IRAQI Academic Scientific Journals

KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES (2025) (02) (04): P(11-23)

خزائن للعلوم الاقتصادية والإدارية
KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES
ISSN: 2960-1363 (Print)
ISSN: 3007-9020 (Online)

A Comparison Between Estimation Methods in Semiparametric and Ordinary Poisson Regression Models for Clustered Count Data

Noor Alhuda Alaa Kazem [1], A. Prof. Dr. Iqbal Mahmoud Alwan [2]

[1,2] college of Administration and Economics / University of Baghdad
Noor.alaa2201m@coadec.uobaghdad.edu.iq
Iqbal.alwan@coadec.uobaghdad.edu.iq

Abstract. Clustered data are considened one of the most common model in applied studies because of it contrase within-cluster variability. In this regard, A semiparametric Poisson regression model was a dopted to analyze this type of data , throughout mergine the nonparametric procedune to represent the nonlinear relations between independent variables and the parametric procedune that models the random effects specific to each cluster. The two estimation methods were considered. The first method involves simultaneous estimation of both the parametric and nonparametric parts using the **Penalized Least Squares (PLS)** and. The second method estimates the nonparametric parts first, followed by the parametric parts using the **Backfitting algorithm**. These two methods were compared with the classical **Maximum Likelihood Estimation (MLE)** method used to fit the ordinary Poisson regression model, based on the **Mean Absolute Percentage Error (MAPE)** as the evaluation criterion. Simulation were done with several , considering three different values for the number of clusters: k = 3,5,10. Also Two types of cluster sizes were used:

- In the first, clusters had equal sizes: m_k = 10,30,50
- In the second, clusters had unequal sizes, drawn from total sample sizes of n = 50,100,250,500 resulting in varying cluster sizes.

The results indicated that the **Backfitting algorithm** superior the other methods in both experimentaly and practicly.

Keywords: Backfitting algorithm,Clustered data, Mean Absolute Percentage Error (MAPE), Penalized Least Squares, Poisson regression, semiparametric Poisson regression

DOI: 10.69938/Keas.2502042

مقارنة بين طريقتي التقدير إنموذج شبه المعلمي وإنموذج التقليدي في انحدار بواسون للبيانات العد العنقودي

نورالهدى علاء كاظم[١] ، أ.م.د. اقبال محمود علوان [٢]
[2,1] كلية الادارة والاقتصاد / جامعة بغداد

المستخلص. تُعد البيانات العنقودية من الأنماط الشائعة في الدراسات التطبيقية، لما تتضمنه من تباين داخل العناقيد. في هذا السياق، تم اعتماد إنموذج بواسون شبه المعلمي لمعالجة هذا النوع من البيانات، من خلال دمج الاسلوب غير المعلمي يُستخدم لتمثيل العلاقات غير الخطية بين المتغيرات التوضيحية، والاسلوب المعلمي يُعبّر عن التأثيرات العشوائية الخاصة بكل عنقود. وكانت هناك طريقتين : الطريقة الاولى يتم تقدير الجزء المعلمي والجزء غير المعلمي في آن واحد باستخدام المربعات الصغرى الجزائية (Penalized Least Squares) اما الطريقة الثانية يتم تقدير الجزء غير المعلمي اولاً من ثم تقدير الجزء المعلمي باستخدام خوارزمية التقدير التكراري (Backfitting), وتمت مقارنة بين الطريقتين مع طريقة الامكان الاعظم لتقدير إنموذج انحدار بواسون التقليدي باستعمال معيار متوسط الأخطاء النسبية المطلقة (MAPE)) , تم تنفيذ عدد من تجارب المحاكاة شملت ثلاث قيم لعدد العناقيد k=3,5,10 اما حجوم العناقيد تم استخدام حالتين: الحالة الاولى يكون فيه الحجوم متساوية في كل عنقود وهي m_k=10,30,50 اما الحالة الثانية يكون حجم كل عنقود مختلفة عن الاخر التي استخدام الحجوم الكلية n=50,100,250,500 ومن

خلالها يكون حجوم مختلفة لكل عنقود , واظهرت النتائج تفوق خوارزمية التقدير التكراري Backfitting على الطرائق الاخرى في جانبي التجريبي والتطبيقي .

الكلمات المفتاحية: البيانات العنقودية , انحدار بواسون , انحدار بواسون شبه المعلمي, خوارزمية التقدير التكراري, متوسط الأخطاء النسبية المطلقة, مربعات الصغرى الجزائية .

Corresponding Author: E-mail: Noor.alaa2201m@coadec.uobaghdad.edu.

1- المقدمة Introduction

في العديد من التطبيقات العلمية، كثيرًا ما لا تتحقق الفروض الأساسية للنماذج الخطية، مثل افتراض التوزيع الطبيعي أو خطية العلاقة بين المتغيرات. وتزداد هذه التحديات وضوحًا في حالات عدّ الأحداث خلال فترة زمنية أو في موقع مكاني معين، كما هو الحال في الدراسات الوبائية والبيئية، حيث يُستخدم إنموذج انحدار بواسون لملاءمته مع هذا النوع من البيانات، نظرًا لاعتماده على فرض تساوي المتوسط والتباين.

إلا أن هذا الإنموذج يفقد كفاءته الإحصائية عند ظهور مشكلة التشتت الزائد (Overdispersion) ، والتي تنشأ غالبًا في البيانات العدّ العنقودية، حيث تكون المشاهدات ضمن كل عنقود مترابطة جزئيًا، كما في بيانات المرضى داخل المستشفيات أو الطلبة داخل المدارس. هذا الترابط الداخلي يؤدي إلى تباين يفوق المتوسط، مما يضعف بفرضيات الإنموذج البواسوني ويحدّ من دقته التنبؤية .

وللتغلب على هذه المشكلة، يتم اللجوء إلى نماذج أكثر مرونة، ومن أبرزها إنموذج انحدار بواسون شبه المعلمي للبيانات العنقودية، الذي يُعد تطويرًا للإنموذج التقليدي، إذ يدمج بين جزء معلمي يصف العلاقة الخطية أو شبه الخطية، وجزء غير معلمي يُقدّر باستخدام تقنيات التمهيد(Smoothing) ، ما يسمح بتمثيل التأثيرات المعقدة وغير الخطية. كما يأخذ الإنموذج في الحسبان التأثيرات الخاصة بالعناقيد، مما يُعزز من القدرة على تمثيل الترابطات الداخلية بين المشاهدات داخل كل عنقود بدقة أكبر.

تتضمن **مشكلة البحث** , ان العديد من النماذج الاحصائية تفترض استقلالية المشاهدات وعدم وجود تأثيرات متغيرة بين العناقيد, هو افتراض قد لا يكون دقيقاً عند التعامل مع البيانات العد العنقودي التي تتميز بوجود ارتباط بين المشاهدات داخل كل عنقود. هذا الارتباط يمكن يؤدي الى تباين غير متجانس في تأثير المتغيرات المستقلة. لذلك, تبرز الحاجة الى تطوير نماذج تحليلية تأخذ بعين الاعتبار الطبيعة العنقودية للبيانات .

يهدف هذا البحث للتخلص من مشكلة التشتت الزائد (Overdispersion) في بيانات العدّ العنقودية والترابط للمشاهدات ضمن كل عنقود ويهدف ايضاً الى تقدير إنموذج انحدار بواسون شبه المعلمي لبيانات العد العنقودي باستعمال طريقتين اولاً عن طريق تقدير الجزء المعلمي والجزء غير المعلمي في آن واحد باستعمال طريقة المربعات الصغرى الجزائية Penalized Least Squares اما الطريقة الثانية تكون عن طريق تقدير الجزء غير المعلمي باستعمال شرائح التمهيد Spline Smoothing اولاً من ثم تقدير الجزء المعلمي باستعمال الطرائق التقليدية بعد ازالة تأثير غير المعلمي ومقارنتها مع تقدير إنموذج انحدار بواسون التقليدي مستخدماً معيار المقارنة متوسط قيم المطلقة للاخطاء النسبية (MAPE) لاختيار افضل نموذج من حيث الأداء والجودة.

