Association des statisticiennes et statisticiens du Québec

Convergence

Volume XVII Numéro 1 **Mai 2012**

Ce document est une reconstitution du journal Convergence initialement publié sous la forme d'un regroupement d'articles sur le site web de l'association.

Le mot de la présidente

Par Nathalie Madore

Voilà enfin le printemps à nos portes. Si vous êtes comme moi, on se sent un peu revivre en cette période de l'année. On a le goût de sortir, de se débarrasser des vêtements noirs et sombres, de voir pousser les fleurs. Mais surtout, on se dit «À quand un nouveau colloque de l'ASSQ?».

C'est donc avec beaucoup de plaisir et de fierté que je vous invite à vous joindre à nous pour cet événement qui, chaque année, nous permet de joindre l'instructif, à l'utile, à l'agréable. En effet, une fois encore, l'activité se déroulera en trois temps : le colloque comme tel qui nous offrira quatre conférences, l'assemblée générale annuelle et enfin, le souper, qui sera précédé d'un cocktail.

L'après-midi sera très varié. Nous aurons le plaisir d'entendre quatre conférenciers :

- Claire Durand, de l'Université de Montréal, nous parlera de l'importance du choix des statistiques dans le cadre d'une étude;
- David Émond, représentant des étudiants au Conseil d'administration de l'ASSQ et étudiant à la maîtrise à l'Université Laval, nous montrera comment optimiser le regroupement des équipes de la LNH au sein des divisions;
- Patrice Mathieu de Statistique Canada fera le point sur le recensement et l'Enquête nationale auprès des ménages de 2011;
- Christian Genest, de l'Université McGill, récipiendaire de la Médaille d'or de la SSC en 2011, nous offrira la conférence préparée à ce titre pour le congrès de la SSC.

Un programme des plus intéressants que vous ne voudrez certainement pas manquer. L'assemblée générale annuelle est toujours une occasion privilégiée pour les membres de s'informer des activités du Conseil d'administration et de proposer des activités, changements, améliorations souhaités. Lieu d'échange par excellence, elle vous permet de vous impliquer concrètement dans la vie de votre association. Chaque membre devrait se faire un devoir de tenter d'assister à cette rencontre. Celle de cette année sera de plus l'occasion de discuter de nombreuses modifications aux statuts proposées par le Conseil mais surtout au règlement de l'ASSQ. L'information à ce sujet vous sera envoyée en même temps que la convocation à l'assemblée. Prenez le temps de bien lire ces documents qui constituent les fondements de notre association. Les membres qui ne pourront être présents à l'assemblée peuvent toujours nous faire part de leurs commentaires, par écrit, avant la rencontre, et nous pourrons en tenir compte lors des débats.

Retenez donc cette date : le vendredi 8 juin, à l'Aquarium du Québec, à Québec. Nous regardons présentement la possibilité d'offrir un accès au site de l'Aquarium avant notre activité pour ceux qui seraient intéressés mais cela reste encore à confirmer. J'espère vraiment vous y voir en très grand nombre.

Nathalie Madore,
Présidente

Les échos du C.A.

Par Lise Charette

C'est sur un nouveau clavier que continue cette chronique potentiellement «assourdissante», si on se fie au titre. Pas trop quand même, je vais faire de mon mieux. Est-ce qu'il y a écho… écho… écho… ? Ma prétention ici est de continuer à vous transmettre ce qui se trame chez vos humbles serviteurs, quelles sont les aspirations, les préoccupations et les actions de vos représentants pour 2012.

Titre de secrétaire de l'Association oblige, mettons d'abord quelque chose au clair. CA ou C.A.? Nature ou givré? À moins que ce ne soit l'inverse… Mon prédécesseur utilisait C.A. (nature?), le prédécesseur de mon prédécesseur utilisait CA (givré?)… Sachez que selon la liste des abréviations de la Banque de dépannage linguistique de l'Office de la langue française, les deux formes sont mentionnées. À défaut d'être rigoureux, nous pouvons donc simplement dire que nous montrons nos différents côtés…

Un nouveau C.A. (nature?) implique un nouvel horaire à mettre en place et à des formalités administratives à régler avant de prendre le rythme de croisière. Trois rencontres plutôt matinales ont eu lieu à Québec dans les bureaux du siège social de l'Association, avec les représentants de l'extérieur de Québec en conférence téléphonique. Nathalie Madore, notre présidente, a dû sortir de sa zone de confort mais elle est toujours à l'heure. Jean-François Plante, notre responsable des communications, a même pu changer la couche de son garçon, tout ça en ligne, sans que nous n'interrompions la discussion, malgré les quelques gazouillis inattendus qui nous ont fait sourire. La preuve est maintenant faite qu'il n'y a pas que les femmes qui ont des aptitudes à faire du multitâches! Tout ça pour dire que nous avons bien hâte de tous se rencontrer en personne une première fois, ça s'en vient.

Comme vous avez pu le constater, le projet de refonte du Convergence est venu à terme en février. Denis Talbot, notre édimestre, a fait du très bon travail pour mettre le tout en forme, après avoir surmonté quelques embûches.

