استخدام طرق معايير المعلومات وطرق تشخيص النموذج لاختيار أفضل نموذج انحدار خطي متعدد مع تطبيق على أطفال مرضى الثيلاسيميا بالموصل

إيمان طارق فتحي

قسم الرياضيات / كلية التربية

جامعة الموصل

الاستلام	القبول
٢٠١١ / ١٠ / ١٧	٢٠١٢ / ٠٣ / ٠٧

ABSTRACT

In this paper we compute Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC) and Schwarz Bayesian Criteria (SBC) for all possible $(2^P - 1)$ for $p \le 10$ independent variables.

The three criteria are measurement of the difference between a given model and the real model, and the model with smallest (AIC), (BIC) and (SBC) compared with other models is the best one.

This study is applied to data gathered from children infected with Thelasymia disease, and the data are analyzed by using SAS to evaluate which of the models $(2^5 - 1 = 31)$ is the best model to determine the best subset of variables that minimize the information criteria among all the variables in the study. We compare the three criteria with model diagnostic like root mean squared error (RMSE), Mallow's C_p and adjusted R^2.

الملخص:

تم في هذا البحث حساب معيار اكاكي للمعلومات (AIC)، معيار بيز للمعلومات (BIC) ومعيار شوارز – بيز للمعلومات (SBC) لجميع النماذج الممكنة $(2^P - 1)$ حيث $P \le 10$ من المتغيرات المستقلة.

ان المعايير الثلاثة أعلاه هي مقياس للفرق بين النموذج المعطى والنموذج الحقيقي، وان النموذج الذي تكون فيه قيم المعايير الثلاثة صغيرة مقارنة بالنماذج الأخرى يعتبر هو النموذج الأفضل.

لقد قمنا بتطبيق الدراسة على بيانات جمعت لأطفال مصابين بمرض الثيلاسيميا في مستشفى ابن الاثير بالموصل وتم تحليلها باستخد ام البرنامج الإحصائي SAS لتقييم أي من النماذج ($2^5-1=31$) هو النموذج الافضل وذلك بتحديد أفضل مجموعة من المتغيرات والتي تعمل على تصغير معايير المعلومات من بين جميع المتغيرات قيد الدراسة . لقد تم مقارنة المعايير الثلاثة مع معايير تشخيص النموذج مث ل الجذر التربيعي لمتوسط مربعات الأخطاء (RMSE)، Mallow's C_p و R^2 المصححة.

١- المقدمة :

ان النموذج الحقيقي نادرا ما يكون معلوما، لذلك فان طرق اختيار النموذج تكون ذات فائدة كبيرة جدا في تحليل الانحدار الخطي ، انظر Draper and Smith (1981) , Linhart and Zucchini (1986).

ان الانحدار الخطي متعدد المتغيرات هو احد الاساليب الاحصائية المستخدمة في كشف العلاقة بين المتغيرات ، اذ انه يستخدم لايجاد النموذج الخطي الافضل للمتغير المعتمد من مجموعة من المتغيرات المستقلة.

هناك عدد كبير من البحوث تناولت مسألة اختيار المتغيرات في نموذج الانحدار الخطي (انظر Hocking(1976) and Thompson(1978 a,b)).

ان البيانات المستخدمة في هذا البحث هي بيانات لخمسة متغيرات مستقلة ($p=5$)، لذا فان هناك ٣١ مجموعة جزئية من النماذج ($2^5-1=31$) وبالتالي فان احد هذه النماذج الجزئية سيكون هو النموذج الافضل للبيانات المدروسة والذي يتصف بان قيمة ال RMSE فيه قليلة وقيمة $\overline{R}^2$ فيه عالية وتقترب من ١.

واخيرا فان هناك طرقا اخرى لاختيار النموذج وذلك بتصغير مقياس المعلومات لمجموعة من النماذج البديلة (المدروسة) (انظر Mallows (1973), Sparks, Zucchini and Coutsourides (1985)).

الجانب النظري

٢- طرق تشخيص النموذج :

قمنا في هذا البحث بدراسة ثلاثة طرق إحصائية لتحديد أفضل نموذج خطي وهي
A- الجذر التربيعي لمتوسط مربعات الأخطاء (RMSE).

B- معامل التحديد المعدل $(\overline{R}^2)$.

C- احصاءة C_p Mallow's .

وسوف نقوم بذكر هذه الطرق بشئ من الإيجاز :

A- الجذر التربيعي لمتوسط مربعات الأخطاء (RMSE)

ان الجذر التربيعي لمتوسط مربعات الاخطاء هي دالة لمجموع مربعات الاخطاء (SSE) وعدد المشاهدات (n) وعدد المتغيرات المستقلة $(k \le p+1)$ حيث k تشتمل على المعلمة الثابتة في النموذج (Intercept) ويمكن كتابتها كما يلي:

$$RMSE = \sqrt{\frac{SSE}{n-k}} \qquad \ldots\ldots(1)$$

حيث يتم حساب RMSE لكل النماذج ، والنموذج الذي تكون فيه قيمة RMSE قليلة هو النموذج الأفضل . ان هذه الطريقة تعتمد على عدد المعلمات في النموذج (k) لذلك فان إضافة معلمات أخرى للنموذج سوف يقلل كل من البسط والمقام انظر Beal (2007).

