Journal of Economics and Administrative Sciences Vol.26 (NO. 119) 2020, pp. 480-493

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

مقارنة بين الانحدار الشجيري وانحدار ذي الحدين السالب باستعمال المحاكاة

م.د. أسماء نجم عبد الله
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الإحصاء
موبايل: 07700300417
الايميل:
asmaanajm92@gmail.com

الباحث/ براء خضير عباس
جامعة بغداد /كلية الإدارة والاقتصاد/ قسم الإحصاء
موبايل: 07705885725
الايميل:
bb2769250@gmail.com

Received :30/10/2019 Accepted :7/1/2020 Published :April / 2020

مستخلص البحث

في هذا البحث نحن بصدد المقارنة بين أنموذج الانحدار الشجيري وانحدار ذي الحدين السالب, حيث شملت هذه النماذج نوعين من الأساليب الإحصائية تمثلت بالنوع الأول "الإحصاء اللامعلمي" وهو الانحدار الشجيري الذي يهدف إلى تقسيم مجموعة البيانات إلى مجاميع فرعية.

أما النوع الثاني "الإحصاء ألمعلمي" وتمثل بأنموذج انحدار ثنائي الحدين السالب والذي عادة ما يستخدم عند التعامل مع البيانات الطبية وخصوصا عند التعامل مع أحجام عينات كبيرة. حيث تمت المقارنة بين هذه الطرق وفق معيار متوسط مربعات الخطأ (MSE) وباستخدام المحاكاة للتجربة وبأخذ أحجام عينات مختلفة حيث أظهرت نتائج المحاكاة بان الانحدار الشجيري هو الأفضل عندما تكون قيمة التباين كبيرة (5) ولجميع أحجام العينات بينما عند اخذ قيم تباين صغيرة ومتوسطة نلاحظ تفوق أنموذج انحدار ثنائي الحدين السالب عند قيم التباين (1 , 0.5 0.01) ولجميع أحجام العينات.

المصطلحات الرئيسة للبحث/ الانحدار الشجيري, انحدار ذي الحدين السالب , طريقة الإمكان الأعظم, المحاكاة.

البحث مستل من رسالة ماجستير

المبحث الأول/ المقدمة العامة

1-1 المقدمة: Introduction

تكمن فلسفة الإحصاء من حيث محاولة إيجاد نماذج (Estimator models) للظواهر المختلفة بحيث تكون قريبة إلى الواقع الفعلي, و الانحدار الخطي وهو احد النماذج الخطية الذي يستخدم في مجالات واسعة حيث يكثر استعماله في تحليل بيانات العديد من البحوث الاقتصادية والبحوث الطبية والعلوم التطبيقية الأخرى, ونحن في هذا البحث بصدد المقارنة بين أنموذجين من نماذج الانحدار لتقدير المعالم وهما الانحدار ذي الحدين السالب (NBR) والانحدار الشجيري (RT) .

2-1 مشكلة البحث: Problem of the Research

تتمثل مشكلة البحث بأنه كثيرا ما يظهر في بيانات المتغير المعتمد Y بأنه يأخذ أكثر من توزيع أي يبتعد عن التوزيع الطبيعي فنلاحظ انه مصفوفة البيانات تارة تأخذ الانحدار ذي الحدين السالب وتارة أخرى الانحدار الشجيري لذا نحن في إطار المقارنة بين الأنموذجين من خلال معيار المقارنة MSE .

3-1 هدف البحث: Purpose of the Research

إن الهدف من البحث هو المقارنة بين نماذج الانحدار ألا وهي: الانحدار الشجيري وانحدار ذي الحدين السالب في تقدير معالم الأنموذج ومن خلال استعمال المحاكاة وبأخذ أحجام عينات مختلفة .