L'autre projet pour amorcer 2012 est le Colloque annuel qui est bel et bien en branle pour le 8 juin à l'Aquarium de Québec. La présidente de l'ASSQ nous donne d'ailleurs plus d'information à ce sujet dans Le mot de la présidente. Il n'y a cependant pas encore de Jeudis de l'ASSQ à l'horaire. Nous faisons appel à vous pour des idées, des contacts, votre intérêt à présenter ce qui est passionnant dans votre travail ou simplement ce qui mériterait d'être connu ou de s'y attarder comme statisticien ou comme individu. Le C.A. a par ailleurs approuvé en mars le formulaire *Offre de support à l'organisation d'activités liées à la statistique*, formulaire disponible sur le site de l'Association et ayant déjà pris vie par une demande du CASUL pour la 25e Journée de la Statistique. Il y a certainement des activités intéressantes dans vos milieux de travail auxquelles des membres pourraient se greffer. Osez!

Un autre sujet qui nous importe est de finaliser la production du «dépliant» sur la Carrière en statistique qui deviendra un outil pour les membres afin de présenter leur profession, projet qui va bon train et pour lequel vous aurez des nouvelles sous peu.

Mentionnons finalement la partie formelle vécue par tout C.A. soit la révision de ses règlements et statuts, pour lesquels nous vous transmettrons des propositions à débattre lors de la prochaine assemblée générale annuelle. Comme membre, vous pouvez évidemment faire des propositions sur ces règlements et statuts. Ce n'est pas le sujet le plus passionnant, mais si le cœur vous en dit, au risque de me répéter, osez !

Lise Charette
Secrétaire

Les commandites ne font pas toutes scandale…

Par Lise Charette

L'ASSQ offre à ses membres une aide financière ou logistique à l'organisation d'un évènement qui va dans la foulée de sa mission de promotion de la statistique. Il peut s'agir de représentation dans une journée carrière, une conférence sur la profession, un colloque ou une conférence qu'organise une organisation et auxquels des membres de l'ASSQ pourraient se greffer, une activité que vous aimeriez organiser pour les membres, etc. Il y a plein de possibilités… Cette offre vous interpelle ? Nous vous invitons à remplir le Formulaire de demande de contribution de l'ASSQ à l'organisation d'un évènement et à nous le faire parvenir. N'hésitez pas à contacter un membre du C.A.!

Le mot du registraire

Par Eric Lacroix

Compte rendu de la campagne d'adhésion et de renouvellement

Voici quelques nouvelles concernant les adhésions à l'Association des statisticiennes et statisticiens du Québec. En date du 30 avril, seulement 83 des 152 membres sollicités avaient renouvelé leur adhésion à l'ASSQ pour 2012. Par ailleurs, à cette date, SOM, l'Université du Québec, l'Université Laval et SolutionStat avaient renouvelé leur adhésion à titre de membres institutionnels.

Au total, en comptant les nouvelles inscriptions (5) et les membres délégués par les membres institutionnels, l'ASSQ comptait 88 membres actifs.

Au cours du mois d'avril, un rappel a été envoyé aux retardataires. À ceux qui ne l'ont pas encore fait, merci de nous expédier votre formulaire de renouvellement le plus tôt possible.

Eric Lacroix,
registraire de l'ASSQ

Un conte de Noël

Par Nathalie Madore

Il était une fois une petite statisticienne qui errait dans les dédales d'Internet à la recherche de renseignements intéressants. Le web foisonnait de données de toutes sortes et l'internaute espérait bien trouver une information qui lui permettrait d'améliorer ses connaissances de la population.

C’est alors qu’elle fut attirée par un titre des plus surprenants : «La moitié des Canadiens prévoient acheter leurs cadeaux en ligne!» Sachant qu’elle-même n’avait pas du tout l’intention de remplacer les séances de magasinage par des visites de sites web des magasins, elle s’étonna de ce pourcentage si élevé. Curieuse, elle poursuivit donc la lecture.

Elle apprit ainsi qu’un sondage démontrait que «46 pour cent des consommateurs iront sur Internet pour acheter leurs cadeaux des Fêtes cette année, comparativement à 41 pour cent l'année dernière». Une première question effleura l’esprit de la petite statisticienne : est-ce qu’aller sur Internet pour acheter un cadeau est la même chose qu’acheter un cadeau sur Internet? Bien que la petite statisticienne avait elle aussi l’intention, avant de se rendre en magasin, d’aller sur Internet pour comparer des prix, vérifier des disponibilités, elle n’envisageait toutefois pas d’y faire ses achats. Quelle était donc la question réellement posée aux répondants? Malheureusement, l’article ne le disait pas. Pourtant, il s’agissait là d’une information très importante. Sachant que, pour le Québec, le CEFRIO (centre facilitant la recherche et l’innovation dans les organisations) estime à près de 80 % la proportion de gens qui utilisent régulièrement Internet et à moins de 25 % celle des cyberacheteurs, on pourrait facilement croire que les données ne parlaient pas ici d’achats sur Internet, mais bien d’utilisation d’Internet dans le cadre du magasinage des Fêtes. Mais on ne peut en être certain.

