المجلة العراقية للعلوم الاحصائية (19) 2011
ص ص [219-232]

نموذج انحدار بيتا مع التطبيق

زكريا يحيى الجمال* اسامة بشير حنون**

الملخص

يهدف هذا البحث إلى عرض مفهوم انحدار بيتا وكيفية استخدامه في التطبيقات الإحصائية على اعتبار أن متغير الاستجابة له قيم محصورة بين الصفر والواحد ويتبع توزيع بيتا ، وقد تم استخدام معياري أكاكي و بيز للمعلومات لغرض الحصول على أفضل نموذج من خلال تطبيق انحدار بيتا المتعدد في دراسة مقدار النسبة المئوية لنسبة السكر في التلف الرطب الناتج إلى النسبة السكرية للشروب الخام الناتج. تم استخدام برنامج R لغرض الحصول على النتائج المرجوة من هذا البحث .

Multiple Beta Regression Model with Application

Abstract

This paper aims to display the concept of beta regression and modality of its utilization in the statistical application by considering the values of the response variable which are constrained by the interval(0,1) and the response variable is following beta distribution. Both Akaike and Bayesian information criteria are used to get the preferable model by applying the multiple beta regression in investigating the percentage of the sugar proportion of the resultant humid spoil to

* مدرس/قسم الاحصاء والمعلوماتية/كلية علوم الحاسوب والرياضيات/جامعة الموصل.
** مدرس مساعد/قسم الاحصاء والمعلوماتية/كلية علوم الحاسوب والرياضيات/جامعة الموصل.

تاريخ التسلم :19/ 7/ 2010 ـــــــ تاريخ القبول : 29/ 9/ 2010

the sugar proportion of the resultant crude juice. We used statistical software R to obtain the expected results.

1- المقدمة

تعتمد العديد من الدراسات والبحوث على أساليب متطورة من أجل الحصول على نتائج تتصف بالفعالية والدقة العاليتين، وقد كان لعلم الإحصاء وفروعه المرتبطة به الأثر الكبير في بناء النماذج الرصينة وتحليل البيانات من خلالها وصولاً للقرارات السليمة. يعد تحليل الانحدار أهم فروع علم الاحصاء، والذي يهتم ببناء العلاقة الرياضية بين متغير الاستجابة والمتغيرات التوضيحية وتتمثل هذه العلاقة كتركيبة خطية تدعى معادلة الانحدار إذ دقتها تعتمد على صحة تقدير معلماتها والتي تشترط توفر فروض التحليل ، وتعتبر طريقة المربعات الصغرى الاعتيادية أحد أكثر طرق تقدير المعلمات استخداماً وتتصف هذه الطريقة بكفاءتها العالية بتقدير معلمات نموذج الانحدار عند توفر فروض التحليل ، أما عند عدم توفر فروض التحليل فتصبح مقدراتها غير متسقة ولا متكاملة الشروط ، إن الفرض الأساس في طريقة المربعات الصغرى هو أن يكون متغير الاستجابة يتبع التوزيع الطبيعي ، ولكن قد يندر هذا الفرض في كثير من المجالات التطبيقية وعليه فيجب البحث عن طرق وأساليب إحصائية أخرى لدراسة نماذج الانحدار.

يعتبر انحدار بيتا أحد أهم الأساليب الإحصائية التي تستخدم في حال كان متغير الاستجابة لا يتبع التوزيع الطبيعي أو بعبارة أدق فإن انحدار بيتا يستخدم عندما يتبع متغير الاستجابة توزيع بيتا وتكون قيمه محصورة ضمن الفترة (1,0).[Korhonen et al. 2007]، وإن من قدم مفهوم انحدار بيتا هما الباحثان Ferrari وCribari-Neto عام 2004 [Ferrari & Cribari-Neto,2004] ثم توالت الدراسات في هذا المجال من قبل العديد من الباحثين أمثال Smithson &]

[Espinheira et al.,2008] و[Simas et al.,2010] وVerkuilen,2006] وغيرهم .