2- توزيع بواسون Poisson Distribution [2]

يُعد توزيع بواسون من التوزيعات الاحتمالية المنفصلة ذات الأهمية الكبيرة في العديد من التطبيقات الإحصائية، لاسيما تلك المتعلقة بنمذجة عدد مرات حدوث ظاهرة نادرة ضمن فترة زمنية أو مكان معين، مثل حوادث الطائرات أو الزلازل. ويُعرف هذا التوزيع أحيانًا بـ"توزيع الأحداث النادرة".

تم تطوير هذا التوزيع على يد العالم الفيزيائي –والرياضي الفرنسي (Simeon-Denis poisson) (1781-1840), حيث تم اشتقاقه كتقريب لتوزيع ثنائي الحدين, ويُعرف التوزيع منذ ذلك الحين باسمه .

ويُشار إلى توزيع بواسون أيضًا باسم "قانون الأعداد الصغيرة (Law of Small Numbers) "، نظرًا لارتباطه بنمذجة الأحداث التي تحدث بمعدلات منخفضة ولكن محتملة الحدوث ضمن فترات زمنية محددة.

وان دالة الكثافة الاحتمالية لتوزيع بواسون تكون :

$$P(y) = \frac{\mu^y e^{-\mu}}{y!} \qquad y = 0,1,2,\dots \qquad \dots\dots(1)$$

حيث ان $\mu > 0$ تمثل معلمة التوزيع .

3- إنموذج انحدار بواسون Poison Regression Model [1]

إنموذج انحدار بواسون هو احد اهم النماذج الخطية اللوغارتيمية. ويُستخدم بشكل أساسي في نمذجة المتغير المعتمد عندما تكون قيمه على شكل قيم بهيئة قيم معدودة (Count Data) او بهيئة معدلات (Rate Data) . فإن الإنموذج يحتوي على متغيرات تفسيرية (توضيحية) كثيرة ما يؤثر سلباً على دقة الإنموذج وبساطته في تفسير النتائج وأن المتغير المعتمد yi يتبع توزيع بواسون وبمعلمة قدرها (μ) كما تتبع الأخطاء العشوائية في الإنموذج توزيع بواسون بمعلمة قدرها (μ) , ويعرف إنموذج انحدار بواسون بالمعادلة العامة :

$$\mathbf{Y_i = Exp(X\beta + \varepsilon)} \quad \dots\dots (2)$$

حيث ان

Yi :- موجه المتغير التابع ذي درجة $(\mathbf{n \times 1})$

X :- مصفوفة المتغيرات المستقلة ذات درجة $(\mathbf{n \times (p+1)})$

$\boldsymbol{\beta}$:- موجه المعلمات ذو درجة $(\mathbf{(p+1) \times 1})$

$\boldsymbol{\varepsilon}$:- موجه الأخطاء العشوائية ذي درجة $(\mathbf{n \times 1})$

n : حجم العينة ، P: عدد المتغيرات المستقلة

❖ فرضيات إنموذج انحدار بواسون [14][17]

يعتمد إنموذج انحدار بواسون على ثلاثة فرضيات اساسية :

1- المتغير المعتمد Y يتبع توزيع بواسون عند معلمة قدرها $\boldsymbol{\mu}$.

2- ان معلمة التوزيع للمتغير المعتمد y تكون متساوية الى :

$$\boldsymbol{\mu_i = e^{x_i'\underline{\beta}}} \quad \dots\dots (3)$$

3- إن أزواج المتغيرين $(\mathbf{y_i , x_i})$ تكون بينهما استقلالية , أن مع الاعتماد هذه الفرضيات الثلاثة بالاضافة الى خاصية توزيع بواسون [11] :

$$\mathbf{E(y_i|x) = var(y_i|x) = \mu_i = e^{x_i'\underline{\beta}}}$$

من أجل تقدير معلمات إنموذج انحدار بواسون يتم استخدام طريقة مقدرات الامكان الاعظم (Maximum Likelihood Method), وذلك بلاعتماد على الفرضيات الاساسية الثلاثة التي تم ذكرها سابقاً , ويُعد توزيع بواسون هو الأساس الذي ينبني عليه الإنموذج الخاص بالمتغير المعتمد (yi) , فتكون دالة التوزيع كما في المعادلة $(\mathbf{1})$ ومن خلال تعظيم المشاهدات لتوزيع المتغير المعتمد (yi) فتصبح دالة الامكان الاعظم كالاتي [1][13] :

$$\mathbf{L(y_1, y_2, \dots\dots, y_n; \mu_i)} = \frac{\mu_i^{\sum_{i=1}^{n} y_i}\ e^{-\sum_{i=1}^{n}\mu_i}}{\prod_{i=1}^{n} y_i!}$$

$$\mathbf{LogL(y_i/x_i, \beta)} = \sum_{i=1}^{n} \mathbf{y_i\left(Log(\mu_i)\right)} - \sum_{i=1}^{n} \boldsymbol{\mu_i} - \mathbf{Log}\prod_{i=1}^{n} \mathbf{y_i!}$$

وبالاعتماد على الافتراض الثاني من الفروض الثلاثة لإنموذج انحدار بواسون في المعادلة $(\mathbf{3})$ يتم تعويض في المعادلة $(\mathbf{6})$ وكما يلي :

$$\mathbf{LogL(y_i/x_i, \beta)} = \sum_{i=1}^{n} \mathbf{y_i}\left(\mathbf{Log}\left(\mathbf{e^{x_i'\underline{\beta}}}\right)\right) - \sum_{i=1}^{n} \mathbf{e^{x_i'\underline{\beta}}} - \mathbf{Log}\prod_{i=1}^{n} \mathbf{y_i!}$$

$$\frac{\partial \mathbf{LogL}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{y_i x_i} - \mathbf{x_i}\sum_{i=1}^{n} \mathbf{e^{x_i'\underline{\beta}}} = \sum_{i=1}^{n} \mathbf{(y_i - e^{x_i'\underline{\beta}})\, x_i}$$

$$\frac{\partial \mathbf{LogL}}{\partial \boldsymbol{\beta}} = \mathbf{0} \quad \rightarrow\rightarrow \sum_{i=1}^{n} \mathbf{(y_i - e^{x_i'\underline{\beta}})\, x_i = 0} \quad \dots\dots (4)$$

إذ ان المعادلة رقم $(\mathbf{4})$ هي غير خطية بالنسبة لموجه المقدرات $(\underline{\boldsymbol{\beta}}')$, ولحل هذه المشكلة تستعمل احدى طرائق التكرارية والمعروفة بخوارزمية المربعات الصغرى التكرارية الموزونة (ltreative weighted Least square) , اذ تكون مقدرات المعلمات لإنموذج انحدار بواسون :

$$\mathbf{\hat{\beta}_{MLE} = (X'\hat{W}X)^{-1}(X'\hat{W}\underline{Z})} \quad \dots\dots (5)$$

حيث ان

$\mathbf{\hat{\beta}_{MLE}}$: موجه معلمات إنموذج انحدار بواسون المقدرة

$\mathbf{\hat{W}}$: هي مصفوفة قطرية عناصر قطرها تساوي القيم المقدرة لمعلمة توزيع بواسون $\hat{\mu}_i$

