Journal of Economics
and Administrative Sciences (JEAS)

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

مقارنة بين انحدار المربعات الصغرى الجزئية و الانحدار الشجيري 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: 13/11/2019 Accepted :7/1/2020 Published :June / 2020

هذا العمل مرخص تحت اتفاقية المشاع الابداعي نسب المُصنَّف - غير تجاري - الترخيص العمومي الدولي 4.0
Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

مستخلص البحث:

ناقش هذا البحث عملية المقارنة بين أنموذج انحدار المربعات الصغرى الجزئية والانحدار الشجيري, حيث شملت هذه النماذج نوعين من الأساليب الإحصائية تتمثل بالنوع الأول "الإحصاء ألمعلمي" وهو انحدار المربعات الصغرى الجزئية والتي يتم اعتمادها عندما يكون عدد المتغيرات اكبر من عدد المشاهدات وكذلك عندما يكون عدد المشاهدات اكبر من عدد المتغيرات أما النوع الثاني فهو "الإحصاء اللامعلمي" المتمثل بالانحدار الشجيري الذي يتمثل بتقسيم البيانات بشكل هرمي , وتم تقدير نماذج الانحدار للأنموذجين ومن ثم المقارنة بينهما .

حيث كانت المقارنة بين هذه الطرق وفق معيار متوسط مربعات الخطأ (MSE) وباستخدام المحاكاة للتجربة وبأخذ أحجام عينات مختلفة حيث أظهرت نتائج المحاكاة بان انحدار المربعات الصغرى الجزئية هي الأفضل عند اخذ قيم التباين الآتية (1 , 0.5 , 0.01) ولجميع أحجام العينات بينما ظهر أن الانحدار الشجيري هو الأفضل عندما تكون قيمة التباين كبيرة (5) ولجميع أحجام العينات .

المصطلحات الرئيسة للبحث/ انحدار المربعات الصغرى الجزئية, تقنية NIPLAS , الانحدار الشجيري, المحاكاة.

البحث مستل من رسالة ماجستير

المبحث الأول/ المقدمة العامة

1-1 المقدمة: Introduction

تكمن فلسفة الإحصاء من حيث محاولة إيجاد نماذج (Estimators 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, في هذا البحث جرت المقارنة بين أنموذجين من نماذج الانحدار لتقدير المعالم وهما الانحدار الشجيري (RT) وانحدار المربعات الصغرى الجزئية(PLSR) .

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 الشجيري وتارة أخرى انحدار المربعات الصغرى الجزئية لذا نحن في إطار المقارنة بينهم من خلال معيار المقارنة MSE .

3-1 هدف البحث: Purpose of the Research

إن الهدف من البحث هو المقارنة بين النماذج الآتية وهي: الانحدار الشجيري وانحدار المربعات الصغرى الجزئية في تقدير معالم الأنموذج وباستعمال المحاكاة.

المبحث الثاني/ الجانب النظري

1-2 المقدمة:

في هذا المبحث سيتم عرض ودراسة أنموذج انحدار المربعات الصغرى الجزئية (PLSR) من خلال تطبيق خوارزمية (NIPLAS(PLS1 و الانحدار الشجيري تقنية CART (RT) .

2-2 انحدار المربعات الصغرى الجزئية:

Partial Least Square Regression (PLSR)

تعد تقنية PLSR أداة سهلة وقوية للغاية للنمذجة الإحصائية وخاصة في المعالجات السريرية أي عندما يتعامل الباحث مع كمية هائلة من البيانات أي بوجود عدد كبير من المتغيرات مصحوبا بقليل من المشاهدات وذلك لاستخلاص العوامل الكامنة من اجل التنبؤ بعامل واحد أو أكثر[5]. و تحاول هذه التقنية بمحاولة فهم العلاقة بين المتغيرات التنبؤية ومتغير الاستجابة, ويعد أول من استخدم هذه الطريقة العالم الاقتصادي HERMON WOLD(1966). وقد أشار Svant إلى تقنية المربعات الصغرى الجزئية "إسقاط الهياكل الكامنة" وذلك من اجل وضع أفضل التحليلات لتقنية PLSR [5,3]. وتوجد العديد من الخوارزميات المتعلقة بالمربعات الصغرى الجزئية وجميعها تستند على خطوتين أساسيتين الأولى: تتمثل بإيجاد المتغيرات الكامنة بين X و Y عن طريق تعظيم مصفوفة التباين والتباين المشترك والخطوة الثانية: فهي انحدار Y على المركبات t.

$$X = T\acute{P} + E \quad \cdots\cdots\cdots (2-1)$$
$$Y = U\acute{q} + F \quad \cdots\cdots\cdots (2-2)$$

حيث أن:

X : تمثل مصفوفة المتغيرات التنبؤية من الرتبة $n \times m$.