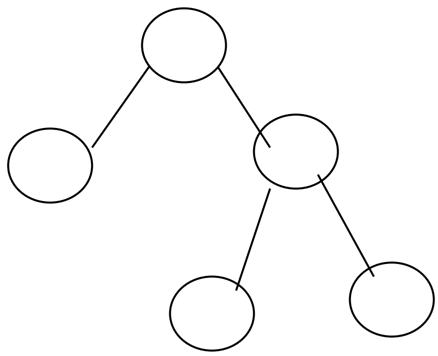
المبحث الثاني/ الجانب النظري

1-2 المقدمة:

في هذا المبحث سيتم عرض ودراسة أنموذج الانحدار الشجيري تقنية CART (RT) وأنموذج انحدار ثنائي الحدين السالب(NBR) وطريقة تقديره باستخدام طريقة الإمكان الأعظم (MLE).

2-2 الانحدار الشجيري: Tree regression

يعد الانحدار الشجيري من الطرق اللامعلمية التي ظهرت في نهاية الستينات من قبل Breiman and Others حيث كان هنالك بعض التحفظات حولها لحداثتها [4], ويوجد لها العديد من التقنيات ومن ضمنها تقنية Cart التي تم تصميمها لتمثيلها قواعد القرار في ما يسمى "الأشجار الثنائية" (Binary Trees) أي أنها أشجار ثنائية التقسيم حيث تكون مجموعة البيانات الناتجة على شكل هرمي تبدأ بعقدة الجذر الكامل وتنتهي بمجموعات صغيرة متجانسة من المشاهدات[2].كما في شكل رقم (1).



الشكل (1) يمثل شكل التصنيف الشجري

ومن خلال الرسم يتضح بأنه لدينا ثلاث مستويات هي العقد , تمثل المستوى الأول حيث العقدة في أعلى الرسم والتي تمثل عقدة الجذر. بينما المستوى الثاني تمثل بالعقدة الداخلية أما المستوى الثالث يتمثل بالعقد الطرفية والتي تتمثل عقد أيسر وأيمن الشجرة[7].

الانحدار الشجري يحتاج إلى عدد كبير من البيانات حيث تكون متغيرات كمية (مستمرة أو متقطعة) أو قد تكون متغيرات وصفية (ترتيبي أو اسمي). حيث يفترض أن البيانات تتمثل بمتغير الاستجابة Y مع متجه من المتغيرات التنبؤية والتي تكون على شكل مصفوفة ثابتة M.

$$X_i = (X_1, X_2, \cdots\cdots, X_m)$$

ويجب العمل بالخطوات الآتية عند كل عقدة (Node) وهي:

1- إيجاد جميع التقسيمات التي يمكن إيجادها للمتغيرات التنبؤية, وفي الغالب التقسيمات الثنائية تولد أسئلة ثنائية.

من النموذج ($X_i > C$) لكل قيم C التي هي تكون ضمن مجال لــ X_i أي أن X_i تأخذ أعداد محدودة .

$$\left(b_1\ \ b_2\ \ b_3\ \ \ldots\ \ b_i\right)$$

ونستطيع السؤال هنا; هل إن ($Xm \in C$) حيث يتراوح C ضمن للمجموعات الفرعية [$b_1, b_2, \ldots\ldots b_i$] وبتلك الحالات في الشجرة T عند الإجابة ب(نعم) ننتقل إلى العقدة اليسار, أما إذا كانت (لا) نتجه إلى يمين العقدة.

2- اعتماد مفهوم (حسن المطابقة) أي اختيار أفضل تقسيم وذلك من خلال معيار مطلق التباين الأصغر أو المربعات الصغرى.

3- إيقاف الانقسام على العقدة, التي لا تتوفر فيها الشروط المطلوبة.

عند نمو الشجرة بشكل كبير جدا (النمو المفرط) وذلك من خلال زيادة تفرعات الشجرة إذ يلاحظ وجود عدد قليل من البيانات عند كل عقدة طرفية مما يستدعي التوقف, ولذلك يتم تقليمها بشكل متكرر.

من اجل تنفيذ خوارزمية عند كل عقدة تبدأ من X_1 إلى X_m ولجميع المتغيرات الواحد بعد الأخر ثم يقارن مع M لاختيار أفضل تقسيم للمتغير, يتم تطبيق كل من الخطوتين الأولى والثانية على كل عقدة (الأبناء) حتى الحصول على شجرة كاملة[9] ,وبذلك يمكن التعبير عن الأنموذج الأساسي لانحدار الشجرة كالأتي[8]:

$$y = F\left(X_1, X_2, \cdots\cdots, X_p\right) + \varepsilon_i \qquad , \varepsilon_i \sim (0, \sigma^2) \quad \cdots\cdots (1-2)$$

حيث أن:

y : تمثل متغير الاستجابة

X_p: المتغيرات التوضيحية للأنموذج.

