اقتراح استعمال مبدأ اعظم دالة انتروبي POME على توزيع كاما العام في تقدير احتمالات البقاء للسكان في العراق

أ.م.د. عمر عبد المحسن علي/كلية الادارة والاقتصاد/جامعة بغداد
الباحث/ رغدة زياد طارق

المستخلص:

تم في هذا البحث تقدير دالة البقاء على قيد الحياة لبيانات تعاني من اضطراب وتشويش للمسح الاجتماعي والاقتصادي للأسرة في العراق (Iraq Household Socio-Economic Survey: 2012 IHSES II 2012) لبيانات فئات خماسية العمر تتبع توزيع كاما العام (Generalized Gamma: GG). واستعملت طريقتين للأغراض التقدير والموائمة fitting وهي طريقة مبدأ اعظم دالة انتروبي Principle of Maximizing Entropy: POME وطريقة تمهيد لامعلمية بدالة لبّية Kernel ، للتغلب على المشاكل الرياضية التي تعتري التكاملات التي يتضمنها هذا التوزيع بالذات المتمثلة بتكامل دالة كاما الناقص، هذا الى جانب استعمال الطريقة التقليدية وهي الامكان الاعظم Maximum Likelihood: ML حيث تتم المقارنة على اساس اسلوب الجهاز المركزي للإحصاء في احتساب دالة البقاء من خلال برنامج MORTPAK كقيم حقيقية. وبعد ذلك القيام بالمقارنة باستعمال معيار جذر متوسط مربعات الخطأ Root Mean Square Error: RMSE ، ومعيار متوسط مطلق نسبة الخطأ Mean Absolute Percent Error: MAPE . وأظهرت النتائج أفضلية طريقة الانتروبي في تقدير دالة البقاء على الطرائق الاخرى.

المصطلحات الرئيسية للبحث / توزيع كاما العام، دالة البقاء، دالة الامكان الاعظم، ،مبدأ اعظم دالة انتروبي، الدالة اللبّية.

مجلة العلوم
الاقتصادية والإدارية
العدد ٩٣ المجلد 22
الصفحات ٤٥٤-٤٦٧

١-١ المقدمة Introduction

ان مسألة التعامل مع دوال البقاء تحتاج الى تمثيلها بدالة احتمالية معينة تسلك سلوك توزيع مناسب يتلائم مع ظاهرة الحياة (أوالموت)، وهو في هذا البحث سيكون توزيع كاما العام ذو المعلمات الثلاث كونه يتميز بدقة وحساسية عاليتين يتكيّف لصياغة دالة البقاء سواء في التطبيقات الديموغرافية أو الطبية. الا أن التعامل مع وصف لهذا التوزيع يتطلّب قدراً كبيراً من الجهد الرياضي عند محاولة تقدير معلماته الثلاثة آنياً فتبرز الحاجة الى طرائق عددية لغرض القيام بعملية التقدير بسبب التعقيد الناجم عن ذلك.

تحليل البقاء على قيد الحياة هي واحدة من التقنيات المستخدمة على نطاق واسع في مجال الإحصاءات السكانية؛ تنشأ أهميته أيضا في مجالات متنوعة مثل: الطب والهندسة وعلم الأوبئة وعلم الأحياء، والفيزياء، والصحة العامة أو تحليل التاريخ حدث في علم الاجتماع وفي التطبيقات الصناعية فان عادة ما يكون الوقت المناسب لفشل وحدة فنية وفي لاقتصاد يدل على الوقت حتى قبول وظيفة بحلول العاطلين عن العمل.

2-1 مشكلة البحث Problem of the Research

استعمال الاسلوب التقليدي في التقدير امراً بالغ الصعوبة حيث غالباً ماتعاني البيانات الديمغرافية من وجود اضطراب وتشوش بالسلوك، خصوصا فيما يتعلق بدالة البقاء لسكان العراق تبعا للفئات العمرية لأسباب تتعلق بالنظام الغذائي المتبّع، وتوفر عوامل الأمن الغذائي والرعاية والتأمين الصحي، ومستوى الرفاهية أو الحرمان، ناهيك عن الأوضاع الأمنية والسياسية على حد سواء. هذا بالاضافة الى وجود عائق آخر يتمثل بصعوبة الصياغة الرياضية للتوزيع الاحتمالي المتمثل بتوزيع كاما العام Generalized Gamma: GG بثلاثة معلمات، اذ يتضمن وجود تكامل كاما الناقص في جوهر عملية التقدير كجزء من الاستدلال الاحصائي.

3-1 هدف البحث The Aim of the Research

يهدف هذا البحث الى مقارنة وايجاد تقدير لدالة البقاء على قيد الحياة بعد تقدير معلمات توزيع كاما العام GG بثلاثة معلمات، اذ هناك حاجة ضرورية الى طريقة عددية للتقريب ومن ثم التقدير باحتمالات البقاء باستعمال عدة طرق معلمية (طريقة الامكان الاعظم Maximum Likelihood: ML وطريقة مبدأ اعظم دالة انتروبي POME)، أضافة الى الطريقة اللامعلمية (طريقة اللبية Kernel) كأحدى طرائق التمهيد لإظهار ودراسة فرص بقاء الفرد من سكان على قيد الحياة خلال العمر t. وإجراء المقارنات بين طرائق التقدير أعلاه بأستخدام جذر متوسط مربعات الخطأ (RMSE) ومتوسط مطلق نسبة الخطأ(MAPE) كمعايير لمقارنة جودة التقدير.

2. الجانب النظري Theoretical Part

1-2 المقدمة

يتم تحديد صيغة دالة الكثافة الاحتمالية (pdf) $f(t)$ ودالة التوزيع التراكمية (CDF) $F(t)$ ودالة البقاء $S(t)$ لتوزيع كاما العام بثلاثة معلمات وعرض بعض خصائصه مثل المتوسط والتباين والحالات الخاصة له. وايضا يبين هذا الفصل كيفية اشتقاق وتقدير معلمات ودالة البقاء باستخدام طرق مختلفة مثل الامكان الاعظم ML ومبدا اعظم دالة انتروبي POME اضافة الى تقدير دالة البقاء باستخدام طريقة اللامعلمية وهي طريقة لبية (Kernel) باستعمال دوال لب (Gaussian) لانسجامه كصيغة اسية مع التوزيع كاما العام GG كأحد توزيعات العائلة الاسية.