La petite statisticienne lu la suite de l’article. Elle trouva alors dans le dernier paragraphe, celui que si peu de gens prennent la peine de lire, une autre explication possible à ce résultat si étonnant : «Le sondage a été effectué en ligne du 17 au 20 octobre 2011, auprès d'un échantillon de 1508 Canadiens choisis parmi l'échantillon Internet de la firme Léger Marketing, LégerWeb». L’inscription dans un panel Internet se faisant sur base personnelle, la petite statisticienne comprit bien que les répondants ne pouvaient absolument pas être déclarés représentatifs des Canadiens à prime abord. Certaines corrections pourraient bien entendu être appliquées, mais l'article n'en fait pas mention. Elle put donc conclure que près de la moitié des internautes suffisamment aguerris pour vouloir s’inscrire dans un panel de répondants à des sondages web pensaient faire leurs achats en ligne à Noël, quoique, sans connaître la question exacte, il restera toujours un doute…

La morale de cette histoire : ne jamais prêter foi à un titre sans prendre la peine de lire l’article jusqu’au tout dernier paragraphe.

Nathalie Madore

Au sujet du « p-value »

Par Denis Talbot

Dans cet article, je vais effectuer un bref résumé d'un essai intéressant sur lequel je suis tombé il y a quelque temps déjà. Cet essai se veut sévère envers une interprétation courante du «p-value» et suggère une manière alternative d'interpréter les tests statistiques. Il s'agit de l'article Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian de Stuart H. Hulbert et Celia M. Lombardi publié dans Annales Zoologici Fennici en 2009.

L'utilisation courante du «p-value»

Les auteurs dénoncent vivement une interprétation fréquente du «p-value», le fameux seuil observé des tests statistiques qu'on utilise si régulièrement. En particulier, les auteurs en ont contre le fait de fixer à l'avance un seuil à partir duquel on décidera que le résultat est «significatif» ou est «non significatif».

D'une part, ils argumentent que cette dichotomisation du résultat est inutile et contribue à une perte d'information. En effet, une pratique courante veut qu'on fixe un seuil alpha avant d'effectuer le test, généralement 5%. Si le seuil observé du test est inférieur à ce seuil alpha, on interprète le résultat comme significatif, alors que si le seuil observé est plus grand que ce seuil, on interprète le résultat comme non significatif. Pourtant, le niveau de preuve en faveur de l'hypothèse alternative est sensiblement le même lorsque le

seuil observé est de 0.049 que lorsque le seuil observé est de 0.051. Alors pourquoi donner une interprétation diamétralement opposée à ces deux résultats?

D'autre part, cette pratique de dichotomisation amène couramment, selon eux, une mauvaise interprétation des résultats. Plusieurs interprètent effectivement, à tort, un résultat non significatif comme étant une preuve en faveur de l'hypothèse nulle. Pourtant, on est averti dès les premiers cours de bases en statistique qu'un résultat non significatif ne constitue pas une acceptation de l'hypothèse nulle, mais simplement un non-rejet de l'hypothèse nulle. La nuance est légère, mais importante. Le non-rejet de l'hypothèse devrait en fait être vu comme une remise à plus tard du jugement sur l'hypothèse testée. La dichotomisation facilite cependant l'interprétation erronée de l'acceptation de l'hypothèse nulle.

Une interprétation alternative

Les auteurs suggèrent alors une utilisation alternative du fameux «p-value». Dans l'approche suggérée, les chercheurs sont appelés à ne pas fixer à l'avance un seuil alpha et à rapporter les seuils observés exacts, (donc à éviter de simplement rapporter des résultats comme significatifs ou non-significatifs). Ils appellent par ailleurs à interpréter le «p-value» comme une valeur sur un continuum du niveau de preuve observé à partir des données pour une hypothèse donnée. On interprétera alors un seuil observé très faible comme étant une preuve forte en faveur de l'hypothèse alternative, un seuil observé modéré comme étant une preuve modérée en faveur de l'hypothèse et un seuil élevé comme une absence de preuve en faveur de l'hypothèse alternative. Les auteurs mettent cependant en garde contre l'interprétation d'une absence de preuve comme étant une preuve de l'absence d'effet.

Cette interprétation du seuil observé selon un continuum, sans fixer de seuil alpha préalablement, peut sembler subjective. Les auteurs mentionnent cependant qu'étant donné que les seuils exacts sont rapportés, le lecteur est libre d'effectuer une interprétation différente des résultats puisqu'il dispose de toute l'information nécessaire pour le faire.

L'hypothèse de recherche et l'hypothèse statistique

Les auteurs discutent également du fait que l'interprétation porte trop fortement sur les résultats des tests concernant l'hypothèse statistique et insuffisamment sur l'interprétation par rapport à l'hypothèse de recherche. Ils proposent ainsi de supplémenter de façon courante les résultats des tests statistiques par des intervalles de confiance, des tailles d'effets standardisées, des tests de puissance statistique et d'autres outils. Selon eux, l'ajout de cette information complémentaire permettrait, par exemple, de voir que dans certains cas, bien que le résultat d'un test statistique donne une preuve forte en faveur de l'hypothèse alternative, le niveau de preuve en faveur de l'hypothèse de recherche est faible puisque l'intervalle de confiance est concentré sur des petites valeurs proches de zéro et que la taille d'effet standardisée est petite.

À l'opposé, on pourrait constater un seuil observé élevé du test statistique et constater sur la courbe de puissance statistique qu'une taille d'effet énorme aurait été nécessaire pour obtenir un seuil observé faible. La courbe de puissance permet donc de constater qu'un jugement concernant cette hypothèse ne peut pas du tout être porté à l'aide des données.