إن الهدف من هذا البحث هو عرض مفهوم انحدار بيتا المتعدد واستخدام معياري أكاكي وبيز للمعلومات للحصول على أفضل نموذج من خلال التطبيق على بيانات السكر لدراسة مقدار النسبة المئوية لنسبة السكر في التالف الرطب إلى النسبة السكرية في الشروب الخام الجديد . يشتمل هذا البحث على جانبين أولهما يتمثل بالجانب النظري وفيه تم عرض مفهوم انحدار بيتا وطريقة التقدير ومفهوم معايير المعلومات في حين يتضمن الجانب الثاني على الجانب التطبيقي و الاستنتاجات.

2- نموذج انحدار بيتا

يعتبر توزيع بيتا من التوزيعات القابلة للتكيف والمهمة في نمذجة المتغيرات العشوائية المستمرة ، والتي تشترط أن تكون قيمها محصورة بين الصفر والواحد ، وهنالك الكثير من هذه المتغيرات في المجالات العلمية المختلفة لها هيئة النسب أو النسب المئوية [Espinheira et al., 2008]، مثل النسبة المئوية لعدد الخلايا السرطانية لحالة مرضية معينة ونسبة الإنفاق على الطعام إلى الدخل الكلي. ليكن y_i متغير عشوائي يتبع توزيع بيتا له دالة كثافة احتمالية معطاة بالشكل الاتي:

$$f(y;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1}\ ;\ 0<y<1 \quad a,b>0 \quad ...(1)$$

إذ أن $\Gamma(.)$ تمثل دالة گاما، a,b تمثلان معلمتي قياس التوزيع، الوسط الحسابي والتباين لتوزيع بيتا يكون بالشكل الاتي:

$$E(y)=\frac{a}{a+b} \quad ...(2)$$

$$var(y)=\frac{ab}{(a+b)^2(a+b+1)} \quad ...(3)$$

[Krishnamorthy, 2006]

ويعد الباحثان [Ferrari and Cribari-Neto, 2004] هما أول من قدما مفهوم انحدار بيتا ، إذ اعتبرا أن متغير الاستجابة y يتبع توزيع بيتا المبين في المعادلة (1) وتكون قيمه محصورة بين الصفر والواحد ($0 < y < 1$) . يعتبر انحدار بيتا امتداد لموضوع النماذج الخطية المعممة (Generalized Linear Models (GLM)) ، إذ أن GLM تختلف عن الانحدار الخطي المعروف كون أن القيم المتوقعة μ_i للمتغير العشوائي y تستبدل بدالة الربط $g(\mu_i) = \eta$ حيث أن η هي تركيبة خطية من المتغيرات المستقلة ،إن الهدف من دالة الربط هو لجعل تباين الخطأ أكثر استقراراً بالإضافة إلى أنه من الممكن اختيار توزيع الخطأ الخاص بالنموذج بصورة مستقلة على عكس الانحدار الخطي الذي يتطلب دائماً أن يكون توزيع الخطأ توزيعاً طبيعياً [Korhonen et al., 2007].

ولغرض عرض نموذج انحدار بيتا فقد قام Gribario-Neto and Ferrari بإعادة كتابة المعادلة (1) بالشكل الاتي:

$$f(y;a,b) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \;;\; 0 < y < 1 \quad \ldots(4)$$

حيث أن

$$\mu = E(y) = \frac{a}{a+b}; \qquad 0 < \mu < 1 \qquad \ldots(5)$$

وأن

$$\phi = a + b; \qquad \phi > 0 \qquad \ldots(6)$$

أي يصبح الوسط الحسابي والتباين للمعادلة (4) بالشكل الاتي:

$$E(y) = \mu \qquad \ldots(7)$$

$$var(y) = \frac{\mu(1-\mu)}{1+\phi} \qquad \dots(8)$$

تسمى المعلمة μ بمعلمة الشكل وتسمى المعلمة ϕ في المعادلة (8) بمعلمة الدقة وذلك أنه كلما كانت قيمتها كبيرة أدى ذلك أن تكون قيم $var(y)$ صغيرة وذلك عند ثبوت μ ، وتدعى المعلمة ϕ^{-1} بمعلمة التشتت . لتكن $(y_1, y_2,\dots, y_n)$ متغيرات عشوائية مستقلة بحيث أن كل متغير منها يتبع توزيع بيتا المبين في المعادلة (4) أي أن $y_i \sim Beta(\mu_i, \phi_i)$ لذا فإن نموذج انحدار بيتا من الممكن الحصول عليه بفرض أن:

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j = \eta_i = x_i'\beta \qquad \dots(9)$$

حيث أن β تمثل متجه معلمات الانحدار وأن $(x_i = (x_{i1}, x_{i2}, \dots, x_{ip})')$ هو متجه المتغيرات المستقلة ، أما $g(\mu_i)$ فهي دالة الربط والتي تكون دالة غير خطية معرفة بالشكل الاتي:

$$\text{Logit } g(\mu_i) = \ln\left(\frac{\mu}{1-\mu}\right) \qquad \dots(10)$$

ومن خلال ربط المعادلتين (9) و(10) بالإعتماد على مفهوم الـ GLM فإن معادلة انحدار بيتا المتعدد هي عبارة عن الوسط الحسابي للمتغير y المبينة دالته في المعادلة (4) .

$$\ln(\mu/(1-\mu)) = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \dots\dots(11)$$

[Korhonen et al., 2007]

3- تقدير المعلمات واختبار جودة النموذج

لغرض تقدير معلمات نموذج انحدار بيتا فقد استخدما Ferrari and Cribari-Neto طريقة الإمكان الأعظم بأخذ اللوغارتم الطبيعي لدالة الإمكان لـــ عينة قوامها n من المشاهدات المستقلة للمعادلة (4)

$$L_i(\mu_i, \phi) = \operatorname{Log}\Gamma(\phi) - \operatorname{Log}\Gamma(\mu_i\phi) - \operatorname{Log}\Gamma\big((1-\mu_i)\phi\big)$$

$$+ (\mu_i\phi - 1)\operatorname{Log} y_i + \{(1-\mu_i)\phi - 1\}\operatorname{Log}(1-y_i), i = 1,2,\dots,n, \quad (12)$$

يمكن الحصول على مقدرات الإمكان الأعظم بأخذ المشتقة الجزئية لـــــ β و ϕ . لتفاصيل أكثر يرجى مراجعة[Ferrari and Cribari-Neto, 2004].

بعد توفيق نموذج انحدار بيتا المتعدد فإنه من المهم الكشف عن جودة النموذج المقدر وذلك بدراسة البواقي والقيم المؤثرة ، إذ أن البواقي التي يمكن استخدامها في نموذج انحدار بيتا المتعدد تدعى بـ(Pearson Residuals) ونرمز لها r_p . أنظر [Espinheira et al., 2008]، والتي يمكن اختبارها كالاتي:

$$r_p = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\widehat{var}(y_i)}} \quad \dots (13)$$

وللعلم فإن:

$$\hat{\mu}_i = g^{-1}(X_i'\hat{\beta}) \quad \dots (14)$$

$$\widehat{var}(y_i) = \frac{\hat{\mu}_i(1-\hat{\mu}_i)}{1+\hat{\phi}} \quad \dots (15)$$

حيث أن $\hat{\mu}_i$ يتم الحصول عليها من خلال صيغة $g^{-1}(.)$ للقيمة الملائمة i من الانحدار الخطي لـ $g(y_1), \dots, g(y_n)$ لـ X.ولغرض معرفة القيم المؤثرة على

النموذج يتم استخدام مقياس البعد لـ Cook وصيغته كالاتي [Yan and Su, 2009 :]

$$D_{Cook} = \frac{h_{ii}}{(1-h_{ii})^2}\left(\frac{y_i - \hat{\mu}_i}{\sqrt{var(y_i)}}\right) \quad \dots (16)$$

حيث أن h_{ii} هي العناصر القطرية في مصفوفة التقدير.

