Ne pas hésiter à envoyer les résultats (sorties) envers un fichier, au lieu de les envoyer directement à l'imprimante. Pratiquement tous les logiciels statistiques créent automatiquement ou sur demande de l'utilisateur des fichiers ASCII, qui après peuvent être incorporés sans problèmes dans le traitement de texte.

Si l'on ne connaît pas les possibilités d'importation de fichier du logiciel en aval, **réaliser la sauvegarde des données sous forme** ASCII. Pour les séries de valeurs numériques, un fichier *ASCII delimited* (utilisation d'un séparateur standard tel que la virgule) est en général préférable.

Pour la création des fichiers-texte en ASCII pur (CONFIG.SYS, AUTOEXEC.BAT, ..), **utiliser un éditeur** au lieu d'un traitement de texte, puisque ce dernier risque de créer un en-tête incompréhensible pour le programme qui doit interpréter le fichier texte.

Références bibliographiques

LAFORE Robert. [1984]. *Assembly language primer for the IBM PC & XT*. The Waite Group, Inc. New American Library.

MERCOUROFF Wladimir. [1990]. *Architecture matérielle et logicielle des ordinateurs et des microprocesseurs*. Armand Colin Editeur, Paris. 258 p.

NORTON Peter, WILTON Richard. [1988]. *The New Peter Norton programmer's guide to the IBM PC & PS/2*. Microsoft Press. 450 p.

VIRGA. [1987]. *L'indispensable pour IBM-PC et compatibles*. Marabout, alleur, Belgique. 403 p.

Critères de validation des équations de régression

R. PALM

Faculté des Sciences agronomiques

Gembloux (Belgique)

1. Introduction

Dans les problèmes de régression, la validation des résultats est une phase importante, car les données disponibles sont souvent de qualité médiocre, en particulier lorsqu'elles proviennent d'études par observation et non de dispositifs expérimentaux et lorsqu'elles ont trait à peu d'individus caractérisés par beaucoup de variables.

Cette validation a pour but de mettre en évidence le non-respect des conditions d'application (non-linéarité des relations, inégalité des variances résiduelles, non-normalité ou autocorrélation des résidus) ou la présence d'observations particulières (données anormales, données influentes). Elle vise également à permettre la comparaison des différents modèles et la quantification de la qualité des prédictions faites à partir des équations de régression.

L'exposé est divisé en quatre parties. La première partie est consacrée à l'examen de critères globaux de mesure de la qualité d'une équation de régression. L'accent est mis sur les risques de surestimation de la qualité des modèles par l'utilisation de la variance résiduelle. Les dangers d'erreurs d'interprétation lors de la comparaison d'équations sur base des coefficients de détermination sont également illustrés.

La deuxième partie concerne la division des données et la validation croisée. La solution la plus directe consiste à répartir les données en deux groupes, un groupe servant à déterminer le modèle et l'autre servant à apprécier la qualité du modèle. Une

autre solution est la méthode du *Jackknife* dont le principe consiste à calculer n équations de régression sur la base des données relatives à n-1 individus, un individu différent étant écarté pour chacune des équations. L'individu écarté est alors prédit par l'équation calculée sur les n-1 autres individus.

La troisième partie est consacrée à diverses représentations graphiques permettant de mettre en évidence certaines situations particulières.

Enfin, la quatrième partie traite de paramètres définis dans le but de chiffrer l'influence de chacun des n individus de l'échantillon sur les résultats de l'ajustement. Pour une observation donnée, ces caractéristiques se basent essentiellement sur la comparaison des résultats obtenus, d'une part, quand on ajuste le modèle à l'ensemble des n individus et, d'autre part, quand on ajuste le modèle après avoir supprimé cette observation.

2. Critères globaux de mesure de qualité de la régression

La qualité d'une équation de régression est le plus souvent appréciée au moyen de l'écart-type résiduel et/ou du coefficient de détermination multiple ajusté.

L'écart-type résiduel mesure la dispersion des écarts entre les valeurs observées et les valeurs estimées de la variable dépendante. Il mesure essentiellement la qualité de l'ajustement réalisé plutôt que la qualité du modèle proprement dit. En fait, il s'agit d'une estimation non biaisée de la variance résiduelle de la population uniquement si les conditions d'application de la régression sont remplies. En particulier, il faut que le modèle théorique soit correct et qu'il soit défini *a priori*. Il ne peut donc y avoir eu de sélection de variables sur la base des mêmes données que celles qui sont utilisées pour l'ajustement. En pratique, on a pu constater que ce critère peut être beaucoup trop optimiste, comme nous le verrons au paragraphe 3.

