استخدام أسلوب ألـ Cross –Validation في نموذج انحدار كاما

م. زكريا يحيى الجمال* م.م. إسراء عبد الجواد صالح** م.م. مازن غانم العناز***

المستخلص

يهدف هذا البحث إلى استخدام أسلوب الـ cross–validation والذي يكون على أساس تقسيم بيانات الدراسة إلى k من المجاميع الجزئية حيث يتم اعتبار إحداها كمجموعة اختبار إلى باقي المجاميع التي سوف تستخدم لغرض تكوين النموذج . تم استخدام هذا الأسلوب في نموذج انحدار كاما عندما يكون متغير الاستجابة يتبع توزيع كاما . لا حظنا من خلال نتائج المحاكاة بأنه كلما تغيرت قيمة معلمة الشكل اختلفت قيمة k الملائمة ولإحجام عينات مختلفة .

الكلمات الدالة: الانحدار الخطي ، انحدار كاما، اسلوب الـ Cross-Validation.

Abstract

The aim of this paper is to use the cross-validation procedure by splitting the data into k subgroups, which one of them will be used as a test to the other group which construct the model. We use this procedure in gamma egression model when the response variable has gamma distribution. We conclude that from simulation that the value of k changes when the shape parameter also changes for different sample sizes.

1ـ المقدمة

يعتبر تحليل الانحدار من أكثر الطرق الإحصائية استخداما في مجالات العلوم والهندسة وغيرها حيث يستخدم لتحليل العلاقة بين متغير توضيحي واحد أو أكثر ومتغير استجابة على هيئة معادلة. إن عملية تقدير المعلمات في نموذج الانحدار تفرض بان يكون متغير الاستجابة ذو توزيع طبيعي ،في كثير من التطبيقات يفشل توفر هذا الفرض وأحد الأساليب التي يمكن استخدامها عندما يكون متغير الاستجابة ذو توزيع كاما هو انحدار كاما. غالبا ما يكون الهدف من تكوين النموذج الذي يلائم البيانات قيد الدراسة هو التنبؤ بقيم جديدة وعليه ولغرض تقييم النموذج من ناحية التنبؤ عادة ما يتم تقدير خطأ التنبؤ (Fushiki,2011).احد أساليب تقدير خطأ التنبؤهوأسلوب الـ cross-validation والذي يقوم على أساس تقسيم بيانات الدراسة إلى k من المجاميع الجزئية والتي تكون متساوية أو قريبة من التساوي بحجم البيانات حيث يستبعد مجموعة واحدة في كل مرة لتمثل مجموعة اختبار للنموذج الذي سوف يتم تكوينه من باقي المجاميع والتي هي (k-1) .يعاد تكرار هذا الأسلوب لـ k من المرات (أي بعدد قيمة k) ،في كل مرحلة تكرار يتم حساب الخطأ ثم يتم حساب خطأ التنبؤ باستخدام Cross-Validation عن طريق حساب المعدل العام للأخطاء الخاصة بـ k من المجموعات .

لقد أولى الكثير من الباحثين اهتماما واسعا في أسلوب الـCross-Validation فمنهم من قارن أساليبه المتعددة (Burman,1989) وآخرين استخدموا هذا الأسلوب في عملية اختيار أفضل نموذج (Zhang,1993) ، (Venter and Snyman,1995). ومنهم من وضح خصائصه في الانحدار الخطي (Bunk and Droge, 1984) و(Picard and Cook,1984).

* كلية علوم الحاسوب والرياضيات /قسم الاحصاء والمعلوماتية
** كلية علوم الحاسوب والرياضيات /قسم الاحصاء والمعلوماتية
*** كلية علوم الحاسوب والرياضيات /قسم الاحصاء والمعلوماتية

يهدف بحثنا هذا الى استخدام اسلوب الـCross-Validation في تقييم نموذج انحدار كاما عن طريق استخدام المحاكاة ولإحجام عينات مختلفة .يشتمل هذا البحث على خمسة مباحث حيث اشتمل المبحث الثاني على مقدمة حول نموذج انحدار كاما في حين اشتمل المبحث الثالث على أسلوب الـ Cross-Validation أماالمبحث الرابع فقد احتوى على تجارب المحاكاة ثم جاءت الاستنتاجات في المبحث الخامس.