ε_i:الخطأ العشوائي.

أما الأنموذج المقدر فيكون [1]:

$$f^{\wedge} = \sum_{Nm}^{n} CmI[(X_1, X_2) \in Rm] \quad \cdots\cdots (2-2)$$

3-2: توزيع ثنائي الحدين السالب

ويسمى بتوزيع باسكال وهو من التوزيعات المتقطعة , حيث يفترض هذا التوزيع أن التجربة لها نتيجتان فقط هما النجاح والفشل ويتمثل p بحالة النجاح و (1-p) الفشل وهو سلسلة من محاولات برنولي للحصول على (r) من حالات النجاح ,للدلالة على (x) من حالات الفشل بمتوسط ($\mu = rq/p$) وتباين, ($\sigma^2 = rq/p^2$) [7],وان دالة الكثافة الاحتمالية للتوزيع هي[2]:

$$Pr(X = x) = \begin{cases} C_x^{x+r-1} p^r q^x & x = 0, 1, 2, \ldots \\ 0 & otherwise \end{cases} \qquad (3-2)$$

كما يعد هذا التوزيع خليط من توزيعين بواسون –كاما وقد تم اشتقاقه من قبل الباحثان (Green wood & Yule) , وتم تطوير هذا التوزيع من قبل الباحث (lordetal 2005) , والذي بدوره يؤدي إلى توزيع ثنائي الحدين ,لنفترض أن لدينا سلسلة من القيم العشوائية التي تتبع توزيع بواسون[6]:

$$g(y_i; \lambda_i) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \qquad \cdots\cdots\cdots\cdots (4-2)$$

أما توزيع كاما فتتمثل دالة الكثافة الاحتمالية لها كالأتي:

$$h(v_i, p, \delta) = \frac{\delta^\theta}{\Gamma\theta} v_i^{p-1} e^{-v_i \delta} \quad \cdots\cdots\cdots\cdots (5-2)$$

وبدمج كلا التوزيعين واخذ التكامل لهما نحصل على :

$$f(y_i,\mu_i,p)=\frac{\left(\frac{p}{\mu_i}\right)^p\left(1+\frac{p}{\mu_i}\right)^{-(p+y_i)}\Gamma(p+y_i)}{\Gamma p\,\Gamma(y_i+1)}\quad\cdots\cdots\cdots\cdots\cdots(6-2)$$

وبعد إجراء تبسيط على الصيغة السابقة تم التوصل إلى دالة الكثافة الاحتمالية لنموذج ثنائي الحدين السالب لنموذج وكالاتي:

$$f(y_i,\mu_i,p)=\frac{\Gamma(y_i+p)}{\Gamma(y_i+1)\Gamma p}\left(\frac{p}{\mu_i+p}\right)^p\left(\frac{\mu_i}{\mu_i+p}\right)^{y_i}\quad\cdots\cdots\cdots\cdots(7-2)$$

وبالمحصلة النهائية نستطيع تعريف دالة الكثافة الاحتمالية لتوزيع ثنائي الحدين السالب بالصيغة الآتية[5]:

$$f(y_i,\mu_i,p)=C_{p-1}^{y_i+p-1}\left(\frac{p}{\mu_i+p}\right)^p\left(\frac{\mu_i}{\mu_i+p}\right)^{y_i}\quad\cdots\cdots\cdots\cdots(8-2)$$

وتم الحصول على تقدير المعالم β عادة بواسطة طريقة الإمكان الأعظم

4-2 طريقة الإمكان الأعظم(5):

والتي يمكن الحصول عليها عن طريق اخذ مضاعف لانكرانج للمعادلة (2-7) وكالاتي:

$$\ln L(p,\beta)=\sum_{i=1}^{n}\left\{\left[\sum_{j=0}^{y-1}ln(j+p\mu_i)\right]-lny_i-(y_i-p\mu_i)ln(1+\mathrm{p}^{-1})+y_i lnp^{-1}\right\}\quad\cdots\cdots\cdots\cdots\cdots(9-2)$$