$\underline{\boldsymbol{Z}}$: موجه الاستجابات الزائفة (pseudo-responses) لكل عنصر i وان العنصر i في الموجه $(\underline{\boldsymbol{Z}})$ مساوٍ الى

$$\mathbf{Z_i = Log(\hat{\mu}_i) + \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i}}$$

4- إنموذج انحدار بواسون شبه المعلمي للبيانات العنقودية
Semiparametric Poisson Regression Model for Clustered Data [4][6]

في دراسة إنموذج انحدار بواسون شبه المعلمي للبيانات العنقودية يفترض ان بيانات متغير الاستجابة $\mathbf{Y_i^k}$ ومجموعة من المتغيرات التوضيحية $\mathbf{X_{ij}^k}$ مؤلفة ب m من العناقيد فنفترض الإنموذج :

$$\mathbf{logE(Y_i^k/\mu_k) = \mu_k + f(X_{ij}^k)} \quad \dots\dots (6)$$

حيث ان

$$\mathbf{i = 1_k, 2_k, \dots, n_k \ , j = 1, 2, \dots, p \ , k = 1, 2, . . , n}$$

اذ ان

$\mathbf{n_k}$: حجم العنقود (يكون ثابت او غير ثابت عبر العناقيد)

$\mathbf{n}$: العدد الكلي للعناقيد

$\mathbf{\mu_k}$: المتغير العشوائي (الجزء المعلمي)

$$\mathbf{\mu_k = U_k + \varepsilon_k} \quad \begin{bmatrix} \boldsymbol{\varepsilon_k \sim N(0, \sigma_\varepsilon^2)} \\ \text{هو التاثير العشوائي للعنقود } \mathbf{U_k} \end{bmatrix}$$

$\mathbf{X_{ij}^k}$: قيمة المشاهدة عند $\mathbf{j^{th}}$ داخل العنقود $\mathbf{k^{th}}$

$\mathbf{f(X_{ij}^k)}$: دالة الممهد (دالة لامعلمية)

$\mathbf{Y_i^k}$: قيمة المتغير المعتمد عند $\mathbf{i^{th}}$ داخل العنقود $\mathbf{k^{th}}$

ان تأثيرات المتغيرات التوضيحية تكون متجانسة داخل كل عنقود , أي أن العلاقة بين المتغيرات المستقلة و التابعة لا تختلف بين المشاهدات ضمن العنقود الواحد. ومع ذلك, يمكن أن تختلف هذه التاثيرات من عنقود إلى اخر, مما يعكس خصوصية كل عنقود في طبيعة العلاقة الإحصائية.

في بعض النماذج, يُفترض أيضاً أن تكون الارتباط داخل كل عنقود متجانسة , أي ان الترابط بين المشاهدات داخل العنقود تتبع نمطاً ثابتاً , فأن تأثير $\mathbf{X_{ij}^k}$ على $\mathbf{Y_i^k}$ يتم افتراضه بشكل غير معلمي .

في هذا الإنموذج, يُفترض ان لكل عنقود k تاثيره الخاص $\mathbf{\mu_k}$ مما يعكس فرضية التجميع, حيث يُسمح باختلاف تأثير المتغيرات المشتركة من عنقود الى اخر .

5- طرائق التقدير Estimation Methods [7]

سيتم التطرق إلى أبرز طرائق التقدير في نماذج انحدار بواسون شبه المعلمية (SPR) عند التعامل مع البيانات العنقودية. وتتمثل هذه الطرائق فيما يلي :

(1-5) التقدير الشبه المعلمي Semiparametric Estimation

يتم في هذه الطريقة تقدير الاجزاء المعلمية وغير المعلمية في وقت واحد لذلك نستخدم طريقة المربعات الصغرى الجزائية Penalized Least Squares

يُعد أسلوب **المربعات الصغرى الجزائية (Penalized Least Squares)** [16] احد الاساليب الاحصائية المستخدمة لتقدير النماذج شبه المعلمية, حيث تعتمد على تقليل دالة الهدف هي عبارة عن مجموع مربعات البواقي (Residual Sum of Squares) مضافًا إليها حد الجزائي (penalty Term) يُفرض على درجة تعقيد الدالة الغير المعلمية. والهدف من هذا الحد الجزائي هو تجنب الإفراط في المطابقة (Overfitting) إذ تزداد قيمة الجزاء مع تعقيد الدالة وتنخفض مع زيادة التمهيد , ويُعبر عن دالة الهدف كما يلي :

$$\mathbf{PLS_\gamma = min \sum_{k=1}^{m} \sum_{i=1}^{m_k} [logY_i^k - \mu_k - f(X_{ij}^k)]^2 + \gamma J(f)} \quad \dots\dots (7)$$

حيث ان

الجزء الاول $\mathbf{min \sum_{k=1}^{n} \sum_{i=1}^{n_k} [logY_i^k - \mu_k - f(X_{ij}^k)]^2}$ يمثل تقليل مجموع مربعات البواقي (RSS)

$\boldsymbol{\gamma}$: تمثل معلمة الممهد smoothing parametrer

الجزء الثاني $\mathbf{J(f)}$ تمثل دالة الجزاء penalty function ويكون مقترن بالقيمة $\boldsymbol{\gamma}$.

ومن بين الطرائق المستخدمة لتمثيل الأجزاء غير المعلمية من الإنموذج، تُعد طريقة **تمهيد الصفائح الرقيقة (Thin Plate Spline)** [15] من أكثر الطرق مرونة. وتمتاز هذه الطريقة بإمكانية استخدامها في حالات متعددة الأبعاد، دون الحاجة إلى تحديد شكل الدالة غير الخطية مسبقًا. وتُبنى الدالة المقدّرة باستخدام دوال أساس (Basis Functions)، حيث يُمكن تمثيل العلاقة بين المتغيرات بالشكل التالي:

$$\mathbf{f(x_{ij}^k) = \sum_{j=0}^{p} \beta_j X_{ij}^k + \sum_{i=1}^{p} \alpha_j B_j(x_{ij}^k)} \quad \dots\dots (8)$$

اي ان

$\mathbf{X_{ij}^k}$: يمثل متجه المتغيرات التوضيحية

$\boldsymbol{\beta_j, \alpha_j}$: يمثلان المعالم التي يجب تقديرها

$\mathbf{B_j(x_{ij}^k)}$: تمثل دالة الاساس basis function وتسمى بعض الاحيان دالة الاساس الشعاعي radial basis function (RBF) والتمهيد باستخدام TPS يختلف عن الطرق الأخرى مثل الشرائح التكعيبية (Cubic Splines) ، حيث إن الجزاء المفروض لا يقتصر فقط على المشتقة الثانية للدالة، بل يمكن أن تُفرض على أي مرتبة من المشتقات التي تُستخدم عند التعامل مع بيانات متعددة الأبعاد .

في حالة البعدين مثلاً، يتم فرض الجزاء من خلال التكامل التالي:

$$\mathbf{J(f) = \iint \left[\left(\frac{\partial^2 f(x)}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial f(x)}{\partial x_1 x_2} \right)^2 + \left(\frac{\partial f(x)}{\partial x_2^2} \right)^2 \right] dx_1 dx_2}$$

حيث ان $\mathbf{(x_1, x_2)}$ احداثيات المتجه $\mathbf{x}$.

في هذا السياق، فإن استخدام تمهيد الصفائح الرقيقة داخل إنموذج GAM يُعتبر تطبيقًا عمليًا لطريقة المربعات الصغرى الجزائية ، حيث يتم تقدير الجزء غير المعلمي عن طريق تقليل دالة هدف تجمع بين ملاءمة البيانات goodness of fit والممهد smoothing .