2ـ نموذج انحدار كاما

يعبر عن نموذج الانحدار الخطي بالشكل الأتي :

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1,2...,n \ , j = 1,2,...,k \qquad .....(1)$$

حيث أن المتغير العشوائي ε_i له توزيع طبيعي بوسط مقداره صفر وتباين مساوي إلى σ_ε^2 إن النموذج (1) يفرض بان تباين متغير الاستجابة يكون ثابت كدالة لوسط متغير الاستجابة (Faraway, 2006) .
يفرض نموذج انحدار كاما بان متغير الاستجابة له توزيع كاما بدالة كتلة احتمالية تعرف بالشكل الأتي:

$$f(y;\nu,\lambda) = \frac{\lambda^\nu}{\Gamma\nu} y^{\nu-1} e^{-\lambda y} \qquad y>0 \qquad ..... (2)$$

حيث ان $\lambda > 0$ تمثل معلمة القياس و $\nu > 0$ بدورها تمثل معلمة الشكل .
في نموذج انحدار كاما يكون معامل الاختلاف ثابت وبالاعتماد على مفهوم النموذج الخطي المعمم وبفرض أن ($\mu = \nu / \lambda$)فيمكن كتابة المعادلة (2)بالشكل الأتي :

$$f(y) = Exp\{\frac{y(-\frac{1}{\mu}) - \log(\mu)}{\frac{1}{\nu}} + \nu\log(\nu) + (\nu-1)\log(y) - \log(\Gamma\nu)\} \qquad .....(3)$$

(Uusipaikka,2009). وعليه فان معادلة انحدار كاما وحسب مفهوم النموذج الخطي المعمم تصبح بالشكل الرياضي الأتي :

$$E(y_i) = Exp\{\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}\} \qquad .....(4)$$

3ـ أسلوب ألـ Cross- Validation

إن أسلوب ألـ Cross-Validation هو طريقة لتقييم النموذج الإحصائي والذي يكون أفضل من معيار اقل قيمة للبواقي حيث أن مشكلة استخدام البواقي تكمن بعدم إعطائها أي دليل حول مدى كفاءة النموذج عند التنبؤ لبيانات جديدة غير موجودة ضمن البيانات الأصلية (Fushiki,2011). يعتبر أسلوب ألـ Cross-Validation احد طرق التغلب على مثل هذه الحالات حيث انه لا يستخدم جميع البيانات قيد الدراسة في تكوين نموذج الانحدار .

في هذا الأسلوب عادة ما يستخدم مصطلحا مجموعة التدريب ومجموعة الاختبار حيث تعرف مجموعة التدريب بأنها مجموعة البيانات التي سوف تستخدم لغرض تكوين النموذج بينما مجموعة الاختبار هي مجموعة البيانات التي سوف تستخدم لغرض التنبؤ بالنموذج الذي تم تكوينه من مجموعة بيانات التدريب .

يقوم أسلوب ألـ Cross-Validation بتقسيم البيانات إلى k من المجاميع وعادة ما تكون متساوية أو قريبة من التساوي وان كل مجموعة سوف تحوي على بيانات للمتغير المعتمد وعلى بيانات المتغير أو المتغيرات المستقلة وهو ما يدعى بأسلوب ألـ k-fold Cross-Validation . عندئذ يتم تكوين النموذج من(k-1) من المجاميع والمجموعة المتبقية سوف تستخدم كمجموعة اختبار للتنبؤ بهذا النموذج عادة ما تكون قيمة k ضمن الفترة (3≤k≤10)(Bunk and Droge, 1984).