4- معايير اختيار أفضل نموذج

إن أسلوب اختيار نموذج معين من بين مجموعة من النماذج هو جانب مهم من جوانب تحليل البيانات إذ يقودنا إلى اختيار أفضل النماذج ، والمقصود بأفضل النماذج هو اختيار مجموعة جزئية من المتغيرات المستقلة من المجموعة الكلية لها تحت الدراسة ، وسنتناول في هذا البحث أفضل معايير المعلومات استخداماً وهي معيار أكاكي AIC ومعيار بيز BIC لغرض اختيار أفضل النماذج ، وكلا المعيارين يستلزمان حاصل جمع جزأين رئيسيين هما اللوغارتم الطبيعي لدالة الإمكان الخاصة بالنموذج تحت الإختيار وعدد معلمات النموذج وبالشكل الاتي:

$$AIC = -2(\text{Ln-likelihood}) + 2(P+1) \quad \dots (17)$$

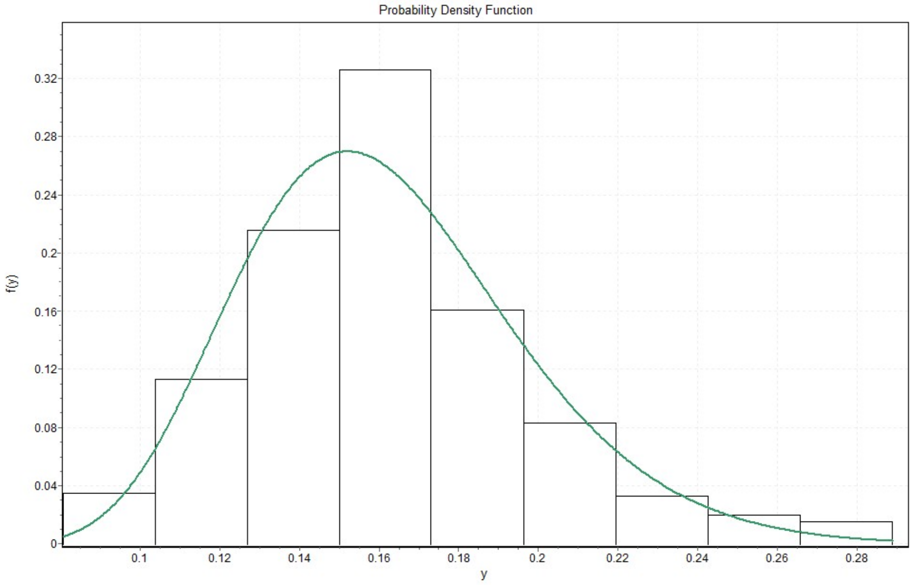
$$BIC = -2(\text{Ln-likelihood}) + (P+1)\,Ln(n) \quad \dots (18)$$

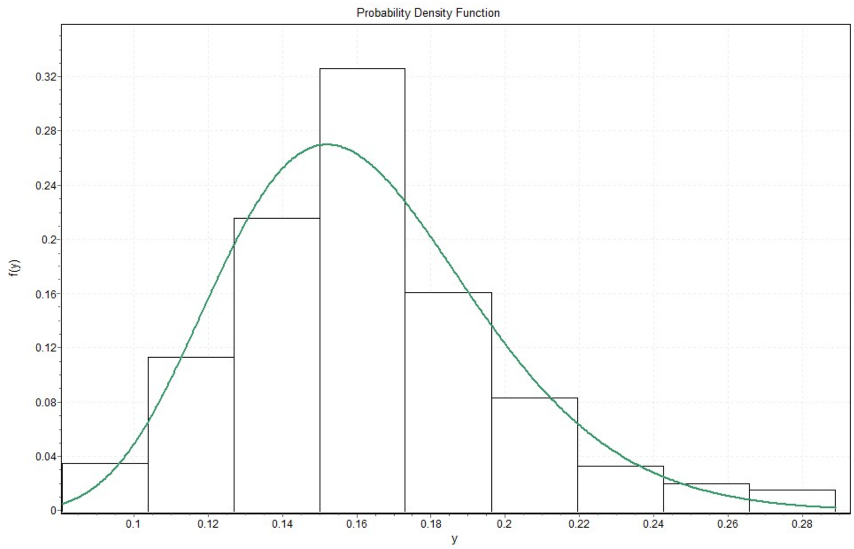
حيث n تمثل حجم العينة قيد الدراسة وأن P تمثل عدد المعلمات المقدرة .