2-2 توزيع كاما العام Generalized gamma: GG [2]

يرجع اصل توزيع كاما العام بأربعة معلمات (k,β,θ,μ) الى العالم Amoroso في عام [1925] [3]ويعتبر الاصل لنموذج العمر حيث ان (β, k) تمثل معلمات الشكل و θ معلمة قياس و μ معلمة موقع او ازاحة الدالة الاحتمالية لتوزيع:

$$\text{Amoroso}^{(k,\beta,\theta,\mu)} = \frac{1}{\Gamma(k)}\left|\frac{\beta}{\theta}\right| \left(\frac{t-\mu}{\theta}\right)^{\beta k-1} e^{-\left(\frac{t-\mu}{\theta}\right)^{\beta}} \quad t,\beta,\theta,\mu \in R$$

$$t \geq \mu \; if \; \theta > 0 \; , t \leq \mu \; if \; \theta < 0$$

وقد تم تقديم في وقت لاحق توزيع كاما العام بثلاثة معلمات من قبل Stacy في عام [1962] [4] حيث افترض ان معلمة الموقع $\mu = 0$.

اما الدالة الاحتمالية لتوزيع GG ذي المعلمات الثلاثة بالصيغة التالية[5] :

$$f(t,\theta,\beta,k) = \frac{\beta}{\Gamma(k)\,\theta^{\beta k}}\; t^{\beta k-1}\; e^{-(\frac{t}{\theta})^{\beta}} \quad t>0, k>0, \theta>0, \beta>0 \quad \dots (1)$$

اذا ان:

$t \Leftarrow$ تمثل المتغير العشوائي (العمر)، $\theta \Leftarrow$ معلمة قياس (Scale Parameter)

$k, \beta \Leftarrow$ معلمات شكل (*Shape Parameter*)

بينما تكون الدالة الاحتمالية التراكمية:

$$F(t) = \frac{IG((\frac{t}{\theta})^{\beta},k)}{\Gamma(k)} \quad \dots (2)$$

اذ ان :

$$IG(S,K) = \frac{1}{\Gamma(k)} \int_0^s u^{k-1} e^{-u}\, du$$

$IG(S,K)$ (incomplete gamma function) دالة كاما الناقصة

وان قيمة المتوسط للمتغير t من مرتبة β :[6]

$$E(t^{\beta}) = \frac{\theta^{\beta}\Gamma(k+1)}{\Gamma(k)} = \theta^{\beta} k \quad \dots (3)$$

ومتوسط اللوغاريتم للمتغير t :[6]

$$E(\ln t) = \frac{\Psi(k)}{\beta} + \ln\theta \quad \dots (4)$$

2-2 تحليل دالة البقاء Survival Function analysis [7]

يعدّ تحليل البقاء على قيد الحياة هو فرع من فروع الإحصاء الرياضي التي تتناول تحليل المدة الزمنية للأحداث إلى حين ظهور واحد أو أكثر من الأحداث مثل الموت في الكائنات البيولوجية والفشل في الأنظمة الميكانيكية. ويسمى هذا الموضوع نظرية المعّولية أو تحليل المعّولية في مجال الهندسة للمواد الصناعية غير الحية.[8]

يحاول تحليل البقاء على قيد الحياة الإجابة على أسئلة مثل: ما هي نسبة عدد السكان والتي سوف تبقى على قيد الحياة إلى وقت معين؟ ومن تلك الفئات العمرية أيضا التي تستطيع البقاء على قيد الحياة؟ ماهي نسبة الموت أو الفشل؟ وكيف يمكن تحديد أسباب متعددة للموت أو الفشل؟ وكيف إن ظروف أو خصائص معينة تزيد أو تقلل من احتمال البقاء على قيد الحياة؟[8]

وللإجابة عن هذه الأسئلة، لا بد من تحديد "وقتاً للحياة". في حالة البقاء البيولوجي، والموت وهو أمر لا لبس فيه، ولكن المعّولية الميكانيكية، والفشل قد لا تكون واضحة المعالم في الجانب الصناعي مثلاً. وبعض الأحداث (على سبيل المثال ،الإصابة بنوبة قلبية أو فشل جهاز عضوي آخر) قد يكون لها نفس الغموض.[8] وتعرّف دالة البقاء ببساطة على أنها احتمال ان حدث (الموت) لم يحدث حتى الان حسب الزمن t وبالتالي فان T يدل على الزمن حتى الموت فان s(t) يدل على احتمال البقاء على قيد الحياة ، أذ ان:[7]

$$S(t)=1-F(t) = P(T > t) \qquad \text{for } t > 0$$

ولنفترض ان $S(0) = 1$ أي أن احتمال بقاء المصاب على قيد الحياة في الزمن (0) يساوي واحد وكذلك يجب أن تكون دالة البقاء غير متزايدة ومستمرة من الجانب الايمن:

$$S(u) \leq S(t) \qquad if u > t$$

فاحتمال البقاء غالبا يفترض انه يقترب من الصفر كلما ازداد عمر الكائن الحي (الانسان مثلاً):

$$S(t) \to 0 \quad as\ t \to \infty$$

ومن خصائص بيانات البقاء على قيد الحياة هو ان وقت البقاء على قيد الحياة لا يمكن ان يكون سالبا.
ودالة البقاء للتوزيع المستمر هي:

$$S(t) = p(T > t) = \int_t^{tmax} p(t)dt$$

فان دالة البقاء لتوزيع كاما العام تمثل حسب المعادلة (1):[7]

$$S_{GG}(t) = 1 - \frac{IG((\frac{t}{\beta})^{\beta}, k)}{\Gamma(k)} \qquad \dots (5)$$