وبما أن الدالة غير المعلمية $\mathbf{f(x_{ij}^k)}$ قد تم تمثيلها في المعادلة (8) فيمكن تعويضه في دالة الهدف المعطاة في المعادلة (7) لنحصل على المعادلة الموسعة الاتية :

$$\mathbf{PLS_\gamma = \min \sum_{k=1}^{m} \sum_{i=1}^{m_k} \left[\log Y_i^k - \mu_k - \sum_{j=0}^{p} \beta_j X_{ij}^k - \sum_{i=1}^{p} \alpha_j B_j(x_{ij}^k) \right]^2 + \gamma \alpha' R \alpha} \quad \dots\dots (9)$$

حيث ان

R: تمثل مصفوفة الجزاء Penalty Matrix التي تكون مربعة من الدرجة ى $\times$ $\mathbf{n}$ وتحسب كالاتي :

$$\mathbf{R = AM^{-1}A'} \quad \dots\dots (10)$$

إذ ان

A : تمثل مصفوفة ذات درجة $\mathbf{n \times (n-2)}$ وتحسب عناصرها كما يلي

$$\mathbf{, a_{(j+1),j} = c_j^{-1} , a_{j,j} = -c_{j-1}^{-1} - c_j^{-1} , a_{i,j} = 0 \quad \forall |i-j| \geq 2 \; a_{(j-1),j} = c_{j-1}^{-1}}$$
$$\mathbf{, i = 1, 2, 3, \dots, n , j = 2, 3, 4, \dots, n-1 \; A_{n,(n-2)} = [a_{i,j}]}$$

M : تمثل مصفوفة ذات درجة $\mathbf{(n-2) \times (n-2)}$ وتحسب عناصرها كما يلي

$$\mathbf{, m_{i,i} = \frac{c_{i-1}+c_i}{3} , m_{i,j} = 0 \quad \forall |i-j| \geq 2 m_{i,(i+1)} = m_{(i+1),i} = \frac{c_i}{6}}$$
$$\mathbf{, i = 2, 3, 4, \dots, n-1 , j = 2, 3, 4, \dots, n-1 M = [m_{i,j}]}$$

حيث ان $\mathbf{n = \sum_{k=1}^{m} m_k}$

c : هي الفرق بين المشاهدتين وتحسب كما يلي

$$\mathbf{c_i = x_{i+1} - x_i}$$

اعادة المعادلة (9)باستخدام المصفوفات

$$\mathbf{RSS_\gamma = (\log Y - \mu_k - \beta X - \alpha B)(\log Y - \mu_k - \beta X - \alpha B)' + \gamma \alpha' R \alpha}$$
$$\mathbf{RSS_\gamma = \log Y' \log Y - 2 \log Y' \mu_k - 2 \log Y' \beta X - 2 \log Y' \alpha B + \mu_k' \mu_k + 2 \mu_k' \beta X + 2 \mu_k' \alpha B + \beta' X' X \beta + 2 \beta' X' \alpha B + \alpha' B' B \alpha + \gamma \alpha' R \alpha}$$

ولأغراض التقدير، نأخذ الاشتقاق الجزئي للدالة في المعادلة اعلاه بالنسبة إلى كل من $\boldsymbol{\alpha , \beta , \mu_k}$ ، ثم نساوي كل منها بالصفر, فنحصل على التقديرات النهائية كالاتي:

اولاً: بالنسبة $\boldsymbol{\mu_k}$

$$\mathbf{\hat{\mu}_k = \log Y - \beta X - \alpha B} \quad \dots\dots (11)$$

ثانياً: بالنسبة $\boldsymbol{\beta}$

$$\mathbf{\hat{\beta} = (X'X)^{-1} X' Y^*} \quad \dots\dots (12)$$

حيث ان

$$\mathbf{Y^* = logY - \hat{\mu}_k - B\alpha}$$

ثالثاً: بالنسبة $\boldsymbol{\alpha}$

$$\mathbf{\hat{\alpha} = S_\gamma \tilde{Y}} \quad \ldots\ldots (13)$$

حيث ان

$$\mathbf{\tilde{Y} = logY - \hat{\mu}_k - X'\hat{\beta}}$$
$$\mathbf{S_\gamma = (B'B + \gamma R)^{-1}B'}$$

$\mathbf{S_\gamma}$: تمثل مصفوفة تمهيد Smoothing Matrix التي تكون معرفة موجبة ومربعة ومتماثلة من الدرجة $\mathbf{n \times n}$

(2-5) التقدير التكراري Backfitting [9][18]

تُعد طريقة Backfitting إحدى الخوارزميات التكرارية الفعّالة لتقدير النماذج الإحصائية شبه المعلمية، وخاصة ضمن إطار النماذج التجميعية (Additive Models) تقوم هذه الطريقة على مبدأ تفكيك الإنموذج إلى دوال جزئية (Partial Functions)، بحيث تمثل كل دالة التأثير غير الخطي لمتغير توضيحي واحد. يتم تقدير كل مكون على حدة، بينما يتم تثبيت المكونات الأخرى مؤقتًا، ويُعاد تكرار هذه العملية تدريجيًا حتى يتحقق التقارب العددي (Numerical Convergence) للإنموذج ككل.

تم تطوير خوارزمية Backfitting من قبل (Hastie & Tibshirani) كطريقة تسلسلية لتقدير الإنموذج التجميعي , بالاعتماد على مبدأ تقدير كل دالة جزائية بشكل منفصل باستخدام أدوات التمهيد الإحصائي. بينما اول من قدم هذه الطريقة هو (Friedman & Stuetze (1981) إذ تعتمد على الاسلوب التكراري .

في التطبيقات شبه المعلمية, يقدر اولاً الجزء غير المعلمي من الإنموذج باستخدام شريحة التمهيد (Spline Smoothing) , ثم يُستخدم القيم المتبقية بعد ازالة تاثير الجزء اللامعلمي $\mathbf{\hat{f}(X_{ij}^k)}$ من المتغير التابع $\mathbf{log\, Y_i^k}$ لتقدير الجزء المعلمي .

خطوات التقدير:

الخطوة الاولى : نفترض قيمة ابتدائية لـ $\boldsymbol{\mu_k}$

$$\mathbf{\mu_k = 0}$$

الخطوة الثانية : نقدر الجزء غير المعلمي باستخدام طريقة Spline Smoothing

تُعد طريقة **تمهيد الشرائح (Spline Smoothing)** [5][3] من الأساليب الشائعة في نماذج الانحدار غير الخطية، وتستخدم للحصول على منحنيات سلسة تقرّب العلاقة بين المتغيرات المدروسة. تعتمد هذه الطريقة على تقليل الخطأ الناتج عن تقدير العلاقة بين المتغير التابع والمستقل، مع الحفاظ على درجة مناسبة من التمهيد. وتُعد مناسبة خاصة عندما تكون العلاقة بين المتغيرات غير معروفة الشكل أو تحتوي على تغيّرات أو انحناءات يصعب على الطرق التقليدية تمثيلها بدقة.