En conclusion

En conclusion, les auteurs proposent plusieurs changements, somme toute mineurs, dans la pratique courante. Évidemment, il serait difficile d'adopter du jour au lendemain l'approche qu'ils proposent. Ces pratiques, sans être radicalement différentes des pratiques courantes, diffèrent des normes utilisées dans plusieurs domaines. Par contre, la lecture de leur texte amène une réflexion sur notre propre pratique en tant que statisticiens et peut nous mener à incorporer certains éléments. Par exemple, on peut facilement rapporter les seuils observés exacts et interpréter les seuils observés entre 0.05 et 0.10 comme étant une tendance à vérifier dans des études ultérieures.

Denis Talbot

La chronique SAS

Par Sylvain Tremblay

Diviser pour régner avec PROC GLMSELECT

À l'époque de l'Empire romain, l'application de la politique extérieure du «divide et impera» a permis de contenir les peuples conquis et d'assurer la suprématie de Rome. En modélisation prédictive, nous pouvons nous inspirer de cette approche lors de la sélection des variables et de la vérification de l'adéquation du modèle dans le but de s'assurer de la puissance prédictive de notre modèle final. Pour ce faire, la procédure GLMSELECT du module SAS/STAT est toute indiquée.

Dans cette chronique, j'aimerais démystifier cette procédure peu connue et illustrer une de ses nombreuses applications: le partitionnement des données dans le processus de sélection d'un modèle linéaire dans un contexte prédictif.

La genèse d'une procédure

La procédure GLMSELECT est un croisement entre deux procédures SAS pour les modèles linéaires: REG et GLM. Ces dernières existent depuis fort longtemps mais ont chacune leurs limites. Par exemple, la procédure REG offre une variété de méthodes de sélection de modèles mais ne permet pas l'utilisation d'une instruction CLASS. À l'inverse, la procédure GLM permet l'utilisation d'une instruction CLASS, mais ne permet pas d'utiliser de méthodes de sélection de modèles.

La procédure GLMSELECT fait le pont entre REG et GLM en offrant de nombreuses fonctionnalités de personnalisation pour la sélection de modèle. En plus des méthodes pas à pas traditionnelles («forward», «backward» et «stepwise»), la procédure GLMSELECT supporte les nouvelles méthodes LAR («Least Angle Regression») et LASSO («Least Absolute Shrinkage and Selection Operator»). De plus, elle permet, grâce à son instruction MODELAVERAGE, d'utiliser des techniques de ré-échantillonnage comme le «bootstrap». La procédure prévoit également un résumé graphique du processus itératif de sélection du modèle.

Pour une méthode donnée de sélection de modèle, de nombreuses options sont offertes pour choisir le modèle final. On peut se baser sur un nombre maximal d'itérations, sur un critère statistique, sur la première itération qui comporte n effets, etc.

La sélection de modèle par le partitionnement des données

Bien connu des adeptes du data mining et des techniques d'apprentissage supervisé, le partitionnement des données consiste à diviser les données de départ en trois partitions disjointes : les données d'entraînement («training»), les données de validation et les données test.

Les modèles seront développés de façon itérative sur les données d'entraînement, l'adéquation de ces modèles à chaque itération se fera sur les données de validation afin d'identifier un modèle candidat et une comparaison finale des modèles candidats s'effectuera avec les données test afin de choisir le meilleur modèle.

Si on utilise par exemple une méthode de sélection pas à pas «forward», le meilleur modèle n'est pas celui que l'on obtient à la dernière étape du processus itératif mais plutôt le modèle qui va obtenir la meilleure performance sur les données de validation. Cette méthode a pour but de choisir un modèle candidat qui est parcimonieux et qui démontre une puissance prédictive sur de nouvelles données.

L'encadré 1 donne un exemple de cette approche avec la procédure GLMSELECT de SAS, en utilisant des données d'entraînement et de validation.

Encadré 1

```
ODS GRAPHICS ON;

PROC GLMSELECT

   DATA=biblio.donnees
   PLOTS=ASEPlot
   SEED=123;

   PARTITION fraction( validate=0.35);

   MODEL y = x1|x2|x3|x4|x5|x6|x7|x8|x9 @3
      /selection=forward(steps=100);

run;

ODS GRAPHICS OFF;
```

L'instruction PROC GLMSELECT contient les options suivantes :

- DATA pour identifier le fichier de données à utiliser;
- PLOTS pour spécifier les graphiques à produire (à utiliser avec ODS GRAPHICS ON). Ici, ASEPLOT demande qu'un graphique de l'erreur quadratique moyenne à chaque itération soit produit;
- SEED pour spécifier le nombre de départ qui va initialiser la production d'un nombre pseudo-aléatoire afin de pouvoir reproduire l'échantillonnage.

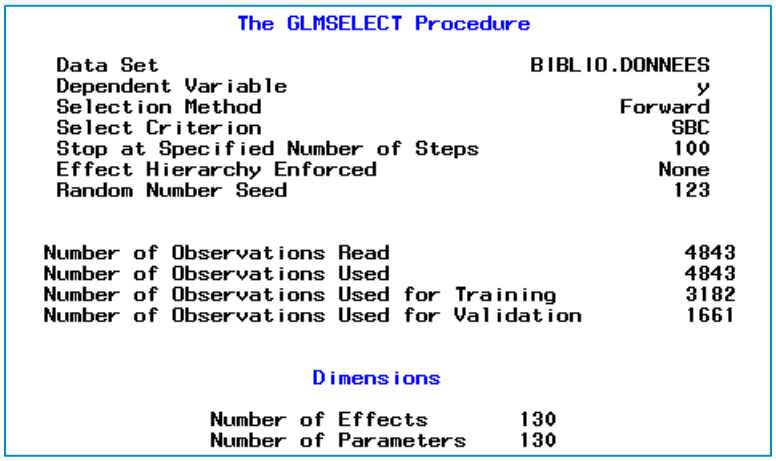
L'instruction PARTITION avec son option FRACTION va s'occuper de faire le partitionnement aléatoire des données en deux: 65% des données originales pour les données d'entraînement et 35% pour les données de validation. Au lieu de l'option FRACTION, il est possible d'utiliser l'option ROLEVAR pour spécifier une variable qui va servir à classifier les observations selon qu'elles doivent être utilisées comme données d'entraînement ou données de validation.