وبأخذ المشتقة الجزئية بالنسبة للمعلمة β ومساواتها للصفر و تطبيق خوارزمية (IRLS) المسماة بالخوارزمية التكرارية لإيجاد مصفوفة المعالم المقدرة وفق انحدار ثنائي الحدين السالب نحصل على:

$$I^{-1}\left(\beta^{(r-1)}\right)=E\left[\frac{d^2 l(x,\beta)}{d\beta d\beta'}\right]=\mathrm{X}'W\left(\beta^{(r-1)}\right)X\quad\cdots\cdots\cdots\cdots(10-2)$$

$$W\left(\beta^{(r-1)}\right)=diag\left[\frac{\mu_i\left(\beta^{(r-1)}\right)}{1+\theta\mu_i(\beta^{(r-1)})}\right]\quad\cdots\cdots\cdots\cdots\cdots(11-2)$$

وفي نهاية الخوارزمية حصلنا على المعادلة التي تمكننا من تقدير معلمات أنموذج ثنائي الحدين السالب وفق طريقة الإمكان الأعظم (MLE) وهي:

$$\hat{\beta}_{\mathrm{MLE}}=\left(X'\widehat{W}X\right)^{-1}\left(\mathrm{X}'\widehat{\mathrm{W}}\hat{\mathrm{Z}}\right)\quad\cdots\cdots\cdots\cdots\cdots\cdots\cdots(13-2)$$

$\hat{\beta}_{\mathrm{MLE}}$: موجه المعالم بطريقة الإمكان الأعظم وفق أنموذج ثنائي الحدين السالب.

$\widehat{W}$: مصفوفة القيم المقدرة وفق توزيع ثنائي الحدين السالب والممثلة بعناصر القطر.

$\hat{\mathrm{z}}$: هو متجه i^{th} لكل الأعداد هي $\frac{y_i+\hat{\mu}_i}{\hat{\mu}_i}+log(\hat{\mu}_i)$.

المبحث الثالث/ الجانب التجريبي

1-3 المقدمة:

في هذا المبحث تم التطرق إلى استخدام أسلوب المحاكاة وذلك من اجل المقارنة بين (انحدار توزيع ثنائي الحدين السالب و الانحدار الشجيري) في تقدير معالم الأنموذج , إذ سنوضح مفهوم المحاكاة التي تستخدم لوصف التجربة من خلال توليد أحجام عينات كبيرة . وقد تمت المقارنة وفق معيار متوسط مربعات الخطأ (MSE) و للنتائج المعروضة والتي تم الحصول عليها من تجربة المحاكاة.

2-3 وصف المحاكـــاة: General Understanding of Simulation

لغرض المقارنة بين نماذج الانحدار المستخدمة في تقدير المعالم تمت المحاكاة باستخدام لغة (R) إذ تم توليد البيانات بأحجام مختلفة (50,100, 150 ,200) .

المرحلة الأولى: تم توليد القيم الافتراضية باستخدام طريقة المربعات الصغرى الاعتيادية (149.09 , 0.183 , -0.392 , -0.132 , 5.209 , 7.733 , 8.895 , -1.634 , -4.071 , 2.645 , 0.332 , 0.583)

المرحلة الثانية: توليد المتغيرات التوضيحية, وبالاعتماد على دالة rand تم توليد نوعين من المتغيرات الا وهي (متغيرات كمية, متغيرات وصفي).

المرحلة الثالثة: توليد الأخطاء العشوائية وتتوزع طبيعيا $N(0,\sigma^2)$ حيث تم اخذ أربع قيم في التباين هي $(\sigma^2 = 0.01 , 0.5 , 1 , 5)$.