Y : يمثل متجه متغير الاستجابة من الرتبة $n \times 1$.

T : مصفوفة إسقاطات x-score من الرتبة $n \times r$.

U : مصفوفة إسقاطات Y-score من الرتبة $n \times r$.

$\acute{P}$: مصفوفة X-loading ذات الرتبة $P \times r$.

q́ : متجه Y-loading ببعد $n \times 1$.

E : متجه البواقي X-residual ذات الرتبة $n \times P$.

F : قيمة البواقي Y-residual ببعد $n \times 1$.

والمصفوفة $\acute{P}$ والمتجه $\acute{q}$ لهما r من الأعمدة وهو محدد بما يأتي:

$$(r < min(n.p))$$

ويمكن التعبير عن العلاقة الداخلية التي تربط بين المتجهات القياسية كالأتي:

$$U = TD + H \quad \cdots\cdots\cdots (2-3)$$

إذ أن :

D : هي مصفوفة قطرية بأوزان الانحدار ذات بعد $r \times r$.

H : مصفوفة البواقي ذات بعد $n \times r$.

تتمثل الفكرة في طريقة المربعات الصغرى الجزئية بإيجاد مصفوفة الأوزان W من مجال X ومتجه C من مجال Y إذ أن:

$$Max\ cov(X_w, Y_c)$$

$$with \quad \|t\| = \|X_w\| = 1 \quad and \quad \|Y_c\| = 1 \cdots\cdots\cdots\cdots (2-4)$$

وان $cov(X_w, Y_c)$,هو تقدير لمصفوفة التباين المشترك ويتم تنفيذ الطريقة بصورة تكرارية متسلسلة ويتم احتسابها الواحدة بعد الأخرى. وبالتالي يتضمن إيجاد كافة المتجهات تحت قيد عدم الارتباط بين هذه المتجهات ,ويوجد هنالك العديد من الخوارزميات لحل المعادلة أعلاه, وفي هذا البحث سوف يتم الاعتماد على خوارزمية NIPLAS(PLS1)[5,6].

3-2 خوارزمية NIPLAS(PLS1)(6):

تتمثل خطوات الخوارزمية بالاتي:

1 – نقوم بتهيئة U_1 عن طريق Y إذ أن

$$U_1 = Y \quad \cdots\cdots\cdots (2-5)$$

2 – حساب أوزان X (X-weight) باستخدام انحدار ols :

$$W_1 = \acute{X}U_1 / \acute{U}_1U_1 \quad \cdots\cdots\cdots\cdots (2-6)$$

وان W_1 هي متجه ببعد $p \times 1$

W_1 تكون normalized بالشكل الأتي:

$$W_1 = W_1 / \|W_1\|$$

3 – لحساب (X-score) نبدأ بإسقاط بيانات X على (X-weight).

$$t_1 = XW_1 \quad \cdots\cdots\cdots\cdots (2-7)$$

وان t_1 هي متجه ببعد $n \times 1$.

4 – حساب (y-weight) بواسطة انحدار ols .

$$C_1 = \acute{Y}t_1 / \acute{t}_1t_1 \quad \cdots\cdots\cdots\cdots (2-8)$$

وان C_1 مصفوفة ببعد 1×1

حيث تكون C_1 normalized بالشكل الأتي:

$$C_1 = C_1 / \|C_1\|$$

5 – لحساب (y-scores) نقوم بإسقاط بيانات Y على (y-weight).

$$U_1^* = YC_1 \quad \cdots\cdots\cdots\cdots\cdots (2-9)$$

وان U_1^* متجه ببعد $n \times 1$.

