omega_i^h(z_۰) \frac{\alpha_{i,j} p_j(z_۰)}{\sum_{k=۱}^{g} \alpha_{i,k} p_k(z_۰)}, j = ۱, ..., g \qquad (۱.۳)$$

[۶] NPMLE

در اجرای رابطه ۱.۳ معمولاً از کرنل نرمال استاندارد و کرنل اپانچنیکف [۷] استفاده میکنیم.

با جایگذاری روابط تعریف شده بالا در الگوریتم ترن بول برآورد متناظر $S(t|z_\circ)$ بصورت زیر می باشد

$$\widehat{S}(t|z_\circ) = \begin{cases} ۱, & t < \tau_۱ \\ \sum_{j:\tau_j > t} \widehat{p}_j(z_\circ), & t \geq \tau_۱ \end{cases} \quad (۲.۳)$$

حال خواص این برآوردگر و اینکه این برآورگر می تواند سایر برآوردگرها را پوشش دهد نشان می دهیم.

۱.۳ خواص برآوردگر GTE

استفاده از این روش به ما در شناسایی داده ها و انتخاب مدل کمک می کند، این برآوردگر یک برآوردگر فراگیر بوده و در شرایطی که داده ها از راست سانسور شده باشند ، برآوردگر تعمیم یافته کاپلان–مایر [۸] و در صورتیکه داده ها سانسور نشده (دقیق) باشند ، برآوردگر تعمیم یافته تجربی تابع بقاء شرطی ناپارامتری و چنانچه z_iها مساوی باشند ، برآوردگر ترن بول (۱۹۷۶) را نتیجه می دهد.این برآوردگر بصورت یکنواخت به احتمال واقعی همگرا می باشد. از این برآوردگر برای تشخیص برازش مدل کاکس به داده ها استفاده می شود. با استفاده از این برآوردگر می توانیم تشخیص دهیم که آیا مدل کاکس برای داده های موجود مناسب است یا خیر. به این صورت که نمودارهای $ln(-ln\widehat{S}(t|z_\circ))$ را رسم میکنیم و اگر این نمودارها بصورت خطوط موازی باشند مدل کاکس برای داده ها مناسب خواهد بود.

۴ برآورد پارامترهای توزیع داده های سانسور فاصله ای

در این بخش به بررسی روشی برای برآورد پارامترهای توزیع داده های سانسور فاصله ای می پردازیم. بدین منظور با استفاده از برآوردگر GTE تابع توزیع دلخواه را برآورد و با نمونه گیری مونت کارلو به تولید داده ها از جامعه هدف می پردازیم، برای این کار با استفاده از نرم افزار R از یک توزیع مانند وایبل [۹] با پارامترهای $shape = ۳$ و $scale = ۱/۵z_\circ$ ، n داده شبیه سازی می کنیم و مشاهدات مستقل $(L_i, R_i, Z_i), i = ۱, ۲, ..., n$ را میسازیم و با استفاده از این مشاهدات و تابع gte مقادیر برآورد تابع بقاء شرطی تعمیم یافته را بدست می آوریم و پس از آن با روش مونت کارلو از توزیع برآورد شده متغیر تصادفی را تولید می کنیم و نشان می دهیم که توزیع این نمونه بسیار نزدیک به توزیع انتخابی اولیه ما یعنی وایبل است. در جداول ۲ و ؟؟ برخی از نتایج این برنامه نویسی آورده شده است.

جدول و شکل

با توجه به جدول ۲ می بینیم که برای مقادیر بزرگتر $Rate$ و $z_\circ$ یعنی $Rate = ۲$ و $z_\circ = ۱۵$ مقادیر ما دقیق تر و به توزیع جامعه نزدیکتر است. $Rate$ برابر است با معکوس میانگین فاصله ها، بطوریکه هر چه مقدار آن بیشتر شود فاصله ها کم تر می شوند.