B- معامل التحديد المعدل $(\overline{R}^2)$

ان معامل التحديد R^2 ليست دائما هي الاحصاءة المقبولة في حساب ملائمة النماذج الجزئية للنموذج الأفضل، اذ ان إضافة متغيرات مستقلة جديدة الى معادلة الانحدار يؤدي الى رفع قيمة R^2 وذلك لثبات قيمة المقا م وتغير قيمة البسط غير ان الاستمرار بإضافة المتغيرات المستقلة سيؤدي الى انخفاض درجات الحرية $(n-k)$، مما يتطلب إيجاد معامل التحديد المعدل $(\overline{R}^2)$ والذي يمثل نسبة التغيرات للمتغير المعتمد والتي تفسر التغيرات الحاصلة في المتغيرات المستقلة، ويمكن كتابته على النحو التالي:

$$\text{Adjusted } R^2 = \overline{R}^2 = 1 - \frac{n-i}{n-k}\cdot(1-R^2) \qquad \ldots\ldots(2)$$

حيث ان n تمثل حجم العينة، $(i=1)$ اذا كان هناك β_0 في النموذج و $(i=0)$ اذا لم يكن هناك β_0 في النموذج، k تمثل عدد المعلمات في النموذج .ان النموذج الذي تكون فيه قيمة R^2 المعدلة عالية هي افضل النماذج المدروسة، مع ملاحظة ان المتغيرات المستقلة الجديدة المضافة الى النموذج غالبا ما تجعل $(\overline{R}^2 = 1)$ في عدة نماذج وبالتالي فان تحديد النموذج الافضل في هذه الحالة سيكون معضلة يصعب حلها، انظر(2008) Stauffer.

C- احصاءة Mallow's C_p

وضع Mallows عام ١٩٧٣ احصاءة عرفت باحصاءة C_p Mallow's واساسها تقييم ملائمة المربعات الصغرى للنماذج بأخطاء طبيعية وذات تباين ثابت وكما يلي:

$$C_P = P + (n-P)\cdot\frac{(\hat{\sigma}^2 - \hat{\sigma}^2_{full})}{\hat{\sigma}^2} \quad \ldots\ldots(3)$$

اذ ان p تمثل عدد المتغيرات في النموذج ، وان النموذج الكامل هو النموذج الذي يأخذ جميع المتغيرات المستقلة بنظر الاعتبار.

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n} = \frac{n-P-1}{n}\cdot S^2_{y|x}$$

فالنموذج الذي تكون فيه قيمة C_p Mallow's صغيرة هو النموذج الافضل او النموذج الاكثر دقة. لاستخدام احصاءة C_p Mallow's سوف نقوم بحساب قيمة C_p لكل مجموعة جزئية من المتغيرات المستقلة ثم نقوم برسم النقاط (C_P, P) لكل نموذج مدروس، والنموذج الذي تكون فيه $C_P \cong P$ يعتبر هو افضل النماذج. انظر Stauffer (2008).

٣- معايير المعلومات Information Criteria

معيار المعلومات هو معيار لحسن المطابقة او عدم التأكدية لمدى قيم مجموعة من البيانات، وفي سياق الكلام عن الانحدار الخطي المتعدد فان معيار المعلومات يقيس الفرق بين النموذج المعطى والنموذج الحقيقي . سوف نقوم بالتعرف على ثلاثة أنواع من طرق معايير المعلومات وهي:

١- معيار اكاكي للمعلومات Akaike's Information Criteria

ان احصاءة نظرية المعلومات المهمة هي اكثر فاعلية في تقييم المقارنات بين النماذج ، ان معيار اكاكي للمعلومات قدم لاول مرة عام (١٩٧٣) من قبل العالم الرياضي الياباني هيتروتكو اكاكي وتعرف بانها مقياس لنظرية المعلومات لـ (كولبك-ليبلر) بين النموذج المعطى والنموذج الحقيقي . ان معيار اكاكي للمعلومات (AIC) ومعيار اكاكي للمعلومات المصحح (AIC_C) للانحدار الخطي المتعدد يعطيان بالصيغتين التاليتين:

$$AIC = n\ln(S^2_{y|x}\cdot\frac{n-p-1}{n}) + 2K \quad \ldots\ldots(4)$$

$$= n\ln(S^2_{y|x}\cdot\frac{n-K+1}{n}) + 2K$$

$$AIC_C = n\ln(S^2_{y|x}\cdot\frac{n-p-1}{n}) + 2K + \frac{2K(K+1)}{(n-K-1)} \quad \ldots\ldots(5)$$

$$= n\ln(S^2_{y|x}\cdot\frac{n-K+1}{n}) + 2K + \frac{2K(K+1)}{(n-K-1)}$$

حيث p تمثل عدد المتغيرات المستقلة و $(K = p+2)$ تمثل عدد المعلمات (من ضمنها β_0, σ). ان معيار اكاكي للمعلومات (AIC) هو تقريب لسلسلة تايلر الخطية لمعلومات كولباك