3-2 طرائق التقدير Estimation Methods

هناك العديد من الطرائق المختلفة لتقدير دالة البقاء على قيد الحياة او منحنى دالة البقاء على قيد الحياة، منها: الطريقة المعلمية التقليدية التي تستخدم لنمذجة البيانات البقاء على قيد الحياة، وانها تختلف من حيث الافتراضات التي يتم اجراؤها حول توزيع البقاء لعدد من السكان ويشمل بعض التوزيعات الاسية الشائعة ومنها توزيع كاما العام. بينما يتم استعمال طريقة اللامعلمية دون فرض من اي القيود حول كيفية احتمال ان يصاب شخص ما في الحدث يتغير مع مرور الوقت.
و سيتم تقدير دوال البقاء باستعمال:

1- الامكان الاعظم لتوزيع كاما العام (GG) حيث يعتبر t المتغير العشوائي.
2- مبدأ تعظيم دالة الانتروبي POME.
3- دالة تمهيد لامعلمية هي دالة لبّية (kernel) من نوع (Gaussian).

1-3-2 تعظيم دالة الامكان Maximizing Likelihood Function

تعدّ طريقة الامكان الاعظم احدى الطرق التقليدية في التقدير على افتراض ان المعلمة المطلوب تقديرها ثابتة وليست متغيرة، اذ تعتبر طريقة الامكان الاعظم واحدة من الطرق المهمة وشائعة الاستعمال في التقدير كونها تتضمن خصائص جيدة منها الثبات والاتساق غالباً وليس دائماً، اذ ان مبدأ هذه الطريقة يكمن في ايجاد تقدير للمعلمات التي تجعل دالة الامكان في نهايتها العظمى.[9]

دالة الامكان لتوزيع كاما العام ذي المعلمات الثلاثة (θ, k, β) يعبر عنها بالمعادلة الأتية[1]:

$$L=\prod_{i=1}^{n} f(t_i, k, \beta, \theta) \qquad \dots (6)$$

$$L=\left(\frac{\beta}{\Gamma(k)\theta^{\beta k}}\right)^n \prod_{i=1}^{n} t_i^{\beta k-1} e^{-\frac{\sum_{i=1}^{n} t_i^{\beta}}{\theta^{\beta}}} \qquad \dots (7)$$

[1] الاشتقاق من قبل الباحثين

لتحويل دالة الامكان (7) الى الشكل الخطي يتم أخذ اللوغاريتم الطبيعي لها:

$$\ln L = n \ln \beta - n\beta k \ln \theta - n \ln \Gamma(k) + (\beta k-1)\sum_{i=1}^{n} \ln t_i - \frac{\sum_{i=1}^{n} t_i^{\beta}}{\theta^{\beta}}$$

$$= n \ln \beta - n\beta k \ln \theta - n \log \Gamma(k) + \beta k \sum_{i=1}^{n} \ln t_i - \sum_{i=1}^{n} \ln t_i - \frac{\sum_{i=1}^{n} t_i^{\beta}}{\theta^{\beta}} \quad \ldots(8)$$

لإيجاد القيم التقديرية للمعلمات الثلاثة (θ, k, β) نجد المشتقات الجزئية للمعادلة (8) نسبة الى المعالم:

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n\beta k}{\theta} + \beta \theta^{-\beta-1} \sum_{i=1}^{n} t_i^{\beta} \quad \ldots(9)$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{n}{\beta} - n k \ln \theta + k \sum_{i=1}^{n} \ln t_i - \left(\frac{\sum_{i=1}^{n} t_i}{\theta}\right)^{\beta} (\ln t_i + \ln \theta) \ldots (10)$$

$$\frac{\partial \ln L}{\partial k} = -n\beta \ln \theta - n \Psi(k) + \beta \sum_{i=1}^{n} \ln t_i \quad \ldots(11)$$

ومن ثم مساواة المشتقات الجزئية الى الصفر نحصل على القيم التقديرية للمعالم التي تجعل دالة الامكان اعظم ما يمكن:

$$\hat{k} = \frac{\hat{\theta}^{-\hat{\beta}} \sum_{i=1}^{n} t_i^{\hat{\beta}}}{n} \quad \ldots (12)$$

$$\hat{\theta} = \left(\exp\left(-\Psi(\hat{k}) + \frac{\hat{\beta} \sum_{i=1}^{n} \ln t_i}{n}\right)\right)^{\frac{1}{\hat{\beta}}} \quad \ldots (13)$$

حيث ان :

$$\Psi(k) = \frac{d}{dk} \ln(\Gamma(k)) = \frac{\Gamma'(k)}{\Gamma(k)}$$

بعد تعويض القيم التقديرية $\hat{\theta}$ و $\hat{k}$ يتم استخراج القيمة التقديرية ل β من المعادلة (10)بطريقة عددية .
وباستعمال خاصية الثبات نحصل على مقدر دالة البقاء لتوزيع كاما العام لثلاثة معلمات بالصيغة الآتية:

$$\hat{S}_{GG}(t) = 1 - \frac{IG\left(\left(\frac{t}{\hat{\theta}}\right)^{\hat{\beta}}, \hat{k}\right)}{\Gamma(\hat{k})} \quad \ldots (14)$$

2-3-2 **مبدأ تعظيم دالة الانتروبي** Principle of Maximizing Entropy [11],[10]

يمكن القول عن دالة انتروبي شانون [12](Shannon Entropy) للمتغير العشوائي المستمر (t) أنها نطاق المعلومات المقترنة بالتوزيع الاحتمالي $f(t,\theta)$، حيث θ تمثل متجه المعلمات التي تصف المتغير العشوائي (t)، ويمكن ان يعبر عنها:

$$H(f) = -\int_{-\infty}^{\infty} f(t,\theta) \ln f(t,\theta)\, dt \quad \ldots (15)$$

أذ ان :

$$\int_{-\infty}^{\infty} f(t,\theta)\, dt = 1 \quad \ldots (16)$$

يمثل الرمز $H(f)$ دالة الانتروبي للدالة $f(t,\theta)$

يمكن كتابة دالة الانتروبي بالشكل الاتي :

$$H(f) = E - \ln f(t,\theta) \quad \dots (17)$$

حيث ان $H(f)$ هو متوسط القيمة $-\ln f(t,\theta)$

فان التوزيع الأقل تحيز بالنسبة للمتغير $t)$) هو ذلك التوزيع الذي يعمل على تعظيم دالة انتروبي طبقاً للمعلومات المعطاة ، أي انه الذي يحقق مبدأ أعظم دالة انتروبي (*Principle of Maximum Entropy*) حسب ما جاء به الباحث (Jaynes, 1961) .