المرحلة الرابعة: توليد متغير الاستجابة Y (متغير كمي) الذي يتبع التوزيع الطبيعي بمتوسط $(m = 0)$ وتباين مقداره $(\mathrm{var} = \sigma^2)$ وفق المعادلات الآتية [8, 3]:

$$y(RT) = f(x\beta) + \varepsilon_i \quad \cdots\cdots\cdots (1-3)$$

$$y(NB) = exp(x_i\beta) = e^{x_i\beta} \quad \cdots\cdots\cdots (2-3)$$

المرحلة الخامسة: اختيار القيم الافتراضية للتباين والتي تمثلت بالقيم الآتية $(\sigma^2 = 0.01 , 0.5 , 1 , 5)$ لكل حجم عينة والتي تم ذكرها سابقا وهي (200 , 150 , 100 , 50).

المرحلة السادسة: تقدير معالم النماذج التي تم التطرق إليها في الجانب النظري ألا وهي :

- انحدار ثنائي الحدين السالب NBR .
- انحدار الشجيري RT .

ومن ثم المقارنة بين الأنموذجين وفق معيار (MSE) ومع تكرار عملية المحاكاة 1000

$$MSE = \frac{1}{n-p-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad \cdots\cdots\cdots (3-3)$$

جدول (1) يوضح تقدير المعالم لحجم عينة $n = 50$ وعندما تكون قيم $(\sigma^2 = 0.01 , 0.5 , 1 , 5)$

RT	NBR	β	σ	n
(X_1) 484.19059	149.090	β_0		
(X_6) 461.83520	0.183	β_1		
(X_3) 425.56048	-0.392	β_2		
(X_5) 197.92937	-0.132	β_3	0.01	
(X_2) 139.16869	5.209	β_4		
(X_{10}) 51.47963	7.733	β_5		
	8.895	β_6		
	-1.634	β_7		

		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
50	0.5	β_0	149.090	(X_1) 477.79351
		β_1	0.183	(X_6) 454.07593
		β_2	-0.392	(X_3) 420.67675
		β_3	-0.132	(X_5) 194.60397
		β_4	5.209	(X_2) 135.95665
		β_5	7.733	(X_{10}) 51.66137
		β_6	8.895	
		β_7	-1.634	
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	1	β_0	149.090	(X_3) 697.9125
		β_1	0.183	(X_1) 623.3473
		β_2	-0.392	(X_5) 484.5784
		β_3	-0.132	(X_6) 484.5784
		β_4	5.209	(X_2) 430.7363
		β_5	7.733	
		β_6	8.895	
		β_7	-1.634	
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	5	β_0	149.090	(X_1) 488.08609
		β_1	0.183	(X_3) 396.58905
		β_2	-0.392	(X_5) 120.34154
		β_3	-0.132	(X_4) 22.95067
		β_4	5.209	(X_{10}) 15.30045
		β_5	7.733	
		β_6	8.895	
		β_7	-1.634	

		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	

جدول (2) يوضح تقدير المعالم لحجم عينة $n = 100$ وعندما تكون قيم $(\sigma^2 = 0.01, 0.5, 1, 5)$

n	σ	β	NBR	RT
100	0.01	β_0	149.090	(X_5) 2154.4670
		β_1	0.183	(X_6) 1844.9211
		β_2	-0.392	(X_1) 1347.0037
		β_3	-0.132	(X_4) 737.9684
		β_4	5.209	(X_8) 737.9684
		β_5	7.733	(X_3) 188.9444
		β_6	8.895	(X_{10}) 54.6795
		β_7	-1.634	(X_{11}) 43.7436
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	0.5	β_0	149.090	(X_5) 2135.73785
		β_1	0.183	(X_6) 1830.53258
		β_2	-0.392	(X_1) 1376.45087
		β_3	-0.132	(X_8) 768.30263
		β_4	5.209	(X_4) 723.29819
		β_5	7.733	(X_3) 233.07583
		β_6	8.895	(X_{10}) 145.83416
		β_7	-1.634	(X_7) 22.28709
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
		β_0	149.090	(X_5) 1986.77943
		β_1	0.183	(X_6) 1189.96851
		β_2	-0.392	(X_1)