الهدف من استخدام الشرائح الممهدة هو تحقيق توازن بين مطابقة الإنموذج للبيانات والتمهيد الكافي لتجنب الإفراط. وقد أظهرت دراسات مثل دراسة Welsh (2002) أن استخدام شرائح ممهدة عند وجود ارتب6اط بين المشاهدات يوفر نتائج أكثر كفاءة مقارنةً بحالة الاستقلال بين المشاهدات

عند تقدير الجزء غير المعلمي من الدالة f على فترة محددة [a,b]، يتم التعبير من خلال تكامل مربع مشتقة f من الرتبة h ، كما يلي :

$$\mathbf{\int_a^b \{f^h(X)\}^2 dX \qquad h \geq 1}$$

ويُعرف هذا المقدار بـ " دالة الجزاء (Penalty Function) "، والتي تُمثّل آلية للسيطرة على تعقيد المنحنى. الهدف منها هو الربط بين التقدير الثابت والمتغير (المرن) للدالة، مما يساعد على ضبط التغيرات الحادة أو غير المنتظمة في البيانات .

بناءً على ذلك، يمكن تعريف ممهد الشريحة $\mathbf{\hat{f}_\theta(X)}$ على أنه مقلل لمعيار المربعات الصغرى الجزائية (Penalized Least Squares) ، كما في المعادلة التالية :

$$\mathbf{\sum_{k=1}^{n}\sum_{i=1}^{n_k}\left[logY_i^k - f(X_{ij}^k)\right]^2 + \theta \int_a^b \{f^h(X)\}^2 dX} \quad \ldots\ldots (14)$$

حيث ان

الجزء الاول $\mathbf{\sum_{k=1}^{n}\sum_{i=1}^{n_k}\left[logY_i^k - f(X_{ij}^k)\right]^2}$ يمثل مجموع مربعات البواقي (RSS)

$\boldsymbol{\theta}$: تمثل معلمة الممهد smoothing parametrer وتكون $\mathbf{\theta \geq 1}$

الجزء الثاني $\mathbf{\int_a^b \{f^h(X)\}^2 dx}$ يمثل الجزاء الغير ممهد roughness penalty ويكون مرجحاً بالقيمة $\boldsymbol{\theta}$ والذي يتحكم في مدى سلاسة التقدير. عندما يكون هذا التكامل كبيرًا. وعند اختيار $\mathbf{h = 2}$ يعني المشتقة الثانية نحصل على شريحة تمهيدية تكعيبية Cubic smoothing spline التي تتمثل بالمعادلة التالية :

$$\mathbf{\sum_{k=1}^{n}\sum_{i=1}^{n_k}\left[logY_i^k - f(X_{ij}^k)\right]^2 + \theta \int_a^b \{f''(X)\}^2 dX} \quad \ldots\ldots (15)$$

ويُشترط حينها أن تكون الدالة f قابلة للاشتقاق مرتين , وأن تكون بالإمكان حساب تكامل مرع المشتقة الثانية.

ولتقدير الجزء اللامعلمي في الإنموذج (12) باستخدام ممهد الشريحة , يمكن كتابة المعادلة (15) بصيغة المصفوفات كما في المعادلة التالية:

$$(\mathbf{logY}-\mathbf{f(x)})'(\mathbf{logY}-\mathbf{f(x)})+\boldsymbol{\theta}\int_a^b\{\mathbf{f''(X)}\}^2\mathbf{dX}$$

وبالاعتماد على العلاقة الرياضية الآتية:

$$\int_a^b\{\mathbf{f''(X)}\}^2\mathbf{dX}=\mathbf{f'(X)Rf(X)}$$

يمكن تعويض التكامل بالعنصر المكافئ له من المعادلة اعلاه , لنحصل على الشكل التالي:

$$(\mathbf{logY}-\mathbf{f(x)})'(\mathbf{logY}-\mathbf{f(x)})+\boldsymbol{\theta}\mathbf{f'(X)Rf(X)}$$

ولتبسيط التعبير أعلاه ، نعيد كتابته كالتالي:

$$\mathbf{logY'logY}-\mathbf{logY'f(X)}-\mathbf{logYf'(X)}+\mathbf{f'(X)f(X)}+\boldsymbol{\theta}\mathbf{f'(X)Rf(X)}$$
$$\mathbf{logY'logY}-\mathbf{2logYf'(X)}+\mathbf{f'(X)f(X)}+\boldsymbol{\theta}\mathbf{f'(X)Rf(X)}$$

بعد اشتقاق المعادلة بالنسبة الى f ومساوتها بالصفر, نحصل على التقدير النهائي كالاتي :

$$\hat{\mathbf{f}}(\mathbf{X})=\mathbf{S}_{\boldsymbol{\theta}}\mathbf{z} \qquad \dots\dots(16)$$

حيث ان

$$\mathbf{z}=\mathbf{logY}$$
$$\mathbf{S}_{\boldsymbol{\theta}}=(\mathbf{I}+\boldsymbol{\theta}\mathbf{R})^{-1}$$

$\mathbf{S}_{\boldsymbol{\theta}}$: تمثل مصفوفة تمهيد Smoothing Matrix التي تكون معرفة موجبة ومربعة ومتماثلة من الدرجة $\mathbf{n}\times\mathbf{n}$

R :هي مصفوفة الجزاء Penalty Matrix ، كما ورد تعريفها سابقاً في المعادلة $(\mathbf{10})$.

الخطوة الثالثة : بعد الحصول على تقدير أولي للجزء غير المعلمي $\hat{\mathbf{f}}(\mathbf{X}_{\mathbf{ij}}^{\mathbf{k}})$ يتم الانتقال الى تقدير الجزء المعلمي $\boldsymbol{\mu}_{\mathbf{k}}$ وذلك من خلال القيم المتبقية الناتجة بعد ازالة تأثير الجزء غير المعلمي $\mathbf{f}(\mathbf{X}_{\mathbf{ij}}^{\mathbf{k}})$ من القيم اللوغارتيمية للمتغير التابع $\mathbf{logY}_{\mathbf{i}}^{\mathbf{k}}$

يتم اعتماد طريقة المربعات الصغرى الاعتيادية لتقدير $\boldsymbol{\mu}_{\mathbf{k}}$، حيث تُصاغ دالة الهدف كما يلي:

$$\mathbf{l}(\boldsymbol{\mu}_{\mathbf{k}},\mathbf{f})=\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n_k}}\left(\mathbf{logY}_{\mathbf{i}}^{\mathbf{k}}-\boldsymbol{\mu}_{\mathbf{k}}-\mathbf{f}(\mathbf{X}_{\mathbf{ij}}^{\mathbf{k}})\right)^2$$

لإيجاد القيمة المثلى لـ$\boldsymbol{\mu}_{\mathbf{k}}$، نقوم باشتقاق المعادلة اعلاه بالنسبة لـ$\boldsymbol{\mu}_{\mathbf{k}}$ وثم نساوي للصفر , فنحصل على التقدير النهائي كالاتي:

$$\hat{\boldsymbol{\mu}}_{\mathbf{k}}=\frac{\mathbf{1}}{\mathbf{n_k}}\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n_k}}\left(\mathbf{logY}_{\mathbf{i}}^{\mathbf{k}}-\hat{\mathbf{f}}(\mathbf{X}_{\mathbf{ij}}^{\mathbf{k}})\right) \qquad \dots\dots(17)$$

الخطوة الرابعة : نكرر الخطوات السابقات حتى يصل التقدير الى حالة التقارب العددي (Numerical Convergence).

6- أختيار معلمة التمهيد Smoothing Parameter

سيتم التطرق إلى أبرز طرائق اختيار معلمة التمهيد المستخدمة في تقديرنماذج انحدار بواسون شبه المعلمية (SPR) :

طريقة العبور الشرعي المعمم Generalized cross-validation [12][8]

تُستخدم هذه الطريقة لتقييم جودة النماذج الإحصائية، كما تُعد امتدادًا طبيعيًا للطريقة التقليدية ولكن بكفاءة عددية محسّنة، حيث لا تتطلب إعادة التقدير المتكرر للإنموذج.