وعلى هذا

$$MSECV = \frac{1}{k}\sum_{i=1}^{k} D_i \qquad .....(5)$$

الأساس يعاد في كل مرة تكوين نموذج بإدخال مجموعة التدريب السابقة إلى مجموعة (k-1) وتسحب مجموعة جديدة أخرى للتنبؤ وهكذا إلى k من المرات ،في كل مرة يتم حساب قيمة خطأ التنبؤ العام لجميع المجاميع والذي يرمز له بـ MSECV حيث أن :

حيث أن

$$D = 2V\sum_{i=1}^{n}\{-\log(\frac{y_i}{\hat{\mu}_i}) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\} \quad .....(6)$$

حيث أن D يمثل مقياس الاختلاف (Deviance) في نموذج انحدار كاما (Jong and Heller,2008)

4. المحاكاة

تم دراسة أسلوب الـ Cross-Validation في نموذج انحدار كاما من خلال استخدام المحاكاة حيث قمنا بتوليد بيانات تتبع نموذج انحدار كاما من خلال فرض أن المتغير أو المتغيرات المستقلة تتبع التوزيع المنتظم في الفترة (0,1).ولإحجام عينات مختلفة (10, 20,30,50,100) ولثلاثة قيم لمعلمة الشكل vهي (1,5,10) حيث تم تكوين ثلاثة نماذج ،شمل النموذج الأول على متغير مستقل واحد (p=1) وبمتجه معلمات(β_0=0,β_1=1) أما النموذج الثاني فقد اشتمل على ثلاثة متغيرات مستقلة (p=3) وبمتجه معلمات (0,1,1,1) ،في حين اشتمل النموذج الثالث على خمسة متغيرات مستقلة (p=5) وبمتجه معلمات (0,1,1,1,1,1).والجداول (1)، (2)، (3)، (4) و (5) توضح نتائج المحاكاة.

جدول رقم (1)

يوضح نتائج المحاكاة لـ n=10:

	k	3	4	5
P=1	v=1	2.888	2.396	2.996
	v=5	4.357	4.265	4.212
	v=10	1.369	1.5656	1.6147
P=3	v=1	1.319	2.282	1.259
	v=5	6.353	4.828	5.406
	v=10	8.44	9.156	10.269
P=5	v=1	0.531	0.602	0.842
	v=5	14.503	9.056	6.244
	v=10	15.534	17.153	20.63

جدول رقم (2)

يوضح نتائج المحاكاة لـ n=20

	k	3	4	5	6	7	8	9	10
P=1	v=1	3.125	3.497	2.927	2.69	3.328	2.78	2.938	2.846
	v=5	3.454	2.75	2.371	2.957	2.718	2.91	2.72	2.77
	v=10	7.954	7.26	8.44	8.467	8.708	8.698	8.713	8.302
P=3	v=1	0.623	0.645	0.706	0.701	0.696	0.723	0.645	0.707
	v=5	10.726	10.047	11.98	9.92	12.22	12.821	10.45	12.29
	v=10	6.575	6.72	5.998	6.44	6.18	6.402	6.408	6.629
P=5	v=1	0.897	0.925	0.662	0.718	0.754	0.809	0.747	0.665
	v=5	6.265	9.163	6.842	7.508	7.123	6.685	7.183	7.162
	v=10	12.714	12.361	10.905	11.98	11.11	11.55	11.58	11.378

جدول رقم (3)

يوضح نتائج المحاكاة لـ n=30

	k	3	4	5	6	7	8	9	10
P=1	v=1	1.753	1.418	1.446	1.472	1.4723	1.5704	1.523	1.479
	v=5	5.734	6.11	5.56	6.43	6.26	6.10	6.02	5.781
	v=10	12.217	10.53	11.32	12.61	10.195	12.166	11.106	10.48
P=3	v=1	0.854	0.61	0.564	0.68	0.646	0.648	0.643	0.729
	v=5	7.31	6.45	5.81	5.91	6.15	5.82	6.11	6.064
	v=10	14.73	18.429	16.88	15.44	18.23	15.48	14.82	15.256
P=5	v=1	0.824	0.885	0.917	0.935	0.836	0.847	0.921	871.
	v=5	6.11	6.211	6.39	6.33	6.86	7.05	6.383	6.455
	v=10	13.411	12.4316	14.73	13.909	12.52	12.89	14.45	13.244