6 – إيجاد U كالأتي:

$$\Delta u = (u\Delta)'(u\Delta) \quad \cdots\cdots\cdots\cdots (2-10)$$

$$\Delta u = u_1^* - u_1$$

فإذا كانت $\Delta u < \varepsilon$ وان ε قيمة صغيرة هذا يعني أننا وجدنا أول مركبة فنتوقف ,عدا ذلك نذهب إلى الخطوة رقم (1) وتستعمل $u_1 = u_1^*$

7 – إيجاد تحميلات X (X-loding) بواسطة انحدار ols وكالاتي:

$$P_1 = \acute{X}t_1 / \acute{t}t_1 \quad \cdots\cdots\cdots\cdots\cdots\cdots (2-11)$$

8 – إيجاد تحميلات Y (y-loding) بواسطة انحدار ols وكالاتي:

$$q = \acute{Y}U_1 / \acute{U}_1U_1 \cdots\cdots\cdots\cdots\cdots\cdots (2-12)$$

وان q متجه ببعد 1×1.

9 – ثم يتم إيجاد التداخل الخطي للمعلمات بواسطة انحدار ols :

$$d_1 = \acute{U}_1 t_1 / t_1 t_1 \cdots\cdots\cdots\cdots\cdots\cdots (2-13)$$

وان d_1 متجه ببعد 1×1.

10 – عمل تفريغ deflate إلى بيانات X.

$$X_1 = X - t_1 \acute{p}_1 \quad \cdots\cdots\cdots\cdots\cdots\cdots (2-14)$$

عمل تفريغ deflate إلى بيانات Y .

$$Y_1 = Y - d_1 t_1 \acute{c}_1 \quad \cdots\cdots\cdots\cdots\cdots (2-15)$$

ثم نستمر بالخطوات من (1-10) لعدد من المرات و باستخدام البيانات المفرغة لكل من X و Y لكي نحصل على عدة مركبات محدد ,وذلك لكي يتم تحديد معاملات الانحدار من خلال المعادلة الآتية:

$$\beta = W(\acute{P}W)^{-1}C \quad \cdots\cdots\cdots\cdots\cdots (2-16)$$

حيث أن:

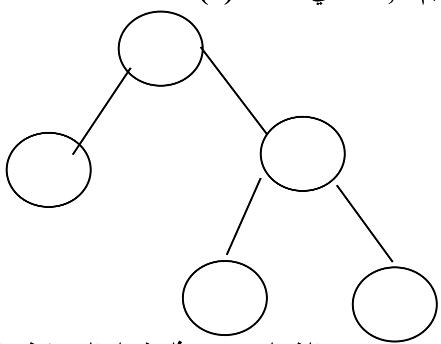
W : هي مصفوفة القيم العشوائية ببعد $p \times r$.

p : مصفوفة التحميلات للمصفوفة X من الرتبة $p \times r$.

C : مصفوفة التحميلات للمصفوفة Y من الرتبة $r \times r$.

4-2 **الانحدار الشجيري**: Tree regression

يعد الانحدار الشجيري طريقة جديدة تم تطويرها بواسطة مجموعة من العلماء الأمريكيين خلال الخمسة وعشرون سنة الماضية وهي من الطرق اللامعلمية , يوجد لها عدة تقنيات ومن بينها تقنية Cart أو ما يسمى ب"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]. حيث انه عقدة الجذر قد تحتوي على بيانات العينة للمشكلة قيد الدراسة أو جزء منها, ومن خلال هذه العقدة نجد أفضل متغير ممكن لتقسيم[4], كما في الشكل (1).



الشكل (1) يمثل شكل التصنيف الشجري

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

الانحدار الشجري يحتاج إلى عدد كبير من البيانات حيث تكون متغيرات كمية (مستمرة أو متقطعة) او 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.

$$\mathrm{X_i} = (X_1, X_2, \cdots\cdots, X_{\mathrm{m}})$$

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

1- إيجاد جميع التقسيمات الممكنة للمتغيرات التنبؤية, وفي أكثر الأحيان التقسيمات الثنائية تولد أسئلة ثنائية. من النموذج ($X_i > C$) لكل قيم C التي هي تكون ضمن مجال لـ X_i أي أن X_i تأخذ أعداد محدودة .

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

ونستطيع السؤال هنا; هل أن ($Xm \in C$) 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 او المربعات الصغرى.