ونتيجة لذلك فأن الحصول على معلمات التوزيع يتحقق من خلال تعظيم $H(f)$ *Maximum.*
ادخال قيود جديدة تتعلق بالمعلمات او الحصول على دالة انتروبي جديدة مقيدة حيث يمكن توظيف مسائل الامثلية مثل مضاعفات لاكرانج عند تقدير المعلمات.

ولتوضيح الفكرة، لنفترض وجود m من القيود الخطية المستقلة (C_i) حيث ان $i = 1,2,\dots m$

$$ci = \int w_i \, f(t,\theta)\,dt \quad \dots(18)$$

حيث w_i هي دوال معلومة معرفة على $f(t,\theta)$ ، فسيتم تطبيق تلك القيود ci لتعظيم دالة $H(f)$ والتي ستنتج الدالة الاحتمالية الاتية:

$$f(t,\theta) = e^{-a_0 - \sum_{i=1}^{m} a_i w_i(x)} \qquad i = 1,2,\dots m \quad \dots(19)$$

اذ ان a_i : هي مضاعفات لاكرانج (Lagrange Multipliers)
وبتعويض المعادلة (19) في (15) يتم التوصل الى:

$$H(f) = a_0 + \sum_{i=1}^{m} a_i c_i \quad \dots(20)$$

عند تعويض دالة الاحتمالية لتوزيع كاما العام كما في المعادلة (1) في المعادلة (15) يتم استخراج القيود[٢]:

$$H(f) = -\int_0^{\infty} f(t) ln \frac{\beta}{\Gamma(k)\,\theta^{\beta k}}\, t^{\beta k-1}\, e^{-(\frac{t}{\theta})^{\beta}}\, dt$$

$$= -ln\,\beta + ln\,\Gamma(k) + k\beta\, ln\,\theta - (k\beta - 1)\, E(\ln t) + \frac{1}{\theta^{\beta}}\, E(t^{\beta}) \quad \dots(21)$$

$$a_0 = \ln\Gamma(k) - \ln\beta + k\beta \ln\theta$$

$$a_1 = \frac{1}{\theta^{\beta}}$$

$$a_2 = -(k\beta - 1) = 1 - k\beta$$

بعد الاشتقاق يتم التوصل الى دالة احتمالية بدلالة القيود ومضاعفات لاكرانج :

$$f(t,a_1,a_2,\beta) = \frac{\beta\, a_1^{\frac{1-a_2}{\beta}}}{\Gamma(\frac{1-a_2}{\beta})}\, t^{-a_2}\, e^{(-\frac{t^{\beta}}{a_1-1})} \quad \dots (22)$$

نوجد المتوسط للمتغير t للمعادلة (22) المرفوع للاس β: [1]

$$E(t^{\beta}) = \frac{\frac{1-a_2}{\beta}}{a_1} \quad \dots (23)$$

نستخرج متوسط اللوغاريتم للمتغير t من المعادلة (22):[1]

$$E(\ln t) = \frac{\Psi(\frac{1-a_2}{\beta})}{\beta} - \frac{\ln a_1}{\beta} \quad \dots (24)$$

[٢] الاشتقاق من قبل الباحثين

نوجد الدالة التراكمية للمعادلة (22):[1]

$$F(t,a_1,a_2,\beta)=\frac{IG(\frac{t^{\beta}}{a_1{}^{-1}},\frac{1-a_2}{\beta})}{\Gamma(\frac{1-a_2}{\beta})} \quad \dots (25)$$

ومن المعادلة (22) نوجد دالة البقاء ل $f(t,a_1,a_2,\beta)$:

$$S(t,a_1,a_2,\beta)=1-\frac{IG(\frac{t^{\beta}}{a_1{}^{-1}},\frac{1-a_2}{\beta})}{\Gamma(\frac{1-a_2}{\beta})} \quad \dots (26)$$

بعد تعويض الدالة الاحتمالية بدلالة القيود والمضاعفات في المعادلة (17) ينتج:

$$H(f)=-\ln\beta-\left(\frac{1-a_2}{\beta}\right)\ln a_1+\ln\Gamma\left(\frac{1-a_2}{\beta}\right)+a_2E(\ln t)+\frac{Et^{\beta}}{a_1{}^{-1}} \quad \dots (27)$$

العلاقة بين معلمات التوزيع والقيود[1]:

$$\frac{dH(f)}{da_1}=\frac{\frac{1-\hat{a}_2}{\hat{\beta}}}{\hat{a}_1}-E(t^{\hat{\beta}})=0 \quad \dots (28)$$

$$\frac{dH(f)}{da_2}=\frac{1}{\hat{\beta}}\ln\hat{a}_1-\frac{1}{\hat{\beta}}\Psi(\frac{1-\hat{a}_2}{\hat{\beta}})+E(\ln t)=0 \quad \dots (29)$$

$$\frac{dH(f)}{d\beta}=-\frac{1}{\hat{\beta}}+\frac{1}{\hat{\beta}^2}(1-\hat{a}_2)\ln\hat{a}_1+\Psi\left(\frac{1-\hat{a}_2}{\hat{\beta}}\right)\left(-\frac{1}{\hat{\beta}^2}\right)+\frac{E(t^{\hat{\beta}})\ln(t)}{\hat{a}_1{}^{-1}}=0 \quad \dots (30)$$

بسبب صعوبة حل المعادلات سيتم اللجوء الى طريقة نيوتن رافسن لتقدير a_1 من المعادلة (28) وa_2 من المعادلة (29) و β من المعادلة (30).