إن القرار حول اختيار أفضل نموذج باستخدام معياري أكاكي وبيز يعتمد على الحصول على أقل قيمة لهذين المعيارين [Burnham and Anderson ,2002] .

5- الجانب التطبيقي

مما لاشك فيه أن نوعية البيانات التي تلائم الجانب التطبيقي من جانب ، ودقتها من جانب آخر ، يعدان ركيزة أساسية في الوصول إلى نتائج يمكن الإعتماد عليها والركون إليها في أي عمل بحثي ، تم أخذ بيانات من معمل سكر نينوى وشملت متغيرات النسب السكرية للماء وشرائح البنجر وتلف البنجر والشروب الخام الناتج وكانت لـ(399) من المشاهدات وتوفرت في البيانات العلاقات المطلوبة لذلك فقد اعتمدت لتطبيق ما أتى به هذا البحث حيث أن المتغيرات المستقلة تمثل النسبة السكرية للشروب الخام الجديد(X_1) والنسبة السكرية لشرائح البنجر المقطع(X_2) وماء العصر(X_3) ، أما متغير الإستجابة (Y) فيمثل النسبة المئوية لنسبة السكر في التلاف الرطب إلى النسبة السكرية للشروب الخام الناتج. من خلال رسم المدرج التكراري لمتغير الإستجابة واستخدام اختبار χ^2 لاختبار حسن المطابقة فقد فشل متغير الاستجابة كونه يتوزع توزيعاً طبيعياً في حين نجح هذا المتغير في اختبار ملائمته إلى توزيع بيتا(χ^2=10.413, p-value=0.237, df=8)





الشكل(1): المدرج التكراري لـ_ _ _ _ _ _ _ _ _ _ _ Y

ولغرض الحصول على النتائج يتم تطبيق ما ورد شرحه في الجانب النظري ، والجدول (1) أدناه يوضح نتائج استخدام انحدار بيتا المتعدد.

الجدول(1): نتائج انحدار بيتا المتعدد باستخدام دالة الربط logit

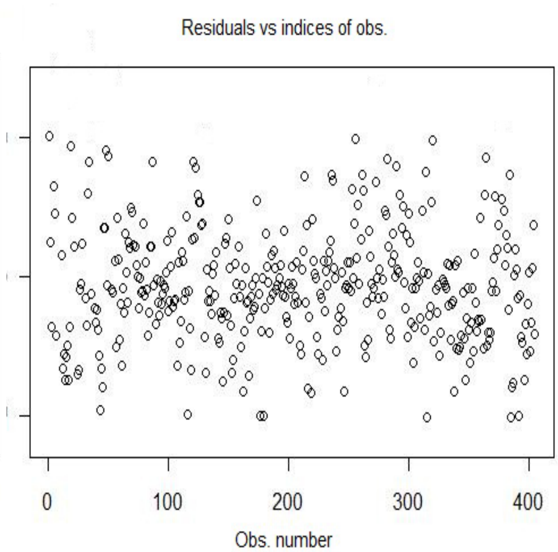
	Estimate	Std. Error	z-test	p-value
Intercept	-1.87214	0.25173	-7.437	0.0000 **
X_1	-0.0366	0.01080	-3.990	0.0006 **
X_2	0.0585	0.01630	3.59	0.0003 **
X_3	-0.0134	0.01825	-0.737	0.046 *

(*):significant at $\alpha = 0.05$, (**):significant at $\alpha = 0.01$

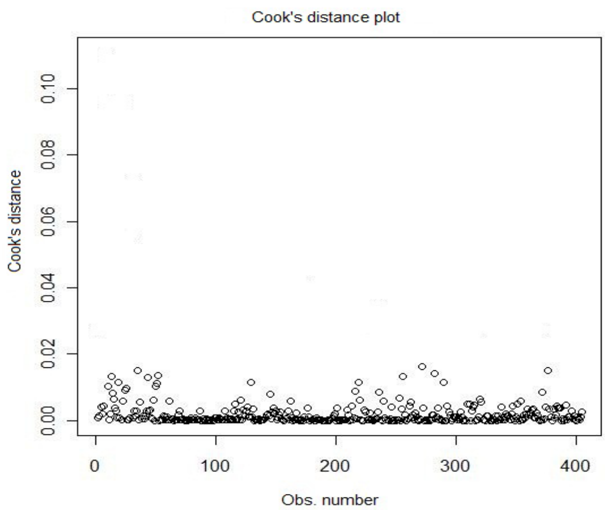
لذا فمن الممكن صياغة نموذج الانحدار وكالاتي:

$$\text{Log}\left(\frac{\hat{Y}_i}{1-\hat{Y}_i}\right) = -1.87214 - 0.0366\,X_1 + 0.0585\,X_2 - 0.0134\,X_3$$