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

عندما لا ينفذ التطبيق بشكل جيد. وذلك بسبب نمو الشجرة بشكل كبير جدا (النمو المفرط) حيث يلاحظ وجود عدد قليل من البيانات عند كل عقدة طرفية , ولذلك يتم تقليمها بشكل متكرر. وبالتالي إذا لم تنفذ هذه التقنية بشكل جيد نتوقف عند الخطوات أعلاه, إذ سيكون عندها عدد البيانات المتوفرة قليلة وعندها ستكون شجرة القرار كبيرة .

أما إذا استمرت تقنية المصنف الشجيري فعندها سيتم 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 [8] وبذلك يمكن التعبير عن الأنموذج الأساسي لانحدار الشجرة كالأتي[7] :

$$y = F(X_1, X_2, \cdots\cdots, X_p) + \varepsilon_i \qquad , \varepsilon_i \sim (0, \sigma^2) \quad \cdots\cdots (2-17)$$

حيث أن:

y : تمثل متغير الاستجابة

X_p :المتغيرات التوضيحية للأنموذج.

ε_i:الخطأ العشوائي.

أما الأنموذج المقدر فيكون بالصورة الآتية[1] :

$$f^{\wedge} = \sum_{Nm}^{n} CmI[(X_1, X_2) \in Rm] \quad \cdots\cdots (2-18)$$

المبحث الثالث/ الجانب التجريبي

1-3 المقدمة:

في هذا المبحث تم التطرق إلى استخدام أسلوب المحاكاة وذلك من اجل المقارنة بين (انحدار المربعات الصغرى الجزئية 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 المحاكــــاة: General Understanding of Simulation

لغرض المقارنة بين انموذج انحدار المربعات الصغرى الجزئية والانحدار الشجيري 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 ألا وهي (متغيرات كمية, متغيرات وصفي).

$$x_1 = round\ (uniform)(19, 86)$$
$$x_2 = round\ (uniform)(0, 1)$$
$$x_3 = round\ (uniform)(48, 87)$$
$$x_4 = round\ (uniform)(0, 1)$$

$$x_5 = round\ (uniform)(0,1)$$
$$x_6 = round\ (uniform)(0,1)$$
$$x_7 = round\ (uniform)(0,1)$$
$$x_8 = round\ (uniform)(0,1)$$
$$x_9 = round\ (uniform)(0,1)$$
$$x_{10} = round\ (uniform)(0,1)$$
$$x_{11} = round\ (uniform)(0,1)$$

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 $(\mu = 0)$ وتباين مقداره $(\text{var} = \sigma^2)$ وفق المعادلات الآتية:

$$y(pls) = x\beta + \varepsilon_i \qquad \cdots\cdots\cdots (1-3)$$

$$y(RT) = f(x\beta) + \varepsilon_i \qquad \cdots\cdots\cdots (2-3)$$

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 (50 , 100 , 150 , 200).

المرحلة السادسة:

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

- انحدار المربعات الصغرى الجزئية.
- انحدار الشجيري RT .

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

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

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

n	σ	β	PLSR	RT
50	0.01	β_0	148.3663675	(X_1) 484.19059
		β_1	0.1907570	(X_6) 461.83520
		β_2	-0.7878616	(X_3) 425.56048
		β_3	-0.1272534	(X_5) 197.92937
		β_4	5.0918982	(X_2) 139.16869
		β_5	7.8500637	(X_{10}) 51.47963
		β_6	8.4388441	
		β_7	-1.9532736	
		β_8	-3.4564555	
		β_9	2.2840332	
		β_{10}	0.9449588	
		β_{11}	0.6565847	
	0.5	β_0	147.7583153	(X_1) 477.79351
		β_1	0.1941648	(X_6) 454.07593
		β_2	-0.5411838	(X_3) 420.67675
		β_3	-0.1195264	(X_5) 194.60397
		β_4	5.0698318	(X_2) 135.95665
		β_5	7.7802550	(X_{10}) 51.66137
		β_6	8.5157218	