تُستخدم طريقة GCV بشكل خاص لتحديد معلمة التمهيد (Smoothing Parameter) ، في سياق تمهيد الشرائح مثل شرائح الصفائح الرقيقة (Thin Plate Splines) وشرائح التمهيدية التكعيبية (Cubic Smoothing Splines)، وتقوم فكرة GCV على تقليل مجموع مربعات البواقي(RSS) ، مع مراعاة درجات الحرية المكافئة (Effective Degrees of Freedom) للإنموذج.

وتحسب دالة $\mathbf{GCV}$ بالشكل :

$$\mathbf{GCV}=\frac{\mathbf{RSS}}{\mathbf{N}\left(\frac{\mathbf{1}}{\mathbf{N}}\mathbf{tr}(\mathbf{I}-\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\theta}})\right)^2} \qquad \dots\dots(18)$$

حيث RSS كما في المعادلة $(\mathbf{7})$ بالنسبة لمعلمة $\boldsymbol{\gamma}$, اما معلمة التمهيد $\boldsymbol{\theta}$ في المعادلة $(\mathbf{14})$

ويتم اختيار معلمة التمهيد التي تحقق اقل قيمة لـ $\mathbf{GCV}$.

7- معيار المقارنة Comparison Standard [10]

هناك عدد من المعايير تستخدم لقياس الجودة ودقة التنبؤ وتقيس مقدار الكفاءة ولغرض ايجاد افضل طريقة لتقدير لإنموذج دالة الانحدار , سيتم استعمال معيار متوسط القيم المطلقة للأخطاء النسبية Mean absolute percentage error لأغراض المقارنة بين الطرائق المستخدمة للنماذج, يُعد هذا المقياس من المؤشرات النسبية الشائعة الاستخدام في تقييم دقة النماذج التنبؤية، ويُستخدم لمقارنة القيم الحقيقية بالقيم المتوقعة بشكل نسبي. يتم حسابه من خلال جمع القيم المطلقة للأخطاء النسبية لجميع المشاهدات، ثم تقسيم المجموع على العدد الكلي للمشاهدات. وتُفضل الطرق التي تحقق أقل قيمة لهذا المعيار، إذ تعني أنها تقدم توقعات أقرب إلى القيم الحقيقية, يُعطى هذا المقياس بالمعادلة التالية:

$$\mathbf{MAPE} = \frac{\sum_{k=1}^{n}\sum_{i=1}^{n_k}\left|\left(\frac{y_i^k - \hat{y}_i^k}{y_i^k}\right)\times 100\%\right|}{\sum_{k=1}^{n} n_k} \quad \ldots\ldots (19)$$

حيث ان

y_i^k : تمثل القيم الحقيقة للمشاهدة.

$\hat{y}_i^k$: تمثل القيم التقديرية .

8- المحاكاة Simulation

تم استخدام ثلاث قيم لعدد العناقيد k=3,5,10 اما حجوم العناقيد تم استخدام حالتين: الحالة الاولى يكون فيه الحجوم متساوية في كل عنقود وهي $m_k = 10, 30, 50$ اما الحالة الثانية يكون حجم كل عنقود مختلفة عن الاخر التي استخدام الحجوم الكلية $n = 50, 100, 250, 500$ ومن خلالها يكون حجوم مختلفة لكل عنقود . تم توليد التاثير العشوائي للعنقود من التوزيع الطبيعي بمتوسط 2.4 وانحراف معياري 1.6 اي ان التباين يساوي 2.7 من ثم تم تطبيقه على إنموذج انحدار بواسون شبه المعلمي المقدر باستخدام طريقة شبه المعلمي) وطريقة التقدير التكراري backfitting وإنموذج انحدار بواسون الذي تم تقديره بطريقة الإمكان الأعظم وباستخدام برنامج (R 4.4.2). تم إعتماد على معيار متوسط القيم المطلقة للأخطاء النسبية (MAPE) للمقارنة وذلك ، فكانت النتائج كما في الجداول ادناه:

جدول (1) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد متساوية

cluster	size cluster	Semiparametric Method	Backfitting Method	Poisson Regression
3	10	0.283368	**0.271307**	34.707091
	30	0.308919	**0.308251**	91.666821
	50	**0.297991**	0.627837	1.741868
5	10	0.148181	**0.146341**	1.012741
	30	0.371188	**0.365663**	17.886578
	50	0.476578	**0.465177**	5.143542
10	10	**0.39218**	0.402538	8.557992
	30	**0.21523**	0.317616	4.006657
	50	0.314508	0.300336	3.422101

نلاحظ من جدول رقم (1):

- أن طريقة التركيب التكراري Backfitting هي الافضل عند جميع حجوم العينة ولجميع عدد العناقيد , وذلك لكونها حققت اقل قيمة لـ MAPE , ماعدا حالتين تفوقت فيه طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE هي : الحالة الاولى عند عدد العناقيد (3) وحجم عينة (50) بلغت قيمة (0.297991) اما الحالة الثانية عند عدد العناقيد (10) وحجم عينة (30,10) بلغت قيمة (0.21523,0.39218).

جدول (2) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد غير متساوية

cluster	size total	Semiparametric Method	Backfitting Method	Poisson Regression
3	50	0.156201	**0.150553**	2.978731
	100	0.13901	**0.124727**	30.328207
	250	0.28823	**0.276133**	1.947009
	500	0.322785	**0.301227**	6.885615
5	50	0.347591	**0.295436**	3.18684
	100	**0.45979**	0.489878	1.11943
	250	0.326094	**0.289006**	3.580831
	500	0.144385	**0.138711**	2.250008
10	50	**0.276277**	0.31071	4.182757
	100	**0.224284**	0.289781	15.551363
	250	0.268773	**0.240498**	9.322468

	500	**0.311889**	0.332517	13.830567

نلاحظ من جدول رقم (2):

- أن طريقة التركيب التكراري Backfitting هي الافضل عند جميع حجوم العينة ولجميع عدد العناقيد , وذلك لكونها حققت اقل قيمة لـ MAPE , ماعدا حالتين تفوقت فيه طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE هي : الحالة الاولى عند عدد العناقيد (5) وحجم الكلي (100) بلغت قيمة (0.45979) اما الحالة الثانية عند عدد العناقيد (10) وحجم الكلي (500,100,50) بلغت قيمة (0.311889,0.224284,0.276277).

في حالة التباين يساوي 1 , فكانت النتائج كما في الجداول ادناه:

جدول (3) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد متساوية

cluster	size cluster	Semiparametric Method	Backfitting Method	Poisson Regression
3	10	0.212543	**0.207437**	1.452534
	30	0.28782	**0.287435**	0.380391
	50	**0.437662**	0.449269	2.162134
5	10	0.275817	**0.267483**	1.278629
	30	0.267612	**0.233028**	2.04971
	50	0.314645	**0.302838**	1.220129
10	10	0.340651	**0.329738**	1.305017
	30	0.354659	**0.33623**	0.949055
	50	**0.272196**	0.277273	1.717384

نلاحظ من جدول رقم (3):

- أن طريقة التركيب التكراري Backfitting هي الافضل عند جميع حجوم العينة ولجميع عدد العناقيد , وذلك لكونها حققت اقل قيمة لـ MAPE , ماعدا حالتين تفوقت فيه طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE هي : الحالة الاولى عند عدد العناقيد (3) وحجم عينة (50) بلغت قيمة (0.437662) اما الحالة الثانية عند عدد العناقيد (10) وحجم عينة (50) بلغت قيمة (0.272196).