وعند تقدير المعلمات (a_1,a_2,β) نقدر دالة البقاء حسب الصيغة الاتية :

$$\hat{S}(t,\hat{a}_1,\hat{a}_2,\hat{\beta})=1-\frac{IG(\frac{t^{\hat{\beta}}}{\hat{a}_1{}^{-1}},\frac{1-\hat{a}_2}{\hat{\beta}})}{\Gamma(\frac{1-\hat{a}_2}{\hat{\beta}})} \quad \dots (31)$$

3-3-3 **دالة تمهيد لبّية** Kernel Smoothing Function

ان المسألة المهمة والرئيسة في التطبيقات الاحصائية تتمثل بمعرفة التوزيع الخاص بالمجتمع المطلوب دراسته ومعرفة خصائص ذلك المجتمع كي يتم تمثيل المجتمع تمثيلاً سليماً من خلال استعمال الاساليب الاحصائية الشائعة.[13]

ان التوزيع الاحتمالي لأي متغير عشوائي يتم وصفه بدلالة دالة الكثافة الاحتمالية، فإذا جمعنا عينة عشوائية مكونة من n من المشاهدات من المجتمع المطلوب دراسته يمكن التعبير عن ذلك المجتمع من خلال دالة الكثافة الاحتمالية التي يتم التعرف من خلالها على خصائص المجتمع المدروس، تمثل دالة الكثافة الاحتمالية مفهوم اساسي في الاحصاء وان تحديد الدالة تعطينا وصفاً لتوزيع ، إذ يمكن تحديد الاحتمالات المرتبطة مع من خلال العلاقة الاتية:[13]،[14]

$$p_r(a\le t\le b)=\int_a^b f(t)dt \quad \forall\ a,b\in R$$

ان الهدف من تقدير دالة الكثافة الاحتمالية اللامعلمية يتمثل بتقريب دالة الكثافة الاحتمالية $f(t)$ للمتغير العشوائي t_i سيتم استعمال مقدر (Kernel) في تقدير دالة الكثافة بالصيغة الاتية:-

$$\hat{f}(t)=\frac{1}{n\hat{h}}\sum_{i=1}^{n}k\ (\frac{t-t_i}{h}) \quad \dots (32)$$

اذ ان h:[15](bandwidth) عرض الحزمــة او تعرف بمعلمة التمهيد (smooth parameter)

$$\hat{h} = 1.06(s)n^{-1/5} \quad \dots (33)$$

تم لاستناد الى طريقة النسبة الذهبية لتقدير عرض الحزمة(Golden Rate method) بقسمة قيمة $\hat{h}$ كما في المعادلة (33)على النسبة الذهبية نحصل على h مثلى وكالاتي:

$$\hat{h}_{GR} = \frac{\hat{h}}{1.618} \quad \dots (34)$$

هناك العديد من دوال اللب تم استعمال في البحث دالة (Gaussian)[13] .

$$k(u) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}} \qquad I(-\infty,\infty)$$

اذ ان :

$$u = \frac{t-t_i}{h}$$

وبعد تقدير دالة الكثافة الاحتمالية سيتم حساب احتمال البقاء على قيد الحياة حسب الصيغة الآتية:[16]

$$\hat{s}(t) = p(T > t)$$
$$= \int_x^{\infty} \hat{f}\,(T)dT$$
$$= \frac{1}{n}\sum_{i=1}^{n} \int_{\frac{t-t_i}{h}}^{\infty} k(u)du \quad \dots (35)$$

4-2 معايير المقارنة Comparison Criteria

يتم استعمال بعض معايير المقارنة للتمييز بين طرائق تقدير دالة البقاء المستعملة في البحث وهي كما في أدناه.

1-4-2 جذر متوسط مربعات الخطأ[17] Root Mean Square Error (RMSE)

أن الصيغة العامة لـ (RMSE) هي:

$$RMSE(\hat{s}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s_i})^2} \quad \dots (36)$$

2-4-2 متوسط مطلق نسبة الخطأ[17] Mean Absolute Percent Error (MAPE)

ان الصيغة العامة لـ (MAPE) هي:

$$MAPE(\hat{s}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{s_i-\hat{s_i}}{s_i}\right| \quad \dots (37)$$

3 الجانب التطبيقي Application Part

1-3 المقدمة Introduction

من خلال الجانب النظري يبين كيفية تقدير دالة البقاء من خلال عدة طرق وهي (ML،Entropy) وطريقة الامعلمية (kernel). في هذا المبحث سيتم استعراض كيفية حساب وتقدير دالة البقاء على اساس دالة البقاء التي يتم حسابها في الجهاز المركزي للإحصاء التي يتم الاعتماد عليها من قبل الامم المتحدة (UN) لبيانات خماسية فئات العمر.

واخيرا يتم حساب معايير المقارنة للطرق الثلاثة للذكور والاناث والذكور والاناث معا، وقد تم اختبار البيانات باستعمال البرنامج الاحصائي (Easy fit) اذ وجد ان البيانات تتبع توزيع كاما العام (GG) ولان توزيع كاما العام من التوزيعات الاسية المستعملة لوصف معدل وفيات البشر تم استعمال توزيع كاما العام في تقدير دالة البقاء .

2-3 وصف البيانات Data Description

تم جمع البيانات من الجهاز المركزي للإحصاء حول الاوضاع المعيشية للأسرة العراقية (المسح الاجتماعي الاقتصادي للأسر)(IHSES II 2012)[18] الذي يوفر منظومة معلومات متكاملة تفصيله، واجمالية عن الوضع المعيشي للفرد والأسرة العراقية، ليقدم صورة رقمية موضوعية عن تطور الأوضاع المعيشية في المجالات المختلفة (الإنفاق والدخل، الوضع الديموغرافي، الصحة، السكن، التشغيل وغيرها). كما إن هذا الملف سيخدم شرائح واسعة من مستخدمي هذا النوع من البيانات على المستويين القطاعي في تنوع جوانب التغطية الإحصائية للمؤشرات، والمكاني في مستوى الشمول والقدرة على توفير مؤشرات على مستوى المحافظات والأقضية. بلغ حجم العينة (25488) أسرة، في جميع محافظات العراق بواقع (216) أسرة لكل قضاء من أقضية العراق ال (118) كما بلغ عدد العناقيد (2832) عنقود ويشمل كل عنقود (9) أسر موزعة على الأقضية والمحافظات وللبيئتين الحضرية والريفية.