Si les données d'entraînement et de validation sont dans des fichiers de données SAS séparés, il est possible de se servir de l'instruction VALDATA= pour lire les données de validation au lieu de se servir de l'instruction PARTITION.

On retrouve ensuite l'instruction MODEL pour spécifier le modèle. Dans la notation, les variables sont séparées par le symbole | et l'opérateur @3 est utilisé pour inclure dans le modèle les effets principaux et toutes les interactions deux à deux et trois à trois. Vient ensuite l'option SELECTION=FORWARD. Cette méthode pas à pas de sélection de modèle débute avec un modèle ne contenant aucune variable et, à chaque itération, ajoute la variable la plus significative (en présence de celles déjà dans le modèle). En utilisant la sous-option STEPS=100, on force cette méthode à faire cent itérations, même si une solution optimale a été atteinte plus tôt.

Dans la sortie SAS de l'encadré 2, on constate que 3182 observations ont été utilisées pour l'entraînement du modèle et 1661 pour la validation, c'est-à-dire respectivement 65% et 35% des données.

Encadré 2



The GLMSELECT Procedure

Data Set	BIBLIO.DONNEES
Dependent Variable	y
Selection Method	Forward
Select Criterion	SBC
Stop at Specified Number of Steps	100
Effect Hierarchy Enforced	None
Random Number Seed	123

Number of Observations Read	4843
Number of Observations Used	4843
Number of Observations Used for Training	3182
Number of Observations Used for Validation	1661

Dimensions

Number of Effects	130
Number of Parameters	130

Dans le sommaire de la sélection «forward» (encadré 3), on retrouve pour chaque itération, l'effet qui est entré dans le modèle, le nombre total d'effets, deux statistiques d'adéquation sur les données d'entraînement (le critère Schwarz Bayésien («Schwarz Bayesian information Criterion») et l'erreur quadratique moyenne («Average Squared Error»)) ainsi que le «Validation ASE». Cette dernière est une statistique d'adéquation sur les données de validation.

Une note à la fin du sommaire indique que le processus de sélection s'est arrêté à l'étape 100. Par contre, un astérisque dans la colonne SBC indique que, pour les données d'entraînement, le modèle choisi selon ce critère serait celui de l'étape 15. De façon similaire, un astérisque dans la colonne Validation ASE indique que, pour les données de validation, le modèle choisi proviendrait de l'étape 11 si on utilise ce critère.

Encadré 3

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	SBC	ASE	Validation ASE
0	Intercept	1	16317.6159	168.2693	129.1468
1	x1	2	16167.6629	160.1172	120.9656
2	x1*x6	3	16067.3252	154.7542	116.2377
3	x3*x6	4	16021.1996	152.1410	115.5667
4	x2*x3*x5	5	16014.7247	151.4474	116.3123
5	x1*x2	6	16011.2034	150.8969	115.1203
6	x5*x6	7	15997.3189	149.8596	114.2577
7	x2*x6	8	15979.2539	148.6340	114.2175
8	x4*x5*x8	9	15973.9479	148.0107	113.9949
9	x5	10	15967.7761	147.3500	113.2011
10	x3*x9	11	15959.7723	146.6078	113.2061
11	x4*x8	12	15955.0731	146.0208	113.0196*
12	x1*x3*x7	13	15941.9121	145.0500	114.8933
13	x3*x4	14	15936.6828	144.4452	114.9276
14	x4*x6*x9	15	15930.7157	143.8096	114.6490
15	x1*x4*x8	16	15925.0077*	143.1885	115.0194
16	x1*x5	17	15926.1079	142.8754	114.6885
17	x4*x6*x8	18	15926.5254	142.5324	114.5737
18	x9	19	15927.9663	142.2360	114.9259
19	x4*x8*x9	20	15926.8350	141.8255	114.5846
20	x5*x9	21	15926.7262	141.4617	116.1585
...					
87	x3*x5*x6	88	16217.9765	130.8089	136.1172
88	x4	89	16222.4353	130.6607	136.3969
89	x2*x8	90	16227.2401	130.5269	137.1140
90	x2*x9	91	16231.5875	130.3745	137.0776
91	x3*x6*x9	92	16237.6258	130.2915	137.9677
92	x3*x5	93	16243.3950	130.1975	138.3479
93	x5*x6*x9	94	16249.8618	130.1321	138.2287
94	x1*x4*x6	95	16255.8179	130.0459	138.1969
95	x1*x9	96	16262.2150	129.9777	138.1180
96	x3*x8*x9	97	16268.4842	129.9044	138.6851
97	x3*x6*x7	98	16274.6963	129.8288	138.3431
98	x8*x9	99	16281.5895	129.7809	139.0483
99	x3*x4*x9	100	16288.4992	129.7338	139.3160
100	x1*x2*x8	101	16295.5380	129.6920	138.8570