Le coefficient de détermination est directement lié au rapport entre la variance résiduelle et la variance marginale, si le modèle comporte un terme indépendant. Il présente les mêmes inconvénients que la variance résiduelle. De plus, en l'absence de terme indépendant, différentes définitions de ce coefficient peuvent être proposées [KVALSETH, 1985].

L'utilisation du coefficient de détermination pour comparer les équations établies sur des données différentes peut conduire à des erreurs d'interprétation. En effet, l'équation qui a le coefficient de détermination multiple le plus grand n'a pas nécessairement la variance résiduelle la plus faible, la valeur plus élevée du coefficient pouvant être, en effet, due à une variance plus grande des variables explicatives.

De même, les coefficients de détermination ne sont pas strictement comparables dans le cas d'ajustements obtenus, pour un même ensemble de données si des transformations non linéaires différentes sont appliquées à la variable à expliquer. Des illustrations par des exemples réels et artificiels sont données par SCOTT et WILD [1991]. Ils montrent, en outre, que, dans le cas de transformations de la variable à expliquer, il y a intérêt à calculer le coefficient de détermination après avoir effectué la transformation inverse.

D'autres critères globaux d'appréciation de la qualité des équations de régression ont été proposés, en relation notamment avec les problèmes de sélection des variables. Il s'agit, par exemple, des critères Jp, Sp et Cp, qui se calculent également à partir de la variance résiduelle estimée. Une définition et une discussion de ces critères sont données par HOCKING [1972]. Comme la variance résiduelle, ces critères ne sont valables que si la sélection des variables est indépendante de l'estimation [ROECKER, 1991].

3. Division des données et validation croisée

Nous avons déjà signalé que la variance résiduelle estimée pouvait donner, en pratique, une idée trop optimiste de la qualité de l'équation de régression lorsque les mêmes données servent à établir l'équation et à chiffrer sa précision. Pour éviter cette situation, on recommande des techniques basées sur la séparation des données en deux ou plusieurs ensembles et sur la validation croisée.

La solution la plus directe consiste à répartir les données en deux groupes, un groupe servant à déterminer le modèle, l'autre groupe servant à apprécier la qualité du modèle, par exemple en calculant le carré moyen de l'erreur de prédiction, si l'objectif est de définir la qualité prédictive du modèle.

L'éclatement des données peut se faire en fonction d'un critère spécifique, tel que la date ou la région d'observation. Ainsi, lors du calcul d'équations de régression en vue de prédire les rendements des cultures à partir de diverses variables climatiques il semble logique de réserver par exemple les cinq à dix dernières années pour valider les modèles construits sur les 15 ou 20 premières années, si on dispose au total de 20 à 30 années d'observation. Dans ce contexte, nous avons pu constater que le carré moyen des erreurs de prédiction ainsi obtenu pouvait être trois à quatre fois plus grand que la variance résiduelle estimée, ce qui montre bien le caractère trop optimiste de ce dernier critère [PALM *et al.*, 1991; PALM et DAGNELIE, 1993].

En l'absence de regroupement naturel des données, l'éclatement se fait le plus souvent de manière aléatoire mais d'autres formes de répartition ont été proposées par SNEE [1977] et l'importance relative de l'échantillon utilisé pour la validation par

rapport à l'échantillon utilisé pour l'estimation a été discutée par PICARD et BERK [1990].

La validation telle que nous venons de la décrire ne peut raisonnablement s'envisager que si la taille de l'échantillon est suffisante. SNEE [1977] considère que cette taille doit être supérieure à 2p+20, p étant le plus grand nombre de coefficients susceptibles d'être introduits dans l'équation. Pour pallier cet inconvénient, une autre forme de validation a été proposée. Il s'agit de la méthode du "*Jackknife*", dont le principe consiste à calculer n équations de régression sur base des données relatives à n-1 individus, un individu différent étant écarté pour chacune des équations. L'individu écarté est alors prédit par l'équation calculée sur les n-1 autres individus. A l'issue de ces calculs on dispose de n prédictions et donc de n erreurs de prédiction, dont on peut calculer le carré moyen. Cette validation croisée a évidemment l'avantage de ne pas éliminer définitivement des observations du fait de la validation. De plus, le calcul des erreurs de prédiction est particulièrement simple car il n'est pas nécessaire, en pratique, de calculer n fois l'équation. En effet, ces erreurs s'obtiennent très facilement à partir des informations liées à la régression sur les n données [PALM, 1988; WEISBERG, 1985].