				925.84842
	1	$\boldsymbol{\beta_3}$	**-0.132**	(X_8) **855.96965**
		$\boldsymbol{\beta_4}$	**5.209**	(X_4) **629.35264**
		$\boldsymbol{\beta_5}$	**7.733**	(X_3) **186.43020**
		$\boldsymbol{\beta_6}$	**8.895**	(X_7) **50.13735**
		$\boldsymbol{\beta_7}$	**-1.634**	(X_{10}) **20.60231**
		$\boldsymbol{\beta_8}$	**-4.071**	
		$\boldsymbol{\beta_9}$	**2.645**	
		$\boldsymbol{\beta_{10}}$	**0.332**	
		$\boldsymbol{\beta_{11}}$	**0.583**	
	5	$\boldsymbol{\beta_0}$	**149.090**	(X_6) **2598.84804**
		$\boldsymbol{\beta_1}$	**0.183**	(X_1) **2255.50536**
		$\boldsymbol{\beta_2}$	**-0.392**	(X_5) **2149.03513**
		$\boldsymbol{\beta_3}$	**-0.132**	(X_8) **1614.64371**
		$\boldsymbol{\beta_4}$	**5.209**	(X_{10}) **1268.6486**
		$\boldsymbol{\beta_5}$	**7.733**	(X_3) **488.23891**
		$\boldsymbol{\beta_6}$	**8.895**	(X_{11}) **32.00195**
		$\boldsymbol{\beta_7}$	**-1.634**	(X_9) **26.11791**
		$\boldsymbol{\beta_8}$	**-4.071**	
		$\boldsymbol{\beta_9}$	**2.645**	
		$\boldsymbol{\beta_{10}}$	**0.332**	
		$\boldsymbol{\beta_{11}}$	**0.583**	

جدول (3) يوضح تقدير المعالم لحجم عينة $n = 150$ وعندما تكون قيم$(\sigma^2 = 0.01 , 0.5 , 1 , 5)$

n	σ	β	NBR	RT
		$\boldsymbol{\beta_0}$	149.090	(X_5) 3145.93930
		$\boldsymbol{\beta_1}$	0.183	(X_1) 2598.94442
		$\boldsymbol{\beta_2}$	-0.392	(X_6) 2525.94151
	0.01	$\boldsymbol{\beta_3}$	-0.132	(X_3) 1233.83369
		$\boldsymbol{\beta_4}$	5.209	(X_4) 786.48482
		$\boldsymbol{\beta_5}$	7.733	(X_8) 43.33676
		$\boldsymbol{\beta_6}$	8.895	(X_{10}) 28.89118

150		β_7	-1.634	(X_{11}) 28.89118
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	0.5	β_0	149.090	(X_5) 3181.02696
		β_1	0.183	(X_1) 2647.47518
		β_2	-0.392	(X_6) 2559.96376
		β_3	-0.132	(X_3) 1252.23480
		β_4	5.209	(X_4) 795.25674
		β_5	7.733	(X_{10}) 28.56402
		β_6	8.895	(X_{11}) 28.56402
		β_7	-1.634	(X_8) 28.56402
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	1	β_0	149.090	(X_5) 3228.83468
		β_1	0.183	(X_1) 2832.00321
		β_2	-0.392	(X_6) 2600.26405
		β_3	-0.132	(X_3) 1395.52757
		β_4	5.209	(X_4) 643.74384
		β_5	7.733	(X_8) 39.61490
		β_6	8.895	(X_9) 10.11548
		β_7	-1.634	
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	5	β_0	149.090	(X_6) 3032.23739
		β_1	0.183	(X_5) 3020.19498
		β_2	-0.392	(X_1) 2389.41273
		β_3	-0.132	(X_3) 1419.98506
		β_4	5.209	(X_4) 1324.20989

		β_5	7.733	(X_{11}) 89.78187
		β_6	8.895	(X_8) 76.95589
		β_7	-1.634	(X_{10}) 39.72508
		β_8	-4.071	(X_9) 32.73558
		β_9	2.645	(X_2) 16.36779
		β_{10}	0.332	
		β_{11}	0.583	