– ليبلر ، بينما معيار اكاكي للمعلومات المصحح (AIC_C) هو تقريب لسلسلة تايلر غير الخطية لذلك فان (AIC_C) هي اكثر دقة من (AIC). ومن الجدير بالذكر انه اذا كان حجم العينة صغير وكانت ($\frac{n}{K} < 40$) فان النموذج الافضل هو النموذج الذي يملك اقل قيمة لـ (AIC_C) (انظر Burnham and Anderson (2002))، كما نلاحظ با نه كلما ازداد عدد المتغيرات المستقلة فان النماذج تؤول الى ان تكون ملائمة للبيانات.

وبصورة عامة فانه لاي نموذج احصائي احتمالي لمجموعة من بيانات عينة ما بدالة الامكان L فان (AIC) و (AIC_C) يعرفان باستخدام الانحراف $D = -2 \cdot \ln L$ وكما يلي:

$$AIC = D + 2 \cdot K$$

$$= -2 \cdot \ln L + 2 \cdot K$$

$$AIC_C = D + 2 \cdot K + 2 \cdot \frac{K(K+1)}{n-K-1}$$

$$AIC_C = -2 \cdot \ln L + 2 \cdot K + 2 \cdot \frac{K(K+1)}{n-K-1}$$

انظر Stauffer (2008).

٢- معيار المعلومات البيزي Bayesian Information Criterion

ان Sawa (1978) قدم معيار لاختيار النموذج والذي اشتق من تعديل بيز لمعيار (AIC). ان معيار بيز للمعلومات (BIC) هو دالة لعدد المشاهدات (n)، SSE وتباين الخطأ (σ^2) وعدد المتغيرات المستقلة $p \le K+1$ وكما مبين بالمعادلة التالية:

$$BIC = n \ln(\frac{SSE}{n}) + 2(K+2)(\frac{n}{\sigma^2}) - 2(\frac{n\sigma^2}{SSE})^2 \quad \ldots\ldots(6)$$

حيث ان $K = p + 2$ وهي عدد المعلمات المقدرة في النموذج ، والنموذج الذي تكون فيه قيمة BIC هي الأصغر هو النموذج الأفضل من بين النماذج قيد الدراسة، انظر Beal (2007).

٣- معيار شوارز البيزي Schwarz's Bayesian Criterion

في اطار الكلام عن بيز، فان Schwarz (1978) وصف معيار اخر سمي باسمه وهو معيار شوارز البيزي للمعلومات لاختيار النموذج الأفضل والذي يرمز له بـ SBC ويمكن كتابته كما يلي:

$$SBC = n \ln(\frac{SSE}{n}) + K \ln n \quad \ldots\ldots(7)$$

ان البسط في الحد الاول لهذا المعيار هو $(n \ln SSE)$ والذي يقل كلما ازداد عدد المتغيرات المستقلة، اما مقام الحد الأول فهو ثابت (بحجم عينة n) والحد الثاني يزداد مع ازدياد عدد المعلمات K في النموذج. اذا كانت $(n \geq 8)$ فان جزاء المعيار لـ SBC $(K \ln n)$ هي اكبر من جزاء المعيار لـ AIC $(2K)$ لذلك فان معيار SBC هو الافضل لمعظم النماذج قيد الدراسة ، انظر Kutner et al. (2004).

الجانب التطبيقي

تم اخذ البيانات من رسالة ماجستير للطالبة اسوان محمد طيب بعنوان (اختيار المتغيرات في انحدار الحرف) حيث قامت بجمع البيانات من مستشفى ابن الاثير التعليمي للولادة والاطفال في الموصل من مرضى مصابين بفقر دم البحر الأبيض المتوسط من نوع بيتا او ما يسمى بالثلاسيميا لـ (١١٠) مشاه دات حيث ان المتغير المعتمد او متغير الاستجابة في البحث هو العمر من العظم مقاسا بالشهر كما واختير عدد من المتغيرات التي يعتقد انها تؤثر فيه وهي:

العمر الحقيقي (X_1): مقاسا بالشهر، ان الأشخاص عرضة للإصابة بالمرض منذ مقتبل العمر ويصل اعلى نسبة للاصابة في عمر (٣٦ الى ٨٠) شهرا وتتركز في عمر (٨٠ الى ٨٥) شهرا.