3-3 ايجاد دوال البقاء

مقارنة دوال البقاء لطرق التقدير الثلاثة (تعظيم دالة الامكان Ml،مبدأ تعظيم الانتروبي Ent ، دالة اللبية Kernel) مع طريقة الجهاز المركزي للإحصاء (UN) للفئات الخماسية تم استخراج دالة البقاء للفئات الخماسية عن طريق برنامج[19] MORT PAK حيث يتم ادخال البيانات الديمغرافية الناتجة من قسمة عدد الوفيات على عدد السكان m(x,n) لكل فئة لسنة (2012) .

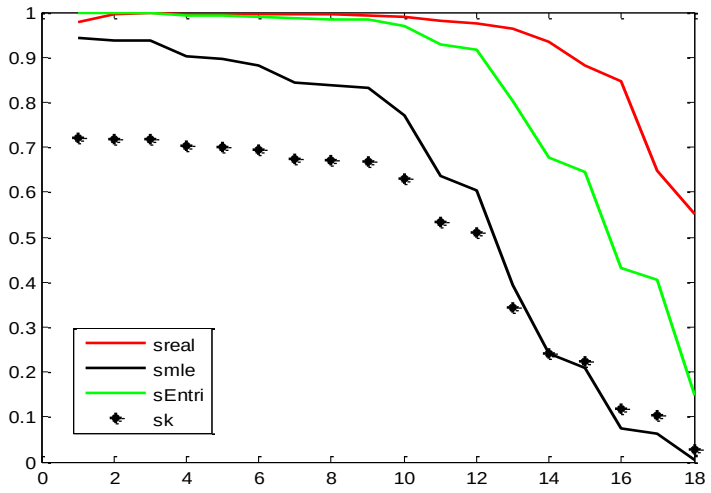
جدول رقم (1) يبين مقارنة الطرق على اساس طريقة لـ UN للذكور والاناث معا

فئات	S_{real}	$\hat{S}_{mle}$	$\hat{S}_{ent}$	$\hat{S}_k$
1-0	0.97333	0.9330	0.9972	0.7294
4-1	0.994059	0.9290	0.9968	0.7279
9-5	0.996414	0.9080	0.9948	0.7197
14-10	0.99579	0.8856	0.9922	0.7099
19-15	0.993778	0.8695	0.99	0.7024
24-20	0.992009	0.8664	0.9895	0.7009
29-25	0.990451	0.8508	0.9871	0.6932
34-30	0.989191	0.8364	0.9846	0.6856
39-35	0.98571	0.7813	0.9733	0.654
44-40	0.98015	0.7324	0.9608	0.6223
49-45	0.965538	0.5763	0.9039	0.5028
54-50	0.954005	0.5710	0.9015	0.4984
59-55	0.93286	0.3462	0.7628	0.3124
64-60	0.889227	0.2752	0.6994	0.2596
69-65	0.8084	0.2342	0.6566	0.2312
74-70	0.752515	0.0672	0.3877	0.1223
79-75	0.580038	0.0524	0.3476	0.1003
80فاكثر	0.500000	0.0075	0.146	0.0278

جدول رقم (2) يمثل المقاييس الاحصائية (RMSE,MAPE) لمقارنة طريقة UN مع الطرق التقديرية:

Methods	RMSE	MAPE
MLE	0.3758	0.3863
Kernel	0.4258	0.4795
Entropy	0.1506	0.1319

من خلال النتائج المبينة في الجدول (2) نلاحظ ان طريقة الانتروبي هي الافضل من بين الطرائق الاخرى في تقدير دالة بقاء السكان للذكور والاناث معا وللمعياريين MAPE،RMSE.



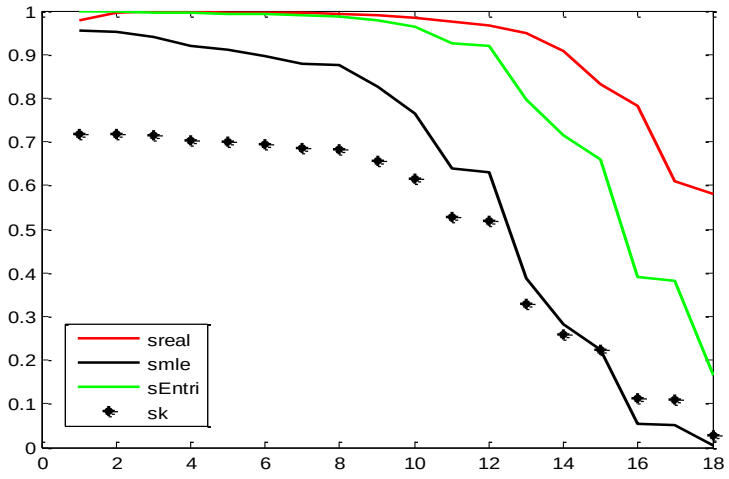
جدول رقم (3) يبين مقارنة الطرق على اساس طريقة UN للإناث

فئات	S_{real}	$\hat{S}_{mle}$	$\hat{S}_{mle}$	$\hat{S}_k$
1-0	0.976923	0.9542	0.9979	0.7192
4-1	0.995222	0.9514	0.9977	0.7181
9-5	0.997663	0.9411	0.9967	0.714
14-10	0.997321	0.9192	0.9944	0.7046
19-15	0.99616	0.9113	0.9934	0.701
24-20	0.995058	0.8956	0.9913	0.6935
29-25	0.994228	0.8786	0.9887	0.6849
34-30	0.993017	0.8760	0.9883	0.6836
39-35	0.990105	0.8249	0.9787	0.6551
44-40	0.984824	0.7642	0.9643	0.6169
49-45	0.974412	0.6396	0.9243	0.5267
54-50	0.966805	0.6304	0.9208	0.5196
59-55	0.948642	0.3880	0.7953	0.3303
64-60	0.906481	0.2830	0.7145	0.2588
69-65	0.832919	0.2250	0.6589	0.2233
74-70	0.782093	0.0553	0.3912	0.1139
79-75	0.608509	0.0521	0.3822	0.1087
80فاكثر	0.580000	0.0056	0.1646	0.0278