Il faut cependant se rendre compte que l'erreur de prédiction calculée par la procédure décrite ci-dessus pour un individu donné n'est pas tout à fait indépendante de l'observation relative à cet individu si le choix des variables explicatives à faire figurer dans l'équation a été réalisé sur l'ensemble des données. Pour le problème des prévisions des rendements déjà évoqué, on a constaté que la technique du "*Jackknife*" surestime très nettement la qualité des modèles par comparaison avec le résultat de la validation réalisée après éclatement des données en fonction de la date d'observation [PALM *et al.*, 1991; PALM et DAGNELIE, 1993].

4. Représentations graphiques

Les exemples artificiels de régression linéaire donnés par ANSCOMBE [1973] et repris par CHATTERJEE et PRICE [1991] et par WEISBERG [1985] ainsi que ceux de TOMASSONE *et al.* [1983] montrent clairement que l'examen des critères globaux est insuffisant et que les représentations graphiques peuvent avantageusement être réalisées en vue de vérifier la validité des modèles de régression.

L'établissement de diagrammes de dispersion de la variable à expliquer en fonction d'une variable explicative particulière et de diagrammes de dispersion des résidus en fonction d'une variable explicative ou bien des valeurs estimées de la variable à expliquer est une pratique courante et ancienne. Ces graphiques permettent

de détecter, dans une certaine mesure du moins, la non-linéarité, l'inégalité de variances résiduelles, la présence de points influents ou anormaux.

Des histogrammes des résidus et des diagrammes des résidus en fonction des scores normaux peuvent également être réalisés en vue de détecter la non-normalité des résidus. On notera cependant que pour ces types de représentation, il est préférable d'utiliser les résidus standardisés, c'est-à-dire des résidus divisés par leur erreur-standard, car les variances des résidus estimés ne sont pas constantes [PALM, 1988].

Des représentations graphiques plus particulières sont également proposées pour étudier l'effet d'une variable explicative donnée sur la variable à expliquer après l'élimination des effets des p-1 autres variables [WEISBERG, 1985; LARSEN et McCLEARY, 1972].

Enfin, des graphiques peuvent encore être réalisés en portant en abscisse une variable explicative particulière ou simplement le numéro d'ordre de l'individu et en ordonnée une caractéristique particulière, calculée pour chaque individu et destinée à quantifier l'influence de cet individu, comme nous allons le voir au paragraphe suivant. Des informations complémentaires concernant les représentations graphiques mentionnées ci-dessus et concernant d'autres représentations graphiques sont données, notamment, par ATKINSON [1986] et par CHATTERJEE et HADI [1986].

5. Mesure de l'influence des individus

Différents paramètres ont été définis dans le but de chiffrer l'influence ou l'importance de chacun des n individus de l'échantillon sur les résultats de l'ajustement. Pour une observation donnée, ces caractéristiques se basent essentiellement sur la comparaison des résultats obtenus, d'une part, lorsqu'on ajuste le modèle à l'ensemble des n individus et, d'autre part, lorsqu'on ajuste le modèle après avoir supprimé cette observation.

Des informations concernant les diverses mesures d'influence sont données notamment dans les synthèses bibliographiques de CHATTERJEE et HADI [1986], HADI [1992], HOCKING [1983] et PALM [1988]. Une raison de la multiplicité de ces mesures tient à la nature de l'influence quantifiée: s'agit-il, par exemple, de l'influence sur les valeurs prédites, sur les coefficients de régression ou sur la matrice des variances et covariances des coefficients de régression?