تضخم الكبد (X_2): مقاسا بالسنتمتر ، اذ يعاني المصاب بالثلاسيميا من تضخم الكبد نتيجة لترسب الحديد بكميات كبيرة في الكبد وحدوث خلل في وظائفه.

الخلايا الشبكية (X_3): في دم المصاب يلاحظ ان نسبة الخلايا الشبكية تكون عالية وعندها يعتبر المريض في وضع صحي سيئ ، ويمكن اعتبار الخلايا الشبكية مقياس لفاعلية نخاع العظم على تكوين كريات دم جديدة.

الارومة الحمراء (X_4): وهي عبارة عن كريات دم حمراء جديدة التكوين وتكون غير ناضجة حاوية على نواة وتكون في نخاع العظم وتنتج لكثرة الحاجة اليها ، فتكون خارج نخاع العظم في الدم للمصابين بالثلاسيميا وهي مؤشر على وجود المرض.

بداية نقل الدم للمريض (X_5): حسب العمر مقاسا بالشهر اذ يلاحظ ان معظم المصابين يبدأ إعطاؤهم وحدات الدم في الأشهر الأولى من العمر وذلك لكون ظهور المرض في مقتبل العمر والكشف عنه منذ بداية ظهوره بالفحص والتحليل لدم المصاب سهلا.

اما المتغير المعتمد (Y): هو العمر من العظم مقاسا بالشهر للمرضى المصابين بالثلاسيميا ، ويعتمد مقياس العمر من العظم على الصور الشعاعية للمرضى المصابين اذ تبدو الصور الشعاعية للمرضى مقارنة باقرانهم الاصحاء اقل من العمر الحقيقي.

وقد تم عرض البيانات في الجدول التالي [جدول (١)] :

جدول (١): البيانات قيد الدراسة

ت	العمر الحقيقي بالعظم Y	العمر الحقيقي بالشهر X_1	تضخم الكبد X_2	الخلايا الشبكية X_3	اورمة حمراء X_4	بداية نقل الدم حسب العمر X_5		ت	العمر الحقيقي بالعظم Y	العمر الحقيقي بالشهر X_1	تضخم الكبد X_2	الخلايا الشبكية X_3	اورمة حمراء X_4	بداية نقل الدم حسب العمر X_5
١	132	147	8	6.4	120	6		٥٦	12	31	0	1	5	10
٢	108	153	6	0.1	17	6		٥٧	15	26	4	5.4	2	10
٣	48	114	4	7.6	1	12		٥٨	72	68	6	7.6	440	30
٤	72	84	7	2	0	36		٥٩	112	176	4	9.2	240	12
٥	72	96	6	1.1	1	7		٦٠	24	34	1	12.8	25	8
٦	108	82	4	2.7	200	12		٦١	60	44	7	2.3	0	12
٧	36	68	3	0.2	2	6		٦٢	30	54	2	0.4	0	7
٨	7	24	0	7.4	5	7		٦٣	9	37	3	1	14	9
٩	24	65	4.5	0.2	0	4		٦٤	120	129	6	13.3	72	3
١٠	68	83	10.5	1.8	18	4		٦٥	96	103	5	0.1	500	36
١١	48	51	6	2.6	4	4		٦٦	96	99	6	13.6	4	6
١٢	132	145	8	3	115	9		٦٧	72	79	10	6	4	4
١٣	60	89	4	2.2	115	12		٦٨	72	85	8	0.7	100	5
١٤	96	95	6.5	2.8	126	18		٦٩	60	76	3	0.9	2	4
١٥	48	70	5	0	0	3		٧٠	8	23	2	1.8	5	7
١٦	30	44	8	2.4	3	3		٧١	96	112	8	5	2	3
١٧	24	39	2	6.4	2	3		٧٢	24	46	2	2	0	3
١٨	60	75	5	0.8	0	11		٧٣	108	117	7	0	82	7
١٩	30	56	0	2.2	1	6		٧٤	36	73	5	5	3	12
٢٠	7	24	0	6.8	10	7		٧٥	76	44	6	2.2	3	6
٢١	64	88	7	3	390	2		٧٦	36	40	3	11.6	3	5
٢٢	108	97	7	3.4	210	2		٧٧	24	45	0	10.2	0	9
٢٣	9	20	3	17.8	35	9		٧٨	36	26	5	11	4	48
٢٤	132	109	6	4.7	100	12		٧٩	12	17	5.5	0	1	4
٢٥	48	45	5	0.1	18	4		٨٠	11	21	3	0	0	7