جدول (4) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد غير متساوية

cluster	size total	Semiparametric Method	Backfitting Method	Poisson Regression
3	50	0.235942	**0.016457**	1.368023
	100	0.354107	**0.140411**	1.744408
	250	0.270248	**0.250105**	1.201511
	500	0.412555	**0.385883**	0.830097
5	50	0.331159	**0.037569**	2.576432
	100	0.185223	**0.174531**	0.484586
	250	0.209196	**0.202240**	1.010680
	500	0.349797	**0.316532**	1.962625
10	50	0.214757	**0.203104**	0.587716
	100	**0.254510**	0.2660004	1.268336
	250	**0.246382**	0.294581	2.121225
	500	0.330058	**0.306567**	1.442907

نلاحظ من جدول رقم (4):

1- أن طريقة التركيب التكراري Backfitting هي الافضل عند جميع حجوم العينة ولجميع عدد العناقيد , وذلك لكونها حققت اقل قيمة لـ MAPE , ماعدا عند عدد العناقيد (10) وحجم عينة (100 , 250) تفوقت فيه طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE التي بلغت قيمة (0.246382 , 0.254510).

في حالة التباين يساوي 4 , فكانت النتائج كما في الجداول ادناه:

جدول (5) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد متساوية

cluster	size cluster	Semiparametric Method	Backfitting Method	Poisson Regression
3	10	0.274801	**0.262185**	11.728176
	30	0.313905	**0.307792**	0.722782
	50	**0.234292**	0.586682	6.380013
5	10	0.252883	**0.232853**	10.154175
	30	0.279262	**0.275899**	2.394799
	50	**0.389723**	0.498907	0.844164
10	10	**0.281249**	0.374222	17.046002
	30	**0.363456**	0.428621	2.106941
	50	**0.280456**	0.353912	28.146851

نلاحظ من جدول رقم (5):

1- بالنسبة لعدد العناقيد k=3
- عند حجم العينة (30,10) كانت طريقة التقدير التكراري Backfitting هي الافضل لامتلاكها اقل قيمة لـ MAPE التي بلغت (0.262185 , 0.307792) .
- اما عند حجم العينة (50) فقد تفوقت طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE وهي (0.234292)

2- بالنسبة لعدد العناقيد k=5
- عند حجم العينة (30,10) كانت طريقة التقدير التكراري Backfitting هي الافضل لامتلاكها اقل قيمة لـ MAPE التي بلغت (0.232853 , 0.275899) .
- اما عند حجم العينة (50) فقد تفوقت طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE وهي (0.389723)

3- بالنسبة لعدد العناقيد k=10
- أن طريقة شبه المعلمية Semiparametric هي الافضل عند جميع حجوم العينة, وذلك لكونها حققت اقل قيمة لـ MAPE .

جدول (6) قيم متوسط القيم المطلقة للأخطاء النسبية (MAPE) عند حجوم العناقيد غير متساوية

cluster	size cluster	Semiparametric Method	Backfitting Method	Poisson Regression
3	50	0.152153	0.148963	4.898733
	100	**0.129330**	0.188996	56.946026
	250	0.287600	**0.271056**	3.071101
	500	0.361803	**0.32917**	14.764998
5	50	0.162246	**0.151735**	3.278847
	100	0.489153	**0.478884**	4.169948
	250	**0.267898**	0.288128	3.002359
	500	0.269549	**0.267293**	1.537668
10	50	**0.275939**	0.334099	4.697885
	100	**0.235451**	0.256663	3.197698
	250	**0.247896**	0.342792	12.222509
	500	**0.232749**	0.249558	4.313855

نلاحظ من جدول رقم (6):

1- بالنسبة لعدد العناقيد k=3
- عند حجم العينة الكلي (500,250,50) كانت طريقة التقدير التكراري Backfitting هي الافضل, وذلك لكونها حققت اقل قيمة لـ MAPE التي بلغت (0.32917,0.271056,0.148963).
- عند حجم العينة الكلي (100) فقد تفوقت طريقة شبه المعلمية Semiparametric لامتلاكها اقل لـ MAPE وهي (0.129330)

2- بالنسبة لعدد العناقيد k=5
- عند حجم العينة الكلي (500,100,50) فقد تفوقت طريقة التقدير التكراري Backfitting لامتلاكها اقل لـ MAPE وهي (0.151735 , 0.478884 , 0.267293).
- عند حجم العينة الكلي (250) كانت طريقة شبه المعلمية Semiparametric هي الافضل, وذلك لكونها حققت اقل قيمة لـ MAPE التي بلغت (0.267898)

3- بالنسبة لعدد العناقيد k=10
- أن طريقة شبه المعلمية Semiparametric هي الافضل عند جميع حجوم العينة , وذلك لكونها حققت اقل قيمة لـ MAPE .

9- الجانب التطبيقي

بناءً على زيارة وزارة الصحة والدوائر المرتبطة بها, تم جمع البيانات اللازمة للدراسة من قسم الكلى والتي تخص مرض الفشل الكلوي المزمن . شملت العينة الكلية (100) مشاهدة موزعة على ثلاث دوائر (دائرة صحة بغداد- الرصافة البالغ عددها (37) مشاهدة ، دائرة صحة بغداد- الكرخ البالغ عددها (36) مشاهدة , مركز تخصصي في امراض الكلى (مركز ابن سينا لغسيل الكلى) البالغ عددها (27) مشاهدة) إذ ان كل دائرة تمثل مجموعة (عنقود) , وقد تم تهيئة استمارة معلومات معدة سابقاً لغرض تسجيل البيانات من طبلات المرضى وباشراف اطباء وممرضين مختصين في هذا المجال. اذ سيتم دراسة توزيع المتغير المعتمد (عدد مرات الغسيل الكلوي بالشهر) مع سبعة متغيرات مستقلة, وتعرف المتغيرات كما يلي :

المتغيرات	الوصف	المستوى الطبيعي	وحدة القياس
Urea	مؤشر يقيس مستوى نتروجين اليوريا	6-21	مليغرام/ديسيلتر
Creatinine	مؤشر يقيس نسبة الكرياتين في الدم	0.6-1.3	مليغرام/ديسيلتر
Alburiben	مؤشر يقيس نسبة الألبومين في الدم	3.4-5.4	غرام/ديسيلتر
Protein	مؤشر يقيس نسبة البروتين في الدم	6-8.3	غرام/ديسيلتر
Calcium	مؤشر يقيس نسبة الكالسيوم في الدم	8.5-10.5	مليغرام/ديسيلتر
Bluriben	مؤشر يقيس نسبة البلروبين في الدم	0.3-1.2	مليغرام/ديسيلتر
age	العمر	—	سنة

واظهر ان المتغير bluriben هو المتغير غير المعلمي وتم اختبار المتغير المعتمد (Response Variable) لمعرفة المتغير يتبع اي توزيع عن طريق برنامج (EasyFit) وهو برنامج إحصائي يُستخدم لتحليل البيانات واختبار مدى توافق مجموعة بيانات مع التوزيعات الاحتمالية , ويسمح البرنامج بتجربة العديد من التوزيعات الاحتمالية على البيانات لتحديد أي توزيع يطابق البيانات بشكل أفضل, كما في الجدول (7):

جدول (7) يبين نتائج اختبار توزيع بيانات المتغير المعتمد

#	Distribution	Kolmogorov Smirnov		Anderson Darling	
		Statistic	Rank	Statistic	Rank
1	D. Uniform	0.33333	2	27.952	5
2	Geometric	0.50173	4	23.199	3
3	Logarithmic	0.50753	5	26.914	4
4	Neg. Binomial	0.33996	3	10.405	2
5	Poisson	0.30844	1	8.5421	1
6	Bernoulli	No fit (data max > 1)			
7	Binomial	No fit			
8	Hypergeometric	No fit			

اظهرت النتائج في الجدول (7) توافق نتائج اختبار حسن المطابقة ان المتغير المعتمد (Y) يتبع توزيع بواسون حسب اختبار (Kolmogorov Smirnov) واختبار (Anderson Darling) .