جدول رقم (4) يمثل المقاييس الاحصائية (RMSE,MAPE) لمقارنة طريقة UN مع الطرق التقديرية:

Methods	RMSE	MAPE
MLE	0.3799	0.3637
Kernel	0.4411	0.4835
Entropy	0.1621	0.1304

من خلال النتائج المبينة في الجدول (4) نلاحظ ان طريقة الانتروبي هي الافضل من بين الطرائق الاخرى في تقدير دالة بقاء السكان للإناث وللمعياريين RMSE,MAPE.



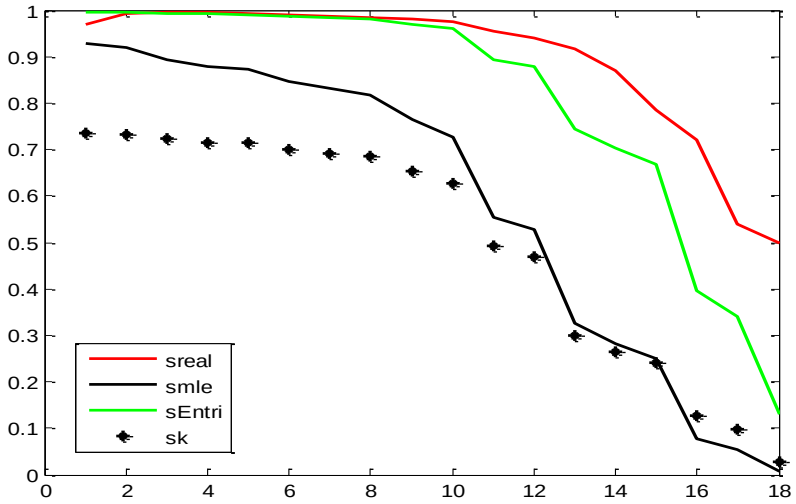
جدول رقم (5) يبين مقارنة الطرق على اساس طريقة UN للذكور

$\hat{S}_k$	$\hat{S}_{ent}$	$\hat{S}_{mle}$	S_{real}	فئات
0.736225	0.996962	0.9278	0.969821	1-0
0.732977	0.996219	0.9192	0.99303	4-1
0.72299	0.993691	0.8946	0.995245	9-5
0.715416	0.991574	0.8775	0.994357	14-10
0.713694	0.991072	0.8738	0.991565	19-15
0.700606	0.987039	0.8469	0.989133	24-20
0.692	0.984204	0.8305	0.986683	29-25
0.68502	0.981812	0.8177	0.98521	34-30
0.653762	0.970239	0.7651	0.981189	39-35
0.627456	0.959578	0.7254	0.975469	44-40
0.493563	0.89385	0.5546	0.955107	49-45
0.470134	0.8801	0.5274	0.93926	54-50
0.299629	0.74323	0.3272	0.916245	59-55
0.265295	0.70247	0.2834	0.871043	64-60
0.240674	0.668248	0.2505	0.784453	69-65
0.125985	0.397521	0.0765	0.722215	74-70
0.096795	0.340771	0.0548	0.53847	79-75
0.027779	0.13105	0.0072	0.50000	80فاكثر

جدول رقم (6) يمثل المقاييس الاحصائية (RMSE,MAPE) لمقارنة طريقة UN مع الطرق التقديرية:

Methods	RMSE	MAPE
MLE	0.3673	0.3910
Kernel	0.4143	0.4754
Entropy	0.1416	0.1269

من خلال النتائج المبينة في الجدول (6) ومن نلاحظ ان طريقة الانتروبي هي الافضل من بين الطرائق الاخرى في تقدير دالة بقاء السكان للذكور وللمعياريين RMSE,MAPE.



4 الاستنتاجات والتوصيات

4-1 الاستنتاجات

بعد التوصل الى نتائج عددية بعد تطبيق الصيغ النظرية للطرائق المستعملة في تقدير احتمالات البقاء البحث، تم التوصل الى أفضلية استعمال تقدير مبدأ اعظم دالة انتروبي (POME) بحسب نتائج الجداول (2)،(4)،(6)،بأستعمال معياري المقارنة RMSE و MAPE على حد سواء برغم حساسية المعيار الأخير.

4-2 التوصيات

يوصي الباحثان بالآتي:

1- توسيع نطاق البحث في الموضوع بأدخال قيود جديدة تتعلق بالمعلمات والحصول على دالة انتروبي جديدة مقيدة، وعند اذ يمكن توظيف مسائل الامثلية مثل مضاعفات لاكرانج عند تقدير المعلمات ومن ثم تقدير دوال البقاء على قيد الحياة او المعولية.

2- استعمال توزيعات جديدة أخرى أكثر تعقيد، مثل توزيع F المعمّم وغيره.

3- استعمال طريقة POME لتقدير دالة البقاء على قيد الحياة لفئات أحادية العمر بدلاً عن الخماسية التي تحتوي على خمس سنوات لكل مجموعة عمرية وذلك لان الفئات الاحادية دقيقة تتحسس لأي عمر يواجه خطر الفناء (الموت) أكثر من غيره وعلى مستوى الجنس.

4- من اجل التغلب على التعقيد الرياضي لتوزيع كاما العام يتم اللجوء الى اعادة هيكلة المعلمات (reparametrization) .

5-نوصي الجهاز المركزي للإحصاء عند عمل مسوحات سكانية الاخذ بنظر الاعتبار الفئات احادية السن.