٢٦	33	18	6	2.4	0	1		٨١	168	148	4	0.2	440	3
٢٧	84	109	5	0.3	170	6		٨٢	36	41	3	0.3	0	18
٢٨	72	77	2	0.2	3	6		٨٣	24	38	4	0.2	5	9
٢٩	72	60	7	0	432	6.5		٨٤	60	69	4	0	2	6
٣٠	30	46	6	2	8	12		٨٥	66	79	5	0.1	0	4
٣١	84	74	5	3.6	14	8		٨٦	7	30	5	4.5	0	5
٣٢	96	113	3.5	4	0	7		٨٧	60	72	0	0.2	22	6
٣٣	30	50	3	9.5	10	6		٨٨	78	76	6	0	3	14
٣٤	48	82	2.5	0	6	18		٨٩	60	65	6	0.4	2	6
٣٥	60	90	4	3.6	20	12		٩٠	61	62	4	3	12	7
٣٦	48	50	10	4	10	48		٩١	63	74	5	3	40	24
٣٧	36	42	5	3	38	40		٩٢	36	63	3	3.4	0	6
٣٨	30	54	5.5	1.7	10	12		٩٣	60	71	8	2.6	10	2
٣٩	144	157	3.5	3.3	300	7		٩٤	36	37	2.5	14	10	9
٤٠	180	175	5.5	0.4	350	7		٩٥	84	111	1	0.2	0	6
٤١	60	67	4	2	10	18		٩٦	144	110	2	1	0	6
٤٢	48	45	4	0.3	0	18		٩٧	48	89	8	1.3	36	3
٤٣	48	109	8	0.1	43	7		٩٨	30	40	0	0.4	3	3
٤٤	66	58	8	4.3	210	48		٩٩	120	161	6	0	0	24
٤٥	132	147	3	19.2	120	14		١٠٠	9	26	0	7.5	1	6
٤٦	96	137	3.5	10.6	600	18		١٠١	18	31.4593	4	2.9	1	4
٤٧	24	40	4.5	3.6	18	6		١٠٢	60	77.1907	3	0.3	16	12
٤٨	48	46	6	8.6	2	30		١٠٣	120	116.1028	8	7.7	5	5
٤٩	42	88	9	2.4	0	6		١٠٤	57	12.3102	4	8.8	112	5
٥٠	18	65	0	7.4	3	6		١٠٥	66	42.7778	6	2	9	5
٥١	72	78	7	0	432	3		١٠٦	48	70.2398	4	1	0	7
٥٢	156	159	0	2	8	6		١٠٧	120	117.9954	8.5	1	25	6
٥٣	72	75	11	3.6	14	5		١٠٨	78	105.6019	4.5	6.6	1	24
٥٤	96	94	4	9.5	10	48		١٠٩	156	166.1943	4.5	1	1	6
٥٥	36	33	3	11.4	172	8		١١٠	78	77.3648	2.5	6	7	4

ان نموذج الانحدار الخطي المتعدد الذي يعتمد على البيانات في جدول (١) أعلاه يمكن التعبير عنه بالشكل التالي:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i$$

حيث ان:

β_j : تمثل معلمات غير معلومة (معاملات الانحدار)، $j = 1,2,3,4,5$

وكانت نتائج التحليل الإحصائي (باستخدام البرنامج الإحصائي SAS) لهذه البيانات كما يلي:

جدول (٢) : نتائج التحليل الإحصائي للبيانات لاختيار أفضل نموذج انحدار خطي متعدد

MODEL	Dependent Variable	Intercept	x1	x2	x3	x4	x5	Independent Variables in Model	Parameters in Model
MODEL1	y	-7.22171	0.833829	1.304645		0.032337		3	4
MODEL1	y	-7.92596	0.835799	1.265563		0.031453	0.075453	4	5
MODEL1	y	-7.99977	0.834684	1.342419	0.149718	0.03194		4	5
MODEL1	y	-8.48646	0.836288	1.300602	0.122331	0.031222	0.067393	5	6
MODEL1	y	-2.61931	0.851666			0.034198		2	3
MODEL1	y	-3.87832	0.853814			0.032792	0.112806	3	4
MODEL1	y	-2.75507	0.851955		0.031521	0.034126		3	4
MODEL1	y	-8.35286	0.868901	1.408538				2	3
MODEL1	y	-3.84779	0.853746		-0.00833	0.032805	0.113285	4	5
MODEL1	y	-9.44756	0.870545	1.340372			0.122679	3	4
MODEL1	y	-9.46328	0.869517	1.461572	0.217565			3	4
MODEL1	y	-10.2107	0.870869	1.388287	0.169946		0.111001	4	5
MODEL1	y	-3.43169	0.890437					1	2
MODEL1	y	-5.21866	0.891244				0.164476	2	3
MODEL1	y	-3.8279	0.891049		0.093178			2	3
MODEL1	y	-5.33601	0.891448		0.032588		0.162523	3	4
MODEL1	y	38.61291		3.79419		0.120022		2	3
MODEL1	y	40.17681		3.881315		0.1218	-0.19801	3	4
MODEL1	y	39.69234		3.734643	-0.22107	0.120477		3	4
MODEL1	y	40.79698		3.838718	-0.1425	0.122007	-0.18844	4	5
MODEL1	y	55.54029				0.131352		1	2
MODEL1	y	57.69239			-0.58465	0.132084		2	3
MODEL1	y	56.49715				0.132351	-0.09704	2	3
MODEL1	y	58.23513			-0.56153	0.132711	-0.06367	3	4
MODEL1	y	41.95079		4.668127				1	2
MODEL1	y	42.27935		4.688507			-0.04032	2	3
MODEL1	y	41.93931		4.668718	0.002324			2	3
MODEL1	y	42.19026		4.694394	0.020352		-0.04173	3	4
MODEL1	y	65.27434			-0.43584			1	2
MODEL1	y	62.57953					0.100695	1	2
MODEL1	y	64.09329			-0.48448		0.129944	2	3