در جدول ؟؟ مقادیر t صدک های مراتب ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ از توزیع $T|Z = ۱۵$ می باشند.

---

[۷] Epanechnikov

[۸] GKM

[۹] weibull

جدول ۱: مقایسه پارامترهای حقیقی با برآورد آنها.

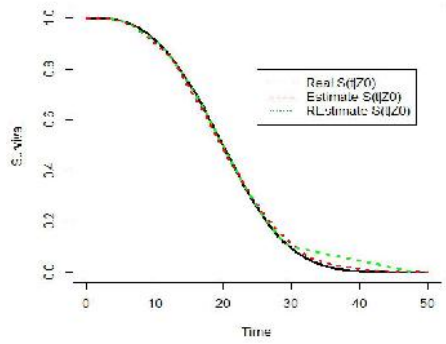
n	$z_\circ$	$Rate$	μ	$\widehat{\mu}$	σ^2	$\widehat{\sigma^2}$	MSE	$Cove\mu$	$Cove\sigma^2$
۵۰	۱۰	۰٫۲۵	۱۳٫۴۰	۱۳٫۶۶	۲۳٫۷۰	۲۹٫۰۲	۰٫۰۰۶۴	۰٫۷۸	۰٫۷۳
		۰٫۵	۱۳٫۳۹	۱۳٫۶۸	۲۳٫۷۳	۳۱٫۱۲	۰٫۰۰۴۶	۰٫۸۶	۰٫۸۲
		۲	۱۳٫۳۸	۱۳٫۶۷	۲۳٫۵۹	۳۱٫۲۸	۰٫۰۰۳۴	۰٫۹۴	۰٫۹۳
	۱۵	۰٫۲۵	۲۰٫۰۸	۱۹٫۵۴	۵۳٫۱۱	۵۷٫۸۳	۰٫۰۰۷۸	۰٫۷۸	۰٫۷۴
		۰٫۵	۲۰٫۱۰	۱۹٫۴۲	۵۳٫۳۴	۶۰٫۲۷	۰٫۰۰۶۳	۰٫۸۶	۰٫۸
		۲	۲۰٫۰۷	۱۹٫۴۴	۵۳٫۰۷	۶۲٫۱۹	۰٫۰۰۵۱	۰٫۹۶	۰٫۹۶
۱۰۰	۱۰	۰٫۲۵	۱۳٫۴۰	۱۳٫۶۷	۲۳٫۷۰	۲۹٫۸۱	۰٫۰۰۶۳	۰٫۸۴	۰٫۷۹
		۰٫۵	۱۳٫۳۹	۱۳٫۸۰	۲۳٫۷۳	۳۱٫۲۲	۰٫۰۰۵۲	۰٫۸۸	۰٫۸۶
		۲	۱۳٫۳۸	۱۳٫۷۱	۲۳٫۵۹	۳۱٫۳۵	۰٫۰۰۳۷	۰٫۹۲	۰٫۹۶
	۱۵	۰٫۲۵	۲۰٫۰۸	۲۰٫۳۳	۵۳٫۱۱	۵۸٫۹۶	۰٫۰۰۸۰	۰٫۸۸	۰٫۸۱
		۰٫۵	۲۰٫۱۰	۲۰٫۳۱	۵۳٫۳۴	۶۰٫۵۴	۰٫۰۰۶۲	۰٫۹۲	۰٫۹
		۲	۲۰٫۰۷	۲۰٫۳۶	۵۳٫۰۷	۶۱٫۳۲	۰٫۰۰۵۴	۰٫۹۸	۰٫۹۸

بحث و نتیجه‌گیری

برآوردگر ناپارامتری تابع بقاء شرطی تعمیم یافته برای داده های سانسور شده برآورد نسبتا دقیقی از تابع بقاء را نتیجه میدهد. واضح است که دقت آن به طول فاصله های سانسور شده بستگی دارد یعنی هرچه طول فاصله ها کمتر باشد دقت برآوردگر افزایش می یابد. از این برآوردگر می توان برای شناسایی خواص اولیه داده ها و اینکه داده ها از مدل کاکس پیروی می کند ، استفاده کرد. ما برای بدست آوردن این نتایج از نرم افزار R استفاده کرده ایم. در اینجا طول پنجره را برابر مقدار بهینه درنظر گرفته ایم یعنی طول پنجره ای که کم ترین خطا (MSE) را دارد.