وبعد أن أظهرت نتائج المحاكاة ان افضل طريقة للتقدير هي (Backfitting , Semiparametric) لإنموذج انحدار بواسون شبه المعلمي فيتم تطبيق الطريقتين واستخدام نفس البرنامج على البيانات الحقيقية فنحصل على النتائج كما في الجداول ادناه :

اولاً: تقدير لانحدار بواسون شبه المعلمي باستخدام طريقة Penalized Least Squares

الجدول (8) نتائج تقدير إنموذج انحدار بواسون شبه المعلمي باستخدام طريقة Penalized Least Squares

Variable	EDF	P_value	MAPE
Intercept	1.15699123	**0.01772476**	0.603039
age	0.00290529	0.48719208	
urea	-0.0039537	**0.00953117**	
ceratinine	0.06190974	0.11222511	
albumin	-0.1146349	0.17393967	
protien	0.05981648	0.4513519	
calcium	-0.0004565	0.78219158	

bluriben	1	0.33568151	

اظهرت نتائج الجدول (8) ان المتغير (urea) التي له تأثير على المتغير المعتمد لكونه متغير معنوي التي بلغت قيمة p_value تساوي (0.00916629) اي اقل من (0.05) ويؤثر بشكل طردي والثابت له تأثير ايضاً.

ثانياً: تقدير لانحدار بواسون شبه المعلمي باستخدام طريقة Backfitting

الجدول (9) نتائج تقدير إنموذج انحدار بواسون شبه المعلمي باستخدام طريقة Backfitting

Variable	EDF	P_value	MAPE
Intercept	1.15699113	**0.01228581**	0.500215
age	0.00290529	0.46948428	
urea	-0.0039537	**0.00604132**	
ceratinine	0.06190973	0.0963723	
albumin	-0.1146349	0.1553704	
protien	0.05981653	0.43288169	
calcium	-0.0004565	0.77373312	
bluriben	1.00001641	0.31560528	

اظهرت نتائج الجدول (9) ان المتغير (urea) له تأثير على المتغير المعتمد لكونه متغير معنوي التي بلغت قيمة p_value تساوي (0.00604132) اي اقل من (0.05) ويؤثر بشكل طردي والثابت له تأثير ايضاً..

بناءً على تحليل نتائج الجداول الثلاثة(8),(9), نجد أن طريقة Backfitting حققت أقل قيمة لـ(MAPE) التي بلغت (0.500215) وهذا يدل على ان هذه الطريقة توفر أفضل أداء , وبالتالي تعتبر الطريقة الأفضل بين الطريقتين .

وكانت قيم التأثير العشوائي للعناقيد كما في الجدول (10)

الجدول (10) قيم التأثير العشوائي الخاص بالعنقود

k	U_k
1	-0.2085941
2	-0.1564721
3	-0.1179184

10- الاستنتاجات

بناءً على ما جاء في البحث فقد تم توصل الى الاستنتاجات الاتية:

1- ان طريقة التقدير باستخدام Backfitting هي الافضل عند إنموذج انحدار بواسون شبه المعلمي للبيانات العد العنقودي بشكل اكبر .
2- طريقة شبه المعلمية Semiparametric كانت الافضل عند الحجوم العينات الكبيرة .
3- اظهرت النتائج ان طريقة انحدار بواسون التقليدي Poisson Regression هي لم تحقق نتائج جيدة .
4- ان اكثر قيم MAPE تزداد مع زيادة حجم العينة لكل عنقود عند الحجوم المتساوية .

11- الاستنتاجات

ندرج اهم التوصيات الاتية:

1- اعتماد انحدار بواسون شبه المعلمي عندما تكون البيانات بيانات عد العنقودية .
2- استخدام طرائق تقدير اخرى غير الطرائق المستخدمة في البحث مثل تقدير الجزء المعلمي باستعمال Maximize Likelihood اما الجزء غير المعلمي تقديره باستعمال Local Polynomial Regression او Nadaraya – Watson Kernel Estimator .
3- نوصي بالدراسات مستقبلية : مقارنة بين انموذج انحدار بواسون شبه المعلمي بوجود التاثيرات العشوائية وانحدار بواسون اللامعلمي للبيانات العنقودية .

المصادر

1- العوادي, ايثار حسين, (2017), " مقارنة بعض طرائق تقدير معلمات إنموذج بواسون الهرمي الجزئي مع تطبيق عملي " رسالة ماجستير, كلية الادارة والاقتصاد – جامعة بغداد .
2- صبري , حسام موفق, (2013), " مقارنة طرائق تقدير معلمات إنموذج انحدار بواسون في ظل وجود مشكلة التعدد الخطي" أطروحة دكتوراه , كلية الادارة والاقتصاد – جامعة بغداد .
3- علي , عمر عبدالحسين (2007) " مقارنة مقدرات النماذج التجميعية المعممة باستخدام الشرائح التمهيدية عند تحليل الانحدار اللامعلمي وشبه المعلمي " اطروحة دكتوراة في الاحصاء, كلية الادارة والاقتصاد , جامعة بغداد
4- Arceneaux,K. and Nickerson,D. (2007), " Modeling certainty with clustered data: A comparison of methods ", Political Analysis 17:177-190 .
5- Aydin, D.(2007), " Estimation of GDP in Turkey by nonparametric regression models , Anadolu University-Eskisehir | TURKEY .
6- Demidnko,E. (2007) "Poisson regression for clustered data " International Statistical Review 75:96-113 .
7- Ernel B,Barrios anf Kristina Celene M.Manalaysay, (2011)," semiparametric Principal Components Poisson regression on Clustered data " , School of Statistics , University of the Philippines Diliman, UPSS Working Paper No.2011-06 .
8- Fattahi, SH. ,(2011), " A Comparative Study of Parametric and Nonparametric Regression ", Iranian Economic Review, Vol.16, No.30.
9- Hastie,T. and Pregibon,D. (1996) " A New Algorithm for Matched Case – Control Studies with Application to Additive Models " Technical Report, Murray Hill, NJ.
10- Hyndman, R.J. & Koehler, A.B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.
11- Long, J.S.(1997), " Regression Models for Categorical and Limited dependent Variables ", SAGE Publicayion Inc, USA .
12- Maharani,M. and Saputro,D.R.S. (2021) " Generalized Cross validation (GCV) in Smoothing Spline Nonparametric Regression Models ", Journal of Physics: Conference Series .
13- Mansson , K. , Kebria, B.M. , Sjolander, P. , Shukur, G (2012) " New Liu Estimates for the Poisson Regression Model " Methods and App ; ication .
14- Mansson , K. , Shukur, G. ,(2011) " A poisson Ridge Regression Estimator" , Economic Modeling, Vol. 28, Issue 4,pp. 1475-1491 .
15- Rossiter, D.G, (2016) " Empirical interpolation: thin plate splines" Cornell University .
16- Wahba,G. (1990) " Spline Models for Observational Data" University of Wisconsin at Madison .
17- Winkelmann, R. (2008) " Economic Analysis of Count Data " , 5 th ed., Springer , Verlag Berlin Heidelberg Germany .
18- Wood,S.N. (2006), " Generalized Additive Models "an introduction with R. CRC Press .