Degree of Freedom	SSE	R - Square	RMSE	Adjusted R - Square	CP	AIC	BIC	SBC
106	34755.75	0.798435	18.10758	0.79273	2.271609	641.1182	643.5516	651.9201
105	34691.76	0.798806	18.17684	0.791142	4.079626	642.9155	645.4778	656.4179
105	34714.9	0.798672	18.18291	0.791002	4.149057	642.9888	645.5443	656.4912
104	34665.21	0.79896	18.25704	0.789295	6	644.8313	647.5169	661.0342
107	35862.82	0.792015	18.30756	0.788127	3.592982	642.5674	644.7006	650.6688
106	35717.57	0.792857	18.35642	0.786994	5.157207	644.1209	646.3318	654.9229
106	35860.97	0.792025	18.39323	0.786139	5.587426	644.5617	646.7401	655.3636
107	36287.58	0.789551	18.41565	0.785618	4.867287	643.8625	645.9247	651.964
105	35717.45	0.792858	18.44359	0.784966	7.156827	646.1206	648.3845	659.623

106	36115.38	0.79055	18.45836	0.784622	6.350677	645.3393	647.4606	656.1412
106	36200.83	0.790054	18.48018	0.784112	6.607027	645.5993	647.7015	656.4012
105	36064.01	0.790848	18.53286	0.78288	8.196558	647.1827	649.3489	660.6851
108	37583.94	0.782033	18.65474	0.780015	6.756529	645.7237	647.6247	651.1246
107	37268.96	0.78386	18.66301	0.77982	7.811564	646.7979	648.7002	654.8994
107	37567.69	0.782127	18.73766	0.778055	8.707781	647.6761	649.5308	655.7776
106	37267.02	0.783871	18.75035	0.777754	9.805734	648.7922	650.6625	659.5941
107	133505.8	0.225737	35.32305	0.211265	296.5341	787.1562	783.7517	795.2576
106	133061.3	0.228315	35.43014	0.206475	297.2005	788.7893	783.9441	799.5912
106	133416.5	0.226255	35.4774	0.204356	298.2662	789.0826	784.2293	799.8845
105	133025.2	0.228524	35.59363	0.199134	299.0923	790.7595	784.4663	804.2619
108	143267.6	0.169123	36.42186	0.16143	323.8209	792.9188	790.8352	798.3198
107	142626.3	0.172842	36.50967	0.157382	323.8969	794.4253	790.8639	802.5268
107	143159.6	0.169749	36.57787	0.154231	325.4969	794.8359	791.2658	802.9373
106	142580.8	0.173106	36.67563	0.149704	325.7604	796.3902	791.3438	807.1922
108	157324.2	0.087602	38.16682	0.079154	365.9925	803.2142	800.97	808.6152
107	157305.6	0.08771	38.34248	0.070658	367.9366	805.2012	801.4233	813.3026
107	157324.2	0.087602	38.34475	0.070548	367.9925	805.2142	801.4361	813.3157
106	157304.9	0.087715	38.52283	0.061895	369.9343	807.2007	801.889	818.0026
108	172072.6	0.00207	39.91572	-0.00717	410.2393	813.071	810.6849	818.472
108	172311.4	0.000685	39.9434	-0.00857	410.9557	813.2236	810.8353	818.6245
107	171880.3	0.003185	40.0794	-0.01545	411.6626	814.9481	810.9902	823.0495

جدول (٣) : نسبة اختيار النموذج الحقيقي من بين جميع نماذج الانحدار الممكنة لمعيار AIC، BIC ومعيار C_P

The real model	AIC	BIC	C_P
Model(1) X_1, X_2, X_4	13 %	13 %	13 %
Model(2) X_1, X_4	26 %	26 %	26 %
Model(3) X_1, X_2, X_4, X_5	6 %	6 %	6 %
AVERAGE	15 %	15 %	15 %