ولغرض معرفة واختبار دقة النموذج قمنا برسم البواقي كما في المعادلة (17) وكذلك مقياس البعد لـ Cook حسب المعادلة(20) والأشكال (2) و(3) يوضحان ذلك .



شكل(2):رسم توضيحي لانتشار البواقي



شكل(3): مقياس البعد لـ Cook

ولغرض الحصول على أفضل نموذج قمنا بتطبيق ما ورد في المبحث الرابع في المعادلتين (21) و (22) على التوالي ، حيث كان عدد النماذج التي يمكن تكوينها هو سبع نماذج (2^3 -1) والموضحة في الجدول (2) أدناه.

الجدول (2): نتائج **AIC** و **BIC** لأختيار أفضل نموذج

النموذج	AIC	BIC
Full Model	-1564.262	-1544.318
X_1	-1555.271	-1543.305
X_2	-1557.191	-1545.224
X_3	-1548.07	-1536.103
X_1X_2	-1565.719	-1549.763
X_1X_3	-1553.546	-1537.59
$X2X_3$	-1555.328	-1539.372

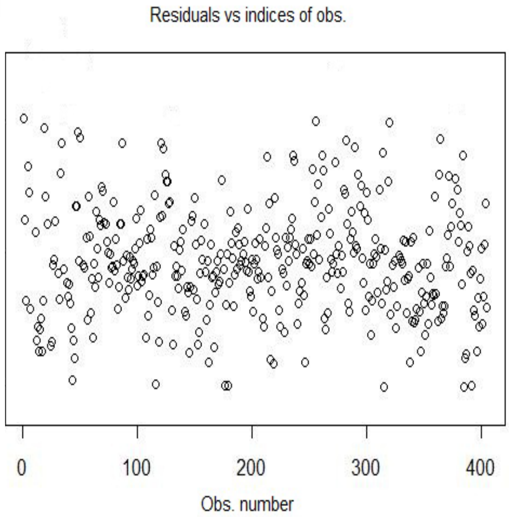
ومن خلال الجدول (2) يتبين بأن أقل قيم لمعياري AIC و BIC كانت عندما يحتوي النموذج على كلا المتغيرين X_1 و X_2 كمتغيرين مستقلين، والجدول (3) يوضح نتائج انحدار بيتا لأفضل نموذج وكالآتي:

الجدول(3): نتائج انحدار بيتا المتعدد للنموذج الذي تم اختياره

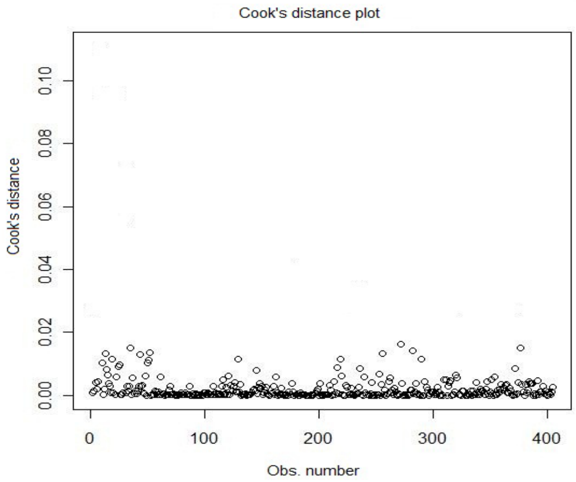
	Estimate	Std Error	z-test	p-value
Intercept	-1.966	0.21744	-9.045	0.0000 **
X_1	-0.0357	0.01077	-3.322	0.0008 **
X_2	0.0578	0.01629	3.552	0.0003 **

(*):significant at $\alpha = 0.05$, (**):significant at $\alpha = 0.01$

بعد اختيارنا لأفضل نموذج قمنا بإعادة حساب البواقي ومقياس البعد لـ Cook's وكما هو موضح بالشكلين (4) و(5) على التوالي.