جدول (4) يوضح تقدير المعالم لحجم عينة $n = 200$ وعندما تكون قيم$(\sigma^2 = 0.01 , 0.5 , 1 , 5)$

n	σ	β	NBR	RT
200	0.01	β_0	149.090	(X_5) 3657.90266
		β_1	0.183	(X_1) 3558.26306
		β_2	-0.392	(X_6) 3002.54127
		β_3	-0.132	(X_9) 1330.43176
		β_4	5.209	(X_3) 1285.04021
		β_5	7.733	(X_4) 76.51312
		β_6	8.895	(X_{11}) 36.83792
		β_7	-1.634	(X_7) 22.89418
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
	0.5	β_0	149.090	(X_5) 3378.1047
		β_1	0.183	(X_1) 3124.0524
		β_2	-0.392	(X_6) 2788.9483
		β_3	-0.132	(X_3) 1396.1453
		β_4	5.209	(X_9) 870.3867
		β_5	7.733	(X_4) 160.2552
		β_6	8.895	(X_8) 18.0028
		β_7	-1.634	
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
		β_0	149.090	(X_5) 3169.28540

		β_1	0.183	(X_1) 3032.00956
		β_2	-0.392	(X_6) 2736.48782
	1	β_3	-0.132	(X_3) 1389.72827
		β_4	5.209	(X_9) 710.62039
		β_5	7.733	(X_4) 138.54031
		β_6	8.895	(X_{10}) 29.19873
		β_7	-1.634	(X_8) 29.19873
		β_8	-4.071	
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	
		β_0	149.090	(X_1) 3644.71499
		β_1	0.183	(X_5) 3613.28088
		β_2	-0.392	(X_6) 2304.05607
	5	β_3	-0.132	(X_9) 1207.03713
		β_4	5.209	(X_4) 640.01557
		β_5	7.733	(X_3) 199.11597
		β_6	8.895	(X_{11}) 157.38003
		β_7	-1.634	(X_2) 73.11029
		β_8	-4.071	(X_8) 40.20928
		β_9	2.645	
		β_{10}	0.332	
		β_{11}	0.583	

الجداول رقم (4) , (3) , (2) , (1) تبين لنا تقدير المعالم لأنموذج الانحدار ذي الحدين السالب (NBR) , بينما الأنموذج اللامعلمي وهو الانحدار الشجيري (RT) يظهر لنا أكثر المتغيرات تأثيرا في أنموذج الانحدار.

جدول (5) يوضح قيم متوسط مربعات الخطأ لأحجام عينات $(n = 50 , 100 , 150 , 200)$ ولقيم $(\sigma^2 = 0.01 , 0.5 , 1 , 5)$

n	σ	RT	NBR
50	0.01	6.773385	0.00009963501
	0.5	6.661607	0.2490875
	1	6.755196	0.9963501
	5	9.639913	24.90875
100	0.01	2.945342	0.0001000602
	0.5	2.99125	0.2501506
	1	3.013333	1.000602
	5	4.618357	25.01506
150	0.01	1.913719	0.0001003594
	0.5	1.937676	0.2508984
	1	1.958644	1.003594
	5	3.05978	25.08984
200	0.01	1.741679	0.00009993852
	0.5	1.758828	0.2498463
	1	1.78529	0.9993852
	5	2.668103	24.98463

من خلال الجدول أعلاه (5) نلاحظ الأتي:

لجميع أحجام العينات ولقيم التباين $(\sigma^2 = 0.01 , 0.5 , 1)$ يعتبر الانحدار ذي الحدين السالب هو الأفضل لامتلاكه اقل متوسط مربعات خطا, بينما عند قيمة التباين $(\sigma^2 = 5)$ فتكون قيمة متوسط مربعات الخطأ اقل قيمة بطريقة الانحدار الشجيري(RT) مقارنة بالانحدار ذي الحدين السالب (NBR) لذا تعتبر هي الأفضل.

أي أن كلما ازدادت قيمة التباين (σ^2) ولجميع أحجام العينات نلاحظ أن نموذج (RT) هو أفضل من أنموذج (NBR).

المبحث الرابع/ الاستنتاجات والتوصيات

يتضمن هذا المبحث أهم الاستنتاجات التي تم التوصل إليها وكذلك بعض التوصيات التي يوصى الأخذ بها.