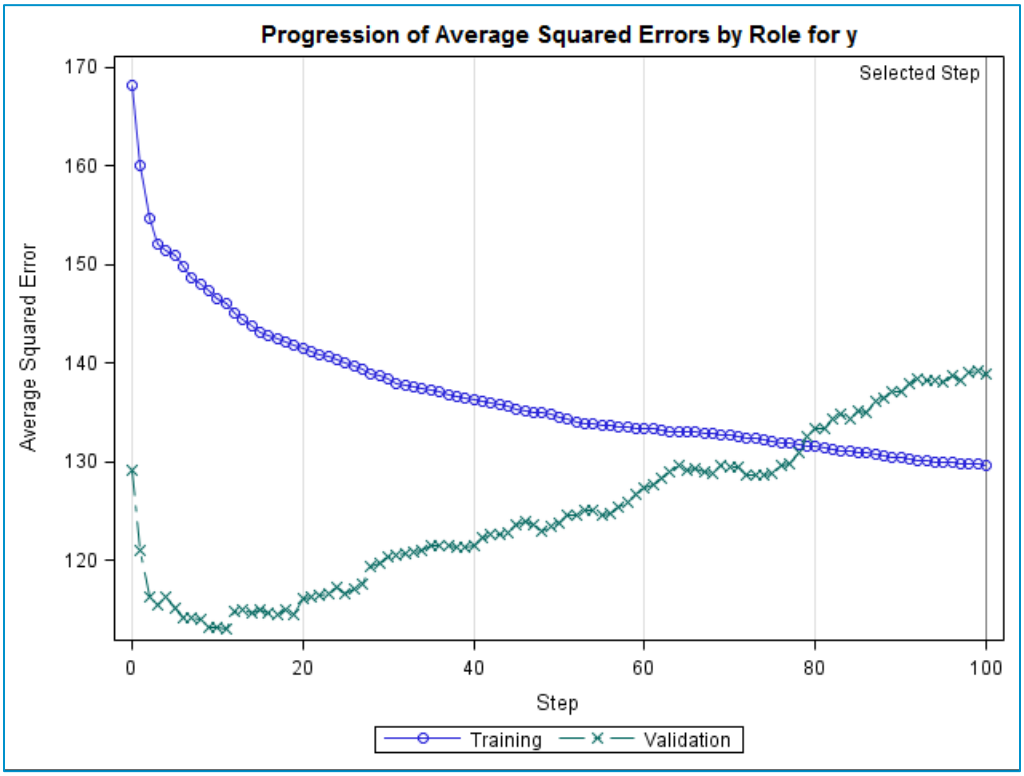
* Optimal Value Of Criterion

Selection stopped at the specified number of steps (100).

Cette information est résumée dans le graphique d'itérations (encadré 4) produit par l'option PLOTS=ASEPLOTS. On y retrouve la progression de l'erreur quadratique moyenne, à chaque étape du processus pas à pas, pour les données d'entraînement et les données de validation.

En examinant ce graphique, on constate que pour les données d'entraînement, plus le nombre d'étapes est grand (i.e. plus il y a de termes dans le modèle, plus il est complexe), meilleur l'ajustement car l'ASE baisse à chaque itération. Pour les données de validation, on observe cette même tendance initialement mais après la 12e étape, l'ASE se met à croître. Ceci est un exemple classique de surajustement.

Encadré 4



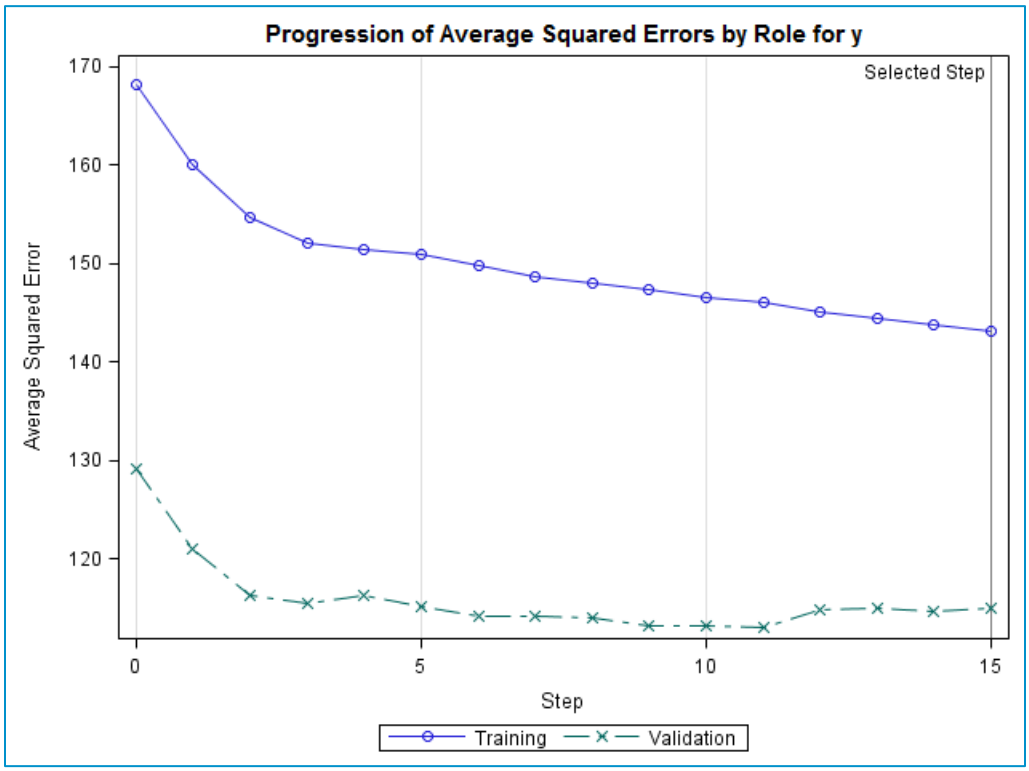
Au lieu de forcer GLMSELECT à faire cent itérations, on peut émuler la méthode de sélection de modèle «forward» de PROC REG en lui demandant de choisir le modèle, en se basant sur un seuil pour le «p-value» en utilisant l'option SLENTRY (voir l'encadré 5).