جدول (٤) : نسبة اختيار النموذج الحقيقي من بين جميع نماذج الانحدار الممكنة لمعياري RMSE و $\overline{R}^2$

The real model	RMSE	$\overline{R}^2$
Model(1) X_1, X_2, X_4	13 %	13 %
Model(2) X_1, X_2, X_4, X_5	6 %	6 %
Model(3) X_1, X_2, X_3, X_4	6 %	6 %
AVERAGE	8.33 %	8.33 %

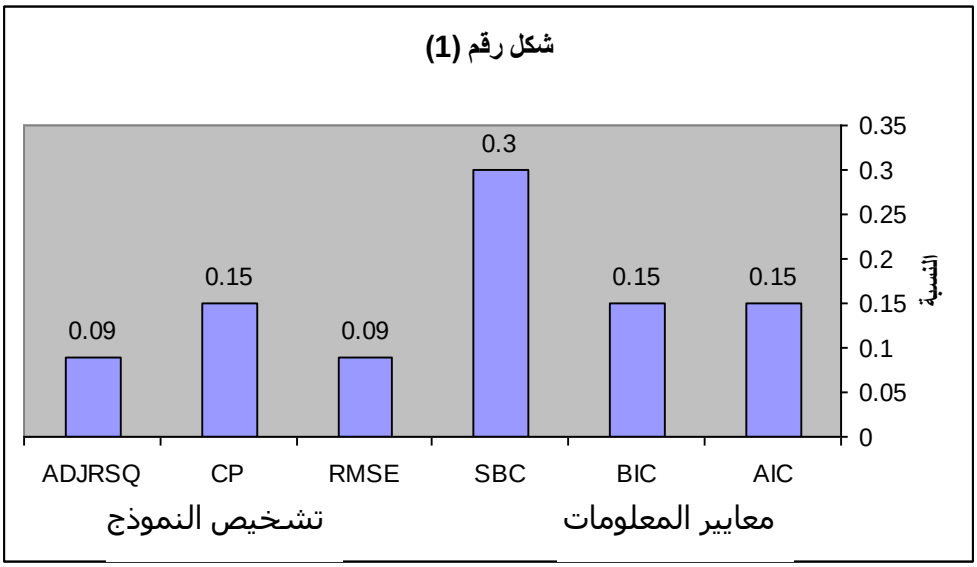
جدول (٥) : نسبة اختيار النموذج الحقيقي من بين جميع نماذج الانحدار الممكنة لمعيار SBC

The real model	SBC
Model(1) X_1, X_4	26 %
Model(2) X_1	51 %
Model(3) X_1, X_2, X_4	13 %
AVERAGE	30 %

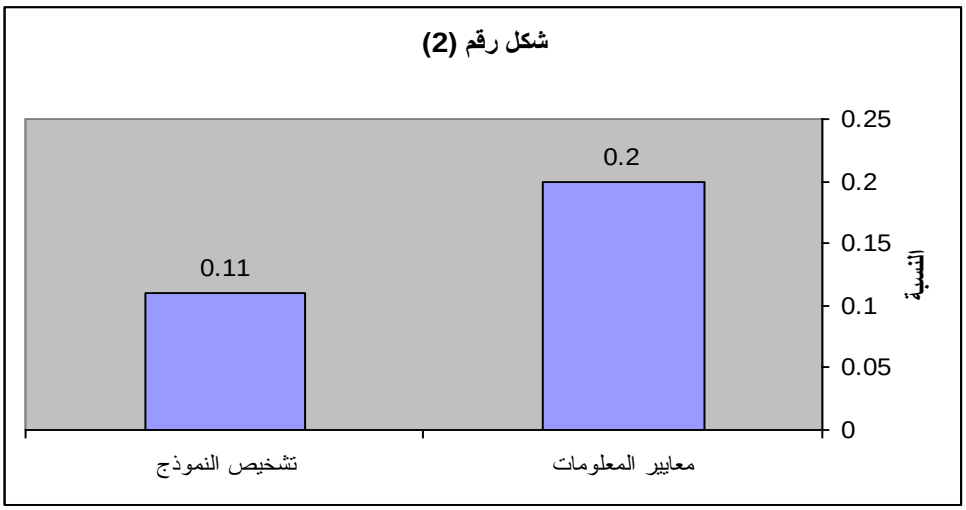
نلاحظ من الجداول (٣)، (٤) و (٥) أعلاه بان اعلى نسبة لاختيار النموذج الأفضل هو استخدام معيار SBC وبالتالي فهو يعتبر من أفضل المعايير المدروسة في هذا البحث ، تليه معايير AIC، BIC ومعيار C_P.