جدول رقم (4)
يوضح نتائج المحاكاة لـ n=50

	k	3	4	5	6	7	8	9	10
P=1	v=1	1.004	0.96	1.03	1.001	0.992	0.999	0.978	0.9924
	v=5	4.614	4.59	4.89	5.005	4.725	4.76	4.92	4.832
	v=10	7.06	6.21	6.29	6.44	6.368	6.328	6.359	6.415
P=3	v=1	0.604	0.621	0.606	0.601	0.641	0.632	0.609	0.611
	v=5	4.708	4.8	4.53	4.87	4.42	4.48	4.63	4.56
	v=10	12.22	11.17	11.343	11.752	11.346	11.763	11.512	11.84
P=5	v=1	0.587	0.559	0.631	0.539	0.54	0.581	0.525	0.548
	v=5	5.67	5.615	7.11	5.739	5.901	6.2	5.99	6.053
	v=10	9.94	11.342	10.03	9.93	9.914	10.125	10.196	10.059

جدول رقم (5)
يوضح نتائج المحاكاة لـ n=100

	k	3	4	5	6	7	8	9	10
P=1	v=1	1.22	1.296	1.273	1.265	1.253	1.228	1.219	1.224
	v=5	4.31	4.937	4.45	4.449	4.397	4.451	4.375	4.415
	v=10	9.529	9.214	9.323	9.083	9.022	9.082	9.0671	9.046
P=3	v=1	1.759	1.659	1.66	1.703	1.832	1.698	1.685	1.671
	v=5	5.692	5.823	5.852	5.945	5.829	6.091	5.757	5.743
	v=10	8.699	8.673	8.751	8.48	8.821	8.561	8.48	8.559
P=5	v=1	0.976	0.941	0.974	0.9416	0.9414	0.957	0.949	0.947
	v=5	4.658	4.794	4.645	4.774	4.647	4.781	4.726	4.731
	v=10	8.927	9.002	8.905	8.985	8.656	8.798	8.733	9.02

5- الاستنتاجات

1- من خلال الجداول (1)، (2)، (3)، (4)و(5) لاحظنا بأنه كلما تغيرت قيمة معلمة الشكل v اختلفت قيم k ولكافة أحجام العينات وكذلك لكافة النماذج الثلاثة .
2- لا يمكن إيجاد قيمة مثلى لعدد المجاميع الجزئية k عندما يكون حجم العينة مساوي إلى 10.
3- لا يمكن القول بأنه كلما زادت قيمة k زادت دقة نموذج انحدار كاما من خلال خطا التنبؤ حيث أعطت نتائج المحاكاة تباينات في عملية اختبار قيمة k المثلى ولكل تجربة ونموذج .

6- المصادر

1- Bunke, O., and Droge, B., 1984, "Bootstrap and Cross- Validation Estimation of the Prediction Error for Linear Regression Model ",The Annals of Statistics ,Vol.12,No.4, pp.1400-1424.

2- Burman, P. , 1989,"A Comparative Study of Ordinary Cross-Validation V-fold Cross-Validation and the Repeated Learning-Testing Methods" , Biometrika,Vol.76,No.3, pp.503-514.

3- Faraway, J.,J.,(2006),"Extending the linear model with R ,Generalized linear ,Mixed Effects and nonparametric Regression Models", Chapman & Hall/CRC, Florida.

4-Fushiki,T.,2011,"Estimation of prediction error by using k-fold cross-validation ",stat. comput. (21), pp.137-146.

5-Jong, P. and Heller ,G.,Z.,2008," Generalized Linear Models for Insurance Data", Cambridge University Press.

6-Picard ,R., R. and Cook, D., R.,1984," Cross-Validation of Regression Models" , Journal of the American statistical Association, Vol.79, No 387 , pp.575-583.

7-Uusipaika, E., 2009, "Confidence Intervals in Generalized regression Models", Chapman & Hall/CRC, Florida.

8-Venter, H., J. and Snyman ,J.,L.,J., 1995, "A Note on the Generalized Cross-Validation Criterion in Linear Model Selection ", Biometrika,Vol.82,No.1,pp.215-219.

9-Zhang, P., 1993, "Model Selection Via Multifold Cross-Validation ",The Annals of Statistics , Vol.21,No.1, pp.299-313.

……..
……
…….