		β_7	-2.3299363	
		β_8	-3.7187807	
		β_9	2.3299363	
		β_{10}	0.8800767	
		β_{11}	0.8131313	
	1	β_0	147.1382649	(X_3) 697.9125
		β_1	0.1976339	(X_1) 623.3473
		β_2	-0.2900744	(X_5) 484.5784
		β_3	-0.1116318	(X_6) 484.5784
		β_4	5.0471009	(X_2) 430.7363
		β_5	7.7079634	
		β_6	8.5939993	
		β_7	-2.1108147	
		β_8	-3.9865086	
		β_9	2.3775807	
		β_{10}	0.8137929	
		β_{11}	0.9738420	
	5	β_0	142.21254011	(X_1) 488.08609
		β_1	0.22506827	(X_3) 396.58905
		β_2	1.70787102	(X_5) 120.34154
		β_3	-0.04837624	(X_4) 22.95067
		β_4	4.85390479	(X_{10}) 15.30045
		β_5	7.09093224	
		β_6	9.20913438	
		β_7	-2.76361741	
		β_8	-6.12832304	
		β_9	2.77533059	
		β_{10}	0.28508273	
		β_{11}	2.29514918	

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

n	σ	β	PLSR	RT
	0.01	β_0	148.9870318	(X_5) 2154.4670
		β_1	0.1838216	(X_6) 1844.9211
		β_2	-0.3535884	(X_1) 1347.0037
		β_3	-0.1308021	(X_4) 737.9684
		β_4	5.2262202	(X_8) 737.9684
		β_5	7.6828727	(X_3) 188.9444
		β_6	8.9159229	(X_{10}) 54.6795
		β_7	-1.6748291	(X_{11}) 43.7436
		β_8	-4.0656349	
		β_9	2.5878586	

100		β_{10}	0.3658727	
		β_{11}	0.5735848	
	0.5	β_0	148.9394561	(X_5) 2135.73785
		β_1	0.1821870	(X_6) 1830.53258
		β_2	-0.2071687	(X_1) 1376.45087
		β_3	-0.1296672	(X_8) 768.30263
		β_4	5.0844269	(X_4) 723.29819
		β_5	7.6430615	(X_3) 233.07583
		β_6	8.9001472	(X_{10}) 145.83416
		β_7	-1.6377265	(X_7) 22.28709
		β_8	-4.0980619	
		β_9	2.4932673	
		β_{10}	0.4831928	
		β_{11}	0.5333697	
	1	β_0	148.89098536	(X_5) 1986.77943
		β_1	0.18051904	(X_6) 1189.96851
		β_2	-0.05781452	(X_1) 925.84842
		β_3	-0.12850988	(X_8) 855.96965
		β_4	4.93972666	(X_4) 629.35264
		β_5	7.60242910	(X_3) 186.43020
		β_6	8.88402959	(X_7) 50.13735
		β_7	-1.59982646	(X_{10}) 20.60231
		β_8	-4.13114903	
		β_9	2.39674403	
		β_{10}	0.60291879	
		β_{11}	0.49232435	
	5	β_0	148.5066085	(X_6) 2598.84804
		β_1	0.1671766	(X_1) 2255.50536
		β_2	1.1345071	(X_5) 2149.03513

		β_3	-0.1192846	(X_8) 1614.64371
		β_4	3.7817866	(X_{10}) 1268.6486
		β_5	7.2770874	(X_3) 488.23891
		β_6	8.7544500	(X_{11}) 32.00195
		β_7	-1.2950775	(X_9) 26.11791
		β_8	-4.3957046	
		β_9	1.6244054	
		β_{10}	1.5609704	
		β_{11}	0.1633664	

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

n	σ	β	PLSR	RT
150	0.01	β_0	148.9355947	(X_5) 3145.93930
		β_1	0.1843038	(X_1) 2598.94442
		β_2	-0.4087445	(X_6) 2525.94151
		β_3	-0.1297836	(X_3) 1233.83369
		β_4	5.1068553	(X_4) 786.48482
		β_5	7.6057658	(X_8) 43.33676
		β_6	9.0124608	(X_{10}) 28.89118
		β_7	-1.6176325	(X_{11}) 28.89118
		β_8	-4.0528975	
		β_9	2.7366246	
		β_{10}	0.2498246	
		β_{11}	0.5425553	
	0.5	β_0	149.0426511	(X_5) 3181.02696
		β_1	0.1871402	(X_1) 2647.47518
		β_2	-0.3445402	(X_6) 2559.96376
		β_3	-0.1356424	(X_3) 1252.23480
		β_4	5.1156125	(X_4) 795.25674
		β_5	7.5974801	(X_{10})