4-1 **الاستنتاجات**:

1 – تبين انه عندما تكون قيمة $(\sigma^2 = 5)$ ولجميع أحجام العينات أن أنموذج الانحدار الشجيري (RT) هو أفضل من أنموذج انحدار ثنائي الحدين السالب (NBR) أي عندما تكون قيمة التباين مرتفعة.

2 – نلاحظ انه لجميع أحجام العينات وعندما تكون قيمة التباين $(\sigma^2 = 0.01 , 0.5 , 1)$ ولجميع أحجام العينات فان أنموذج انحدار ثنائي الحدين السالب أفضل من الانحدار الشجيري.

أي من خلال نتائج الجانب التجريبي تبين لنا بأنه عند ازدياد قيمة التباين فانه أنموذج الانحدار الشجيري هو الأفضل , بينما تكون طريقة الانحدار ذي الحدين السالب هي الأفضل عند قيم التباين الأخرى .

4-2 **التوصيات**:

1 – تقدير أنموذج انحدار ثنائي الحدين السالب باستخدام طرق أخرى لتقدير معالم الأنموذج غير طريقة الإمكان الأعظم ومنها (انحدار الحرف , مقدر ليو).

2 - قد تقام دراسة بخصوص المقارنة بين نفس النماذج المدروسة مع إضافة مشكلة التعدد الخطي.

المصادر:

1- Ali .O , Ahmed .S(2016), "Using Classification Regression Trees and Logistic Regression to Estimate Additive Model Comparison With Application", The Journal of A dministration & Economics, vol. 30 , no.109.

2- Andriyashin .A (2005),"Financial Application of classification and regression trees", Humboldt University, Berlin.

3- Asar .Y(2016), "Liu-type Negative Binomial Regression: A Comparison of Recent Estimators and Applications", Department of Mathmatics Necmettin Erbakan University Konya Turkey.

4- Fox .J(2000), " MULTIPLE AND GENERALIZED NONPARAMETRIC REGRESSION ", Series: Quantitative Applications in the Social Sciences, aSAGE UNIVERSITY PAPER.

5- Lord .D and Park .B(2012), "Negative Binomial Regression Models and Estimation Methods" , Texas A&M University , Korea Transport Instiute.

6- Rasheed .D and AL-Wakil .A, "introducation to mathematical statistics" ,university Baghdad, Al-hiwar University for Printing Publishing and Translation.

7- Sarraf .N , AL-rawi .A and Ismail .G(2016),"Estimate Statistics", University Baghdad.

8- Sepulveda .J(2012), "Comparacion entre Arboles de Regresion CART y Regresion Lineal", Universidad Nacional de Colombia.

9- Yuhong, Wu , Hakon, Tjelmeland & Mike, West (2006), "Bayesian CART: Prior Specification and Posterior Simulation". Institute of Statistics and Decision Sciences, Duke University Department of Mathematical Sciences, Norwegian University of Sciences and Technology.

Comparison Between Tree regression (TR), and Negative binomial regression (NBR) by Using Simulation.

D.Assma Najm Abd-allah
University of Baghdad/ College of A dministration & Economics/Dept statistics
Mobille: 07700300417
Gmail: asmaanajm92@gmail.com

Baraa Khudhair Abbas
pupils of University Baghdad/ College of A dminstration & Economics/Dept Statistic
Mobile: 07705885725
Gmail: bb2769250@gmail.com

Received :30/10/2019 Accepted :7/1/2020 Published :April / 2020

Abstract:

In this paper, the process of comparison between the tree regression model and the negative binomial regression. As these models included two types of statistical methods represented by the first type "nonparameter statistic" which is the tree regression that aims to divide the data set into subgroups, and the second type is the "parameter statistic" of negative binomial regression, which is usually used when dealing with medical data, especially when dealing with large sample sizes. Comparison of these methods according to the average mean squares error (MSE) and using the simulation of the experiment and taking different sample sizes where the results of simulation showed that the tree regression is best when the value of variance is large (5) and for all sample sizes model negative binomial regression when variance values (0.01, 0.5, 1) for all sample sizes, this method is superior to tree regression only when we take medium sample sizes.

Key words/ tree regression(cart), negative binomial regression, maximum likelihood methods, simulation.