Encadré 5

```
MODEL y = x1|x2|x3|x4|x5|x6|x7|x8|x9 @3
   /selection=forward(SLENTRY=0.15);
```

On obtient le graphique d'itérations suivant (encadré 6). On remarque que le modèle choisi provient de l'étape 15. Cette sélection utilise les données d'entraînement et le seuil SLENTRY. Par contre, on remarque que pour les données de validation, le ASE atteint son minimum à l'étape 11.

Encadré 6



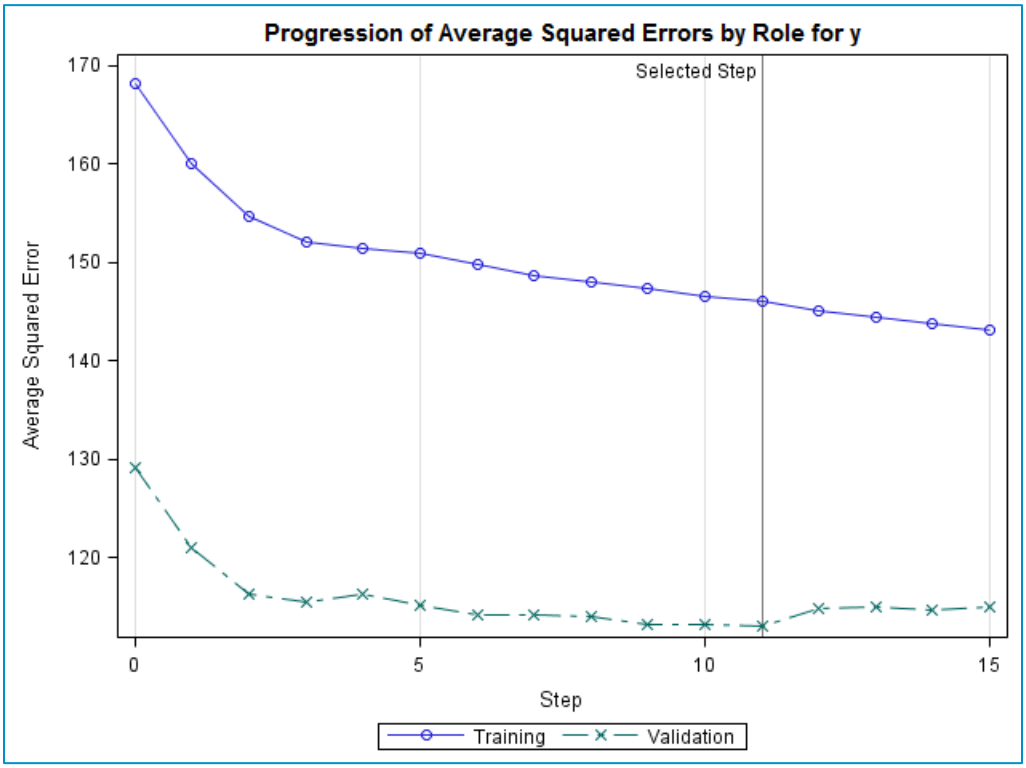
Pour demander à GLMSELECT de choisir le modèle qui a la meilleure performance (plus petit ASE) sur les données de validation, on n'a qu'à utiliser l'option CHOOSE=VALIDATE (encadré 7).

Encadré 7

```
MODEL y = x1|x2|x3|x4|x5|x6|x7|x8|x9 @3
   /selection=forward(CHOOSE=validate);
```

On remarque alors, dans le graphique d'itérations (encadré 8), que le modèle choisi est celui de l'étape 11, étape où le minimum de l'ASE est atteint sur les données de validation. Nous obtenons donc un modèle plus simple (que celui de l'étape 15, donné par l'approche traditionnelle sur les données d'entraînement) et qui exhibe une meilleure performance sur les données de validation, ce qui est une bonne nouvelle dans un contexte prédictif où l'on désire utiliser le modèle sur de nouvelles données.

Encadré 8



Pour déployer le modèle et produire des prédictions pour un nouvel ensemble de données, rien de plus simple! La procédure GLMSELECT est munie d'une instruction SCORE qui facilite le tout (encadré 9).