				28.56402
		β_6	8.9221369	(X_{11}) 28.56402
		β_7	-1.6603018	(X_8) 28.56402
		β_8	-4.1307677	
		β_9	3.0092622	
		β_{10}	0.2330799	
		β_{11}	0.6886937	
	1	β_0	149.1517790	(X_5) 3228.83468
		β_1	0.1900345	(X_1) 2832.00321
		β_2	-0.2790701	(X_6) 2600.26405
		β_3	-0.1416197	(X_3) 1395.52757
		β_4	5.1245212	(X_4) 643.74384
		β_5	7.5889630	(X_8) 39.61490
		β_6	8.8299009	(X_9) 10.11548
		β_7	-1.7037654	
		β_8	-4.2100565	
		β_9	3.2874207	
		β_{10}	0.2159896	
		β_{11}	0.8378857	
	5	β_0	150.01990999	(X_6) 3032.23739
		β_1	0.21317962	(X_5) 3020.19498
		β_2	0.24285663	(X_1) 2389.41273
		β_3	-0.18936942	(X_3) 1419.98506
		β_4	5.19472519	(X_4) 1324.20989
		β_5	7.51920177	(X_{11}) 89.78187
		β_6	8.09056227	(X_8) 76.95589
		β_7	-2.04886389	(X_{10}) 39.72508
		β_8	-4.83928878	(X_9) 32.73558
		β_9	5.51033697	(X_2) 16.36779
		β_{10}	0.07932406	
		β_{11}	2.03284937	

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

n	σ	β	PLSR	RT
200	0.01	β_0	149.1034405	(X_5) 3657.90266
		β_1	0.1837620	(X_1) 3558.26306
		β_2	-0.3781489	(X_6) 3002.54127
		β_3	-0.1322885	(X_9) 1330.43176
		β_4	5.1352417	(X_3) 1285.04021
		β_5	7.7520354	(X_4) 76.51312
		β_6	8.8956567	(X_{11}) 36.83792
		β_7	-1.5692452	(X_7) 22.89418
		β_8	-4.1150529	
		β_9	2.5935667	
		β_{10}	0.2978400	
		β_{11}	0.6161766	
	0.5	β_0	149.2530231	(X_5) 3378.1047
		β_1	0.1838304	(X_1) 3124.0524
		β_2	-0.3137039	(X_6) 2788.9483
		β_3	-0.1368942	(X_3) 1396.1453
		β_4	5.1143659	(X_9) 870.3867
		β_5	7.7518653	(X_4) 160.2552
		β_6	8.9271185	(X_8) 18.0028
		β_7	-1.4815577	
		β_8	-4.0996470	
		β_9	2.7291445	
		β_{10}	0.2830550	
		β_{11}	0.6569534	
	1	β_0	149.4056835	(X_5) 3169.28540
		β_1	0.1839000	(X_1) 3032.00956
		β_2	-0.2479497	(X_6) 2736.48782
		β_3	-0.1415941	(X_3) 1389.72827
		β_4	5.0930770	(X_9) 710.62039
		β_5	7.7516931	(X_4)

				138.54031
		β_6	8.9592005	(X_{10}) 29.19873
		β_7	-1.3920991	(X_8) 29.19873
		β_8	-4.0839131	
		β_9	2.8675193	
		β_{10}	0.2679962	
		β_{11}	0.6985188	
	5	β_0	150.6277353	(X_1) 3644.71499
		β_1	0.1844472	(X_5) 3613.28088
		β_2	0.2777175	(X_6) 2304.05607
		β_3	-0.1791940	(X_9) 1207.03713
		β_4	4.9232988	(X_4) 640.01557
		β_5	7.7504673	(X_3) 199.11597
		β_6	9.2151109	(X_{11}) 157.38003
		β_7	-0.6771078	(X_2) 73.11029
		β_8	-3.9575788	(X_8) 40.20928
		β_9	3.9753088	
		β_{10}	0.1484524	
		β_{11}	1.0293544	

من خلال الجداول رقم (4) , (3) , (2) , (1) يظهر لدينا قيم المعالم المقدرة لأنموذج المربعات الصغرى الجزئية (PLSR) , بينما طريقة الانحدار الشجيري (RT) اللامعلمية فيظهر لدينا المتغيرات الأكثر تأثيرا في الأنموذج ومرتبة حسب الأهمية.