3-4 **الدراسات المستقبلية**

1-القيام بتحليل على اساس توزيع كاما العام متحيز الحجم (Size-biased Generalized Gamma) (SBGG) للبحث.
2- بحث في توزيع المعمم F (generalized F) (GF) يمثل مظلة لمعلمة تحليل البقاء على قيد الحياة.
3-مقارنة توزيعات اخرى مع توزيع كاما العام من خلال طريقة الانتروبي حيث يتم تفضيل التوزيع الاكثر انتروبية بمعنى ان التوزيع الأقل انتروبية يسمح بالقيود الاحتمالية الاقل.

المصادر

١. طارق، رغدة زياد ؛(2016)؛"استعمال دالة الانتروبي في تقدير توزيع كاما العام لتنبؤ باحتمالات البقاء لسكان في العراق" رسالة ماجستير غير منشورة علوم في الاحصاء ،كلية الادارة والاقتصاد، جامعة بغداد.

2.Lienhard, J.H.;Meyer, P.L.; (1967) ;" A Physical Basis for the Generalized Gamma Distribution"; Quarterly of Applied Mathematics,vol.25,No. 3, pp.330–334.

3.Amoroso, L.; (1925) ;" Ricerche intorno alla curve dei redditi "; Annali di Matematica Puraed Applicants ,vol.2(1), pp.123–159.

4.Stacy,. E.; (1962);" A Generalization of the Gamma Distribution". The Annals of Mathematical Statistics,vol.33,No. 3 ,pp.1187–1192.

5.Schutz, A.; Bombrum, L.; Berthoumieu, Y.; and Najim, M. ;(2013); "Centroid Based Texture Classification Using the Generalized Gamma Distribution"; In: Proc. EUSIPCO Marrakech, Morocco.

6.Korteza,K.;Ahmadabadi,A.;(2010);"Some properties of generalized gamma distribution ";Mathematical Sciences,Vol.4,No.1,pp.9-28.

7. Qamruz, Z.; and Karl, P.;(2011);"Survival Analysis Medical Research";http://interstat.statjournals.net/YEAR/2011/abstracts/1105005.php.

٨. سليمان، عباس نجم؛ فرحان، ابتهال حسين؛ (2014) ؛"تقدير دالة البقاء لبيانات حقيقية كاملة لمرضى سرطان الرئة "؛ مجلة ابن لهيثم للعلوم الصرفة والتطبيقية، المجلد، 27العدد (3)، ص:531-541.

٩. الجواد، ياسمين عبد الرحمن؛ (2013) ؛"بناء جداول الحياة الذاتية في العراق بإستعمال احتمالات البقاء "؛ اطروحة دكتوراه في علوم الاحصاء ،كلية الادارة والاقتصاد، جامعة بغداد.

10. Karmeshu;pal,N.R.;(2003);" Entropy Measures, Maximum Entropy Principle and Emerging Applications"; Springer, Verlag, Berlin ,Heidelberg.

11. Levine, R.D.; Tribus, M.;(1979) ;"The Maximum Entropy Formalism" ;MIT Press, Cambridge ,Massachusetts, USE.

12. Shannon,C.E.; (1948) ;" System Technical" ;PP.379-623.

١٣. حمود، مناف يوسف؛ (2005) ؛"مقارنة المقدرات الامعلمية لتقدير دوال الكثافة الاحتمالية" اطروحة دكتوراه في علوم الأحصاء، كلية الأدارة والأقتصاد، جامعة بغداد.

١٤. حمود، مناف يوسف؛ نايف، قتيبة نبيل؛ عباس، تهاني مهدي؛ (2008) ؛"تقدير لامعلمي كثافة الاحتمالية متعددة المتغيرات "؛ مجلة جامعة النهرين للعلوم، المجلد(11)،العدد(2)، ص:55-63.

١٥. ابراهيم ،مروة خليل؛ (2013) ؛"تقويم بيانات العمر والجنس للتعدادات السكانية مع تطبيق عملي لبيانات التعداد العام للسكان لسنة 1997في العراق " رسالة ماجستير علوم في الاحصاء، كلية الادارة والاقتصاد، جامعة بغداد.

16. Qiao, H. ;. Tsokos, C. P. ;(1994) ;" Non-parametric approach to reliability analysis"; IEEE, Dept. of Math. , Phys., Fort Valley State Coll., GA, USA.

١٧. ابراهيم، زينب علاوي؛(2013)؛" الدقة في تخطيط كمية انتاج مادة السمنت في معامل الشركة العامة للسمنت العراقية (دراسة اختبارية لطرق التنبؤ باستخدام معايير الخطأ) "؛ مجلة ديالى للعلوم الهندسية ،المجلد(7)،العدد(1)،ص:39-59 .

١٨. الجهاز المركزي للإحصاء، وزارة التخطيط ؛ (2014) ؛"المسح الاجتماعي والاقتصادي للأسرة في العراق 2012 IHSES II تقرير الجداول "؛ طبعت في مطبعة الجهاز المركزي للإحصاء، العراق. WWW.Cost.gov.iq

19. United Nations;(2013);"MORT PAK For Windows"; version 4.3;New-York;USA.

Proposal of Using Principle of Maximizing Entropy of Generalized Gamma Distribution to Estimate the Survival probabilities of the

Abstract

In this research we been estimated the survival function for data suffer from the disturbances and confusion of Iraq Household Socio-Economic Survey: IHSES II 2012 , to data from a five-year age groups follow the distribution of the Generalized Gamma: GG. It had been used two methods for the purposes of estimating and fitting which is the way the Principle of Maximizing Entropy: POME, and method of booting to nonparametric smoothing function for Kernel, to overcome the mathematical problems plaguing integrals contained in this distribution in particular of the integration of the incomplete gamma function, along with the use of traditional way in which is the Maximum Likelihood: ML. Where the comparison on the basis of the method of the Central Bureau of Statistics to stay through the program MORTPAK real function values calculated. And then compared to the use of Root Mean Square Error: RMSE, and Mean Absolute Percent Error: MAPE. The results showed preference entropy as optimal method to estimate survival function on other methods.

Keywords\ Generalized Gamma Distribution, survival function, the Maximum Likelihood, the principle of the Maximizing function entropy, kernel function.