وكما موضح بالاشكال التالية:

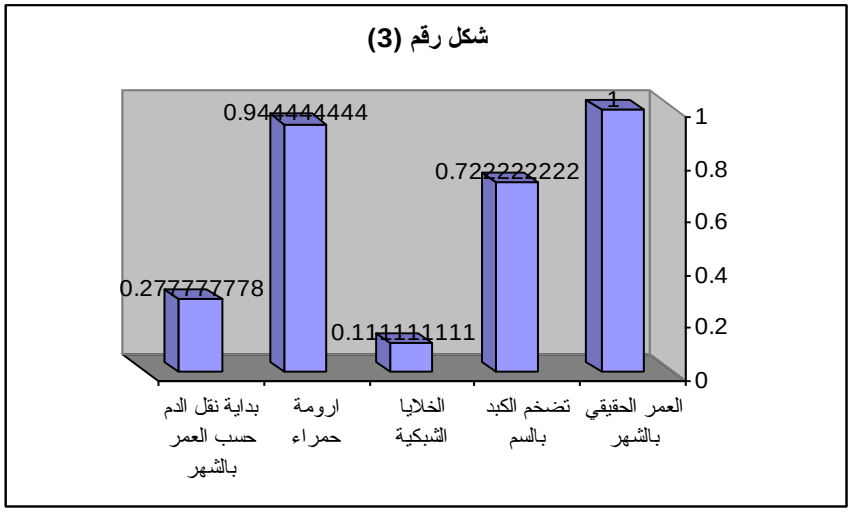
شكل رقم (1)



شكل رقم (2)



شكل رقم (١) و (٢) هما مدرجان يمثلان نسبة كل من طرق معايير المعلومات وطرق تشخيص النموذج لاختيار النموذج الحقيقي (الأفضل) من بين ٣١ نموذج



شكل رقم (٣): يمثل نسبة كل متغير في تأثيره على العمر الحقيقي بالعظم ولجميع المعايير المدروسة وفقا للنموذج الأفضل

الا

١. من خلال التحليل الإحصائي للبيانات نلاحظ بان هناك ثلاثة عوامل تؤثر بشكل مباشر على العمر الحقيقي بالعظم والاصابة بالمرض وهي (العمر الحقيقي بالشهر ، تضخم الكبد والارومة حمراء) مع التركيز على عاملين مهمين نلاحظ ظهورهما في جميع المعايير المحسوبة وهما (العمر الحقيقي بالشهر والارومة الحمراء).

٢. نلاحظ من الأشكال (١) و (٢) بان استخدام طرق معايير المعلومات (AIC,BIC,SBC) لاختيار أفضل نموذج انحدار خطي متعدد تعتبر أفضل من طرق تشخيص النموذج (, RMSE C_p ADJRSQ,).

٣. إن معيار (SBC) يعتبر أفضل المعايير المدروسة حيث حصل على نسبة (0.30) في اختيار النموذج الأفضل للبيانات يليه ثلاثة معايير (C_p AIC,BIC,) بنسب متساوية وهي (١٥ %) في اختيار النموذج الأفضل.

٤. نوصي باستخدام معايير أخرى مثل معيار NIC (New Information Criteria) ومقارنته مع المعايير التي تطرقنا إليها في بحثنا هذا.

المصادر :

1) Beal, D.J., (2007). Information Criteria Methods in SAS for Multiple Linear Regression Models. Proceedings of the Fifteenth Annual Conference of the South East SAS Users Group, Hilton Head, SC.

2) Burnham, K.P., and Anderson, D.R. (2002). Model selection and multimodel inference: A practical information-theoretic approach, 2nd-ed. Springer-Verlag, New York.

3) Draper, N.R., and Smith, H. (1981). Applied Regression Analysis. 2nd-ed. Wiley, New York.

4) Hocking, R.R., (1976). The analysis and selection of variables in linear regression. Biometrics 32, pp. 1-49.

5) Kundu, D., and Murali, G., (1996). Model selection in linear regression. Computational Statist. & Data Analysis 22, pp. 461-469.

6) Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2004). Applied linear statistical models. 4th-ed. McGraw Hill, Chicago, IL.

7) Linhart, H., and Zucchini, W. (1986). Model Selection. Wiley, New York.

8) Mallows, C. L. (1973). "Some Comments on Cp", Technometric, Vol. 15,No.4, pp 661-675.

9) Sparks, R. S.; Zucchini, W. and Coutsourides, D. (1985). "On Variable Selection in Multivariate Regression", Commun Statistis-Theor. Meth., 14(7), 1569-1587.

10) Stauffer, H.B., (2008). Contemporary Bayesian and Frequentist Statistical research methods for natural resource scientists. Wiley, New York.

11) Thompson, M.L., (1978a). Selection of variables in multiple regression. Part I, Int. Statist. Rev. 46, pp. 1-19.

12) Thompson, M.L., (1978b). Selection of variables in multiple regression. Part II, Int. Statist. Rev. 46, pp. 126-146.