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	PLSR
50	0.01	6.773385	0.1439604
	0.5	6.661607	0.3330733
	1	6.755196	0.9009595
	5	9.639913	19.07254
100	0.01	2.945342	0.02459587
	0.5	2.99125	0.2447227
	1	3.013333	0.9054052
	5	4.618357	22.04786
150	0.01	1.913719	0.007326059
	0.5	1.937676	0.2380356
	1	1.958644	0.9304177
	5	3.05978	23.08646
200	0.01	1.741679	0.003305305
	0.5	1.758828	0.2380505
	1	1.78529	0.9425629
	5	2.668103	23.48691

من خلال الجدول أعلاه (5) نلاحظ الأتي:

نلاحظ عند جميع أحجام العينات (200 , 150 , 100 , 50) ولقيم التباين $(\sigma^2 = 0.01, 0.5, 1)$يعتبر أنموذج انحدار المربعات الصغرى الجزئية (PLSR) هي الأفضل لامتلاكها اقل متوسط مربعات خطا, بينما لقيمة تباين $(\sigma^2 = 5)$فتكون قيمة متوسط مربعات الخطأ اقل قيمة عند أنموذج الانحدار الشجيري (RT) لذا يعتبر هو الأفضل مقارنة مع أنموذج انحدار المربعات الصغرى الجزئية, وذلك لكون أساس عمل أنموذج الانحدار الشجيري يقوم على تقسيم مجموعة البيانات المتجانسة ووضعها في قطاعات وبذلك سوف تقل قيم التباين عندها

المبحث الرابع/ الاستنتاجات والتوصيات

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

1-4 الاستنتاجات:

1 – تبين انه عندما تكون قيمة $(\sigma^2 = 5)$ولجميع أحجام العينات أن أنموذج الانحدار الشجيري (RT) هو أفضل من أنموذج انحدار المربعات الصغرى الجزئية (PLSR) , أي كلما زادت قيمة التباين فان أنموذج الانحدار الشجيري هو الأفضل.

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

3 – يتضح لنا انه كلما زادت قيمة التباين فان أنموذج الانحدار الشجيري هو الأفضل.

2-4 التوصيات:

1 – تقدير أنموذج الانحدار الشجيري باستخدام خوارزميات أخرى غير خوارزمية (Cart) ثنائية التقسيم التي تم استعمالها في هذا البحث.

2 - قد تقام دراسة بخصوص المقارنة بين نفس النماذج المدروسة مع إضافة مشكلة عدم التجانس.

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- Hussien .E (2012), "Comparison of "Partial Least Squares Regression and The Effect Circumstances About Cement Stretch" ,The Education and Science Magazine, Vol .25, no.2.

4- Lewis .R(2000), "AnIntroducation to Classification and Regression Tree (CART) Analysis", Harbor-UCLA Medicel Center Torrance, California.

5- Roon.p , Zakizadeh.J and Chartier.S(2014),"Partial Least Squares tutorial for analyzing neuroimaging data" Carleton university and university of Ottawa, The quantitative Methods for Psychology, vol.10, no.2, pp. 200-215.

6- Saleh .R (2016), "Comparison of Partial Least Squares and Principle Components Methods by Simulation", The management and economic/ University Baghdad.

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

8- 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 Partial Least Square Regression(PLSR) and Tree Regression by Using Simulation(RT).

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: 13/11/2019 Accepted :7/1/2020 Published :June / 2020

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

Abstract:

This research discussed, the process of comparison between the regression model of partial least squares and tree regression, where these models included two types of statistical methods represented by the first type "parameter statistics" of the partial least squares, which is adopted when the number of variables is greater than the number of observations and also when the number of observations larger than the number of variables, the second type is the "nonparametric statistic" represented by tree regression, which is the division of data in a hierarchical way. The regression models for the two models were estimated, and then the comparison between them, where the comparison between these methods was according to a Mean Squares Error (MSE) and using the simulation of the experiment and by taking different sample sizes. where the results of the simulation showed that the regression of partial least squares is best when taking the following contrast variance values (0.01, 0.5, 1) and for all sample sizes, whereas tree regression is the best when it is The variance value is large (5) and for all sample sizes.

Key words/ partial least squares regression, technique NIPLAS, tree regression, simulation.