جدول ۲: تابع بقاء شرطی، میانگین، واریانس و فاصله اطمینان زمان های بقاء برای $n = ۵۰$ و $Z_۰ = ۱۵$ و $Rate = ۰/۲۵$

t	$S(t\|z_۰)$	$\widehat{\mu}$	$\widehat{\sigma}^۲$	CI_μ	$CI_{\sigma^۲}$
۱۰/۸۱۶۸	۰/۸۹۳۲	۰/۸۹۴۸	۰/۰۰۶۹	(۰/۸۸۳۲, ۰/۹۰۶۴)	(۰/۰۰۵۷, ۰/۰۰۸۵)
۱۲/۶۹۶۲	۰/۷۵۵۱	۰/۷۶۴۳	۰/۰۱۰۷	(۰/۷۴۹۹, ۰/۷۷۸۶)	(۰/۰۰۸۸, ۰/۰۱۳۱)
۱۴/۰۱۲۱	۰/۴۶۳۷	۰/۴۷۸۴	۰/۰۲۱۲	(۰/۴۵۸۲, ۰/۴۹۸۶)	(۰/۰۱۷۶, ۰/۰۲۶۱)
۱۷/۱۵۷۴	۰/۲۵۴۷	۰/۲۵۸۷	۰/۰۱۹۹	(۰/۲۳۹۲, ۰/۲۷۸۳)	(۰/۰۱۶۵, ۰/۰۲۴۴)
۲۰/۹۱۹۲	۰/۱۳۳۳	۰/۱۳۷۲	۰/۰۰۸۲	(۰/۱۲۴۶, ۰/۱۴۹۸)	(۰/۰۰۶۸, ۰/۰۱۰۱)



شکل ۱: نمودار تابع بقاء حقیقی در برابر تابع بقاء شرطی تعمیم یافته و تابع بقاء شرطی تعمیم یافته بعد از نمونه گیری برای $n = ۱۰۰$ و $Rate = ۲$ ، $z_۰ = ۱۵$.

مراجع

Betensky RA, Rabinowitz D, TsiatisAA (2001) Computationally simple accelerated failure time regression for interval censored data. Biometrika 88:703–711

Beran, R. (1981). Nonparametic regression with randomly censored survival data. Technical report. Department of Statistics, University of California, Berkeley

Böhning D, Schlattmann P, Dietz E (1996) Interval censored data: a note on the nonparametric maximum likelihood estimator of the distribution function. Biometrika 83:462–466

Efron B (1967) The two sample problem with censored data. In: Proceedings of the fifth Berkeley symposium. University of California, Berkeley

Gentleman R, Geyer CJ (1994) Maximum likelihood for interval censored data: consistency and computation. Biometrika 81:618–623

E. L. Kaplan and Paul Meier. (1958). Nonparametric Estimation from Incomplete Observations,Journal of the American Statistical Association

Pan W (1999) Extending the iterative convex minorant algorithm to the cox model for interval-censored data. J Comput Graph Stat 8:109–120

PanW (2000) Smooth estimation of the survival function for interval censored data. Stat Med 19:2611–2624

Sun J (2001) Variance estimation of a survival function for interval-censored survival data. Stat Med 20:1249–1257

Turnbull BW (1976) The empirical distribution function with arbitrary grouped, censored and truncated data,J R Stat Soc Ser B *TEST*, **38**, 290–295.