شكل(4): رسم توضيحي لانتشار البواقي للنموذج الذي تم اختياره



شكل(5): مقياس البعد لـ Cook للنموذج الذي تم اختياره

6- الاستنتاجات

1- عندما تكون قيم متغير الاستجابة محصورة بين الصفر والواحد فقد يكون من غير الملائم استخدام الانحدار الخطي الاعتيادي عن طريق استخدام طريقة المربعات الصغرى للتقدير وذلك بسبب كون متغير الاستجابة لا يتوزع توزيعاً طبيعياً ، لذا يفضل استخدام انحدار بيتا وحسب الشروط .

2- في ما يخص الجانب التطبيقي فقد لوحظ من خلال الجدول (1) ومعادلة الانحدار التي تمثله أن المتغيران التوضيحيان X_1, X_3 لهما تأثير عكسي على متغير الاستجابة وهو مقدار النسبة المئوية لنسبة السكر في التالف الرطب إلى النسبة الكسرية في الشروب الخام الناتج Y ، على عكس المتغير التوضيحي X_2 والمتمثل في النسبة السكرية لشرائح البنجر المقطع والذي أظهر تغيراً طردياً مع متغير الاستجابة .

3- من خلال الشكلين (2) و(3) فقد كانت قيم البواقي متجانسة حول الصفر وأن القيم المؤثرة المحسوبة باستخدام مقياس البعد لـ *Cook* في الشكل (3) كانت غائبة تماماً وهذا يشير إلى أن نموذج انحدار بيتا للبيانات المشار إليها والمطبقة في هذا البحث كان ملائماً جداً .

4- تم اختيار النموذج المحتوي فقط على المتغيرين X_1, X_2 كأفضل نموذج مقارنةً ببقية النماذج الموضحة في الجدول (2) لكون أن كلا المعيارين AIC و BIC لهما اقل قيمة ، وكان النموذج أيضاً يحوي على تجانس كبير لقيم البواقي وأيضاً غياب القيم الشاذة وكما موضح في الشكلين (4) و(5) وهذا كله دلالة على أنه أفضل نموذج.

7- المصادر

1- Burnham, K., P. and Anderson, D., R., (2002), " Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach", 2nd ed., Springer-Verlag, New York.

2- Espinheira, P., L., Ferrari, S., L., P. and Cribari-Neto, F., (2008), " On Beta Regression Residuals", Journal of Applied Statistics, Vol. 35, No. 4, pp.407-419.

3- Espinheira, P., L., Ferrari, S., L., P. and Cribari-Neto, F., (2008),
" Influence Diagnostics in Beta Regression", Computational Statistics & Data Analysis, Vol.52, No.2, pp.4417-4431.

4- Ferrari, S., L., P. and Cribari-Neto, F., (2004), "Beta Regression for Modelling Rates and proportions" , Journal of Applied Statistics, Vol.31,No.7, pp. 799-815.

5- Korhonen, L., Korhonen, K.,T., Stenberg, P., Maltamo, M., and Rautiainen, M., (2007), " Local Models for Forest Canopy Cover with Beta Regression", Silva Fennica, 41,4, pp.671-685.

6- Krishnamoorthy, K., (2006), " Handbook of Statistical Distributions with Applications", Chapman & Hall/ CRC, Florida.

7- Simas, B., A., Barreto-Souza, W., and Rocha, A.,R., (2010), "Improved Estimators for a General Class of Beta regression Models", Computational Statistics & Data Analysis, Vol.54, No.2, pp.348-366.

8- Smithson, M. and Verkuilen, J., (2006), "A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables", Psychological Methods, vol. 11, No. 1, pp.54-71.

9- Yan, X. and Su, X.,G.,(2009), "Regression Analysis, Theory and Computing", World Scientific publishing Co., Uk.