Encadré 9

```
PROC GLMSELECT

   DATA=biblio.donnees
   PLOTS=ASEPlot(stepaxis=number)
   SEED=123;

   PARTITION fraction( validate=0.35);

   MODEL y = x1|x2|x3|x4|x5|x6|x7|x8|x9 @3
      /selection=forward(CHOOSE=validate);

   SCORE DATA=biblio.nouvelles_donnees
         OUT=work.predictions;

run;
```

Conclusion

En conclusion, si vous travaillez avec des modèles linéaires dans un contexte prédictif et si vous désirez avoir le plein contrôle sur le processus de sélection de modèle, la procédure GLMSELECT du module SAS/STAT est toute indiquée. De nombreuses méthodes de sélection s'offrent à vous, dont celle du partitionnement des données. Si la taille initiale de vos données est trop petite pour cette approche, GLMSELECT vous permet même d'utiliser la validation croisée. La procédure GLMSELECT mérite d'être découverte et peut s'avérer une alliée de taille de votre recherche du meilleur modèle.

Sylvain Tremblay,

Groupe de formation – Institut SAS (Canada) inc.

Référence

La procédure GLMSELECT - documentation

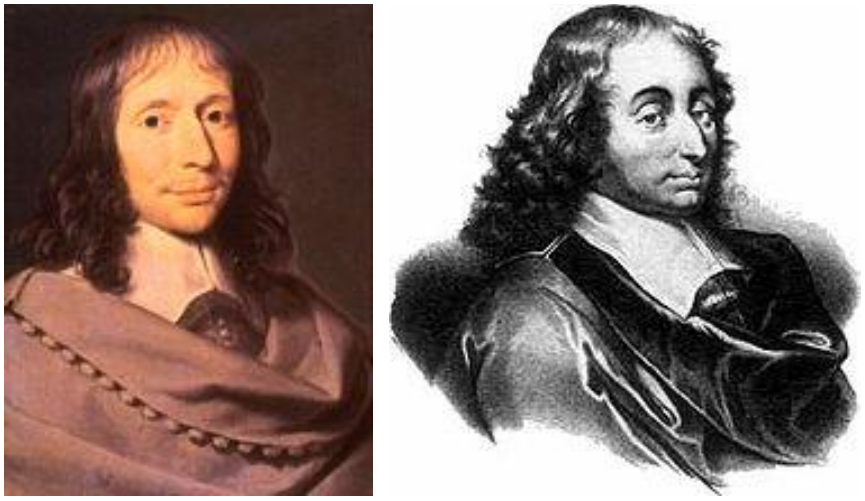
La chronique historique

Par Pierre Lavallée

Erreur de probabilités

Jusqu'au début de 1654, Antoine Gombaud, chevalier de Méré, gagnait très souvent aux tables de jeux. Puis la chance tourna… et, grâce à lui, la science naissante des calculs de probabilités trouva sa première application pratique.

L'un des jeux préférés de Méré consistait à lancer un dé à quatre reprises; pour gagner, il fallait faire sortir le six au moins une fois. Le chevalier réussissait si souvent que plus personne n'accepta bientôt de le défier. Alors, Méré modifia le jeu : il fallut jouer avec deux dés et réussir à faire un double-six en vingt-quatre coups. Méré avait calculé qu'à ce nouveau jeu il gagnerait deux fois sur trois.



Blaise Pascal

Ce jeu remporta beaucoup de succès, mais le chevalier perdit plus souvent que ses calculs ne l'avaient prévu. Il était convaincu que ces échecs répétés provenaient d'une erreur de raisonnement; mais laquelle? Perplexe, Méré écrivit à son ami Blaise Pascal, le célèbre mathématicien français : « Combien de lancers, lui demanda Méré, faut-il faire pour être à coup sûr plus souvent gagnant que perdant? »

Pascal ne fréquentait pas les tables de jeux, mais le problème philosophique et pratique que pose tout pari – comment se comporter en se confrontant au hasard? – l'intéressait au plus haut point. Il soumit le problème au

philosophe Pierre de Fermat, et, pendant quatre mois, de juillet à octobre 1654, les deux grands mathématiciens étudièrent la question. Après d'enrichissantes discussions avec Fermat, Pascal rédigea un Traité du triangle arithmétique, qui fit date dans l'histoire du calcul des probabilités.

Pierre de Fermat

Pascal et Fermat découvrirent que Méré, avec son jeu de dés, était très loin de pouvoir gagner deux fois sur trois, comme il le pensait : les calculs de Pascal révélèrent que le chevalier avait seulement 49 % de chance de gagner.

Mais Pascal démontra que, pour arriver à inverser la tendance, il suffisait de lancer les dés une fois de plus. En vingt-cinq coups, Méré avait cinquante et une chances sur cent de faire un double-six.

L'histoire ne dit pas si Méré mit en pratique le conseil de Pascal, mais il ne trouva sans doute aucun jeu de dés susceptible de le faire gagner deux fois sur trois, car, pour atteindre cet objectif, il faudrait lancer les dés… trente-neuf fois !

[Tiré de La Chance et le hasard, Éditions Time-Life (1992), Collection « L'Univers de l'étrange ».]

Suivre son cours

Par Joseph Nader

Lien vers le document

Nouvelles publications

Par Joseph Nader

Lien vers le document

Événements à venir

Par Joseph Nader

Lien vers le document

Lien intéressant

Par Simon Olivier Fournier

Calculate your survival chance on the Titanic