Plusieurs mesures de l'influence sont actuellement disponibles dans des logiciels tels que SAS ou Minitab et l'utilisateur occasionnel risque d'être dérouté

devant le volume important des documents imprimés que peuvent fournir ces logiciels. Pour éviter que ces documents ne compliquent finalement l'interprétation des données au lieu de la simplifier, l'utilisateur a sans doute intérêt à se limiter à l'analyse d'un petit nombre de caractéristiques. On peut par exemple lui conseiller de repérer les observations les plus influentes sur base de l'examen de la distance proposée par COOK [1977], qui est une fonction croissante du carré du résidu et d'une mesure de l'éloignement de l'individu par rapport au centre de gravité du nuage des points dans l'espace des variables explicatives, du moins si sa régression comporte un terme indépendant. Cette distance, habituellement désignée par h_i ou h_{ii} dans la littérature, est directement liée à la distance de MAHALANOBIS entre un individu et le vecteur des moyennes [DAGNELIE, 1982; WEISBERG, 1985].

Pour les observations les plus influentes, l'examen du résidu et de la valeur de h_i permet de détecter la cause de l'influence: l'individu est-il influent du fait de l'importance du résidu ou du fait de l'éloignement de l'individu par rapport au centre de gravité du nuage de points dans l'espace des variables?

Après avoir opté pour l'une ou l'autre mesure de l'influence, le problème qui se pose à l'utilisateur est, d'une part, de préciser à partir de quelle valeur de la mesure il doit considérer l'observation comme influente et, d'autre part, de définir l'attitude qu'il doit adopter en présence d'une donnée influente. Des éléments de réponse à ces questions peuvent être trouvées dans PALM [1988].

Les différentes mesures de l'influence évoquées ci-dessus ne prennent en compte qu'un individu à la fois. Il peut cependant arriver que des groupes d'individus soient influents en bloc et cette influence peut passer inaperçue si les individus sont examinés un par un. La recherche de ces groupes soulève cependant des difficultés, car les temps de calcul seraient rapidement prohibitifs si on voulait considérer tous les groupes possibles de 2, 3, *etc.* individus. Des informations à ce sujet sont données par COOK et WEISBERG [1982].

Enfin, ces mesures sont calculées pour un modèle déterminé et ne donnent pas d'information concernant l'influence des individus quant au choix des variables.

6. Conclusions

Le rapide survol des critères de validation qui vient d'être présenté, n'est nullement exhaustif. En particulier les différents test statistiques qui peuvent être utilisés pour vérifier les conditions d'application particulières (égalité des variances, normalité, *etc.*) n'ont pas été décrits mais constituent aussi des critères de validation. Des informations complémentaires peuvent être trouvées dans les livres de BELSLEY *et al.* [1980], CHATERJEE et HADI [1988], COOK et WEISBERG [1982], notamment.

Pour conclure, rappelons que la régression se présente souvent comme un processus complexe à caractère itératif, incluant l'examen des données disponibles et d'un ensemble de modèles potentiellement valables. Dans cette optique, suite aux développements théoriques, à la commercialisation de logiciels statistiques plus complets et de moyens de calculs plus performants, une série d'outils nouveaux ont été mis à la disposition des utilisateurs pour leur permettre d'affiner les analyses. L'utilisation de ces outils doit être encouragée car ils permettent incontestablement un examen des données sous divers angles de vue, ce qui compense la tendance naturelle qui est de moins regarder les données, du fait de l'utilisation des moyens informatiques disponibles.

Par ailleurs, on ne saurait trop recommander l'utilisation d'un ensemble de données indépendantes des données utilisées pour l'établissement du modèle. Cette façon de procéder constitue, en effet, un contrôle de tout le processus de construction du modèle.

7. Références bibliographiques

ANSCOMBE F.J. [1973]. Graphs in statistical analysis. *Amer. Stat.* 27, 17-21.

ATKINSON A.C. [1986]. *Plots, transformations and regression: an introduction to graphical methods of diagnostic regression analysis.* Oxford, Clarendon Press, 282 p.

BELSLEY D.A., KUH E. et WELSCH R.E. [1980]. *Regression diagnostics: identifying influential data and sources of collinearity.* New York, Wiley, 310 p.

CHATTERJEE S., HADI A.S. [1988]. *Sensitivity analysis in linear regression.* New York, Wiley, 336 p.

CHATTERJEE S., PRICE B. [1991]. *Regression analysis by example*. New York, Wiley, 298 p.

COOK R.D. [1977]. Detection of influential observations in linear regression. *Technometrics* 19, 15-18.

COOK R.D. et WEISBERG S. [1982]. *Residuals and influence in regression*. New York, Chapman et Hall, 230 p.

DAGNELIE P. [1982]. *Analyse statistique à plusieurs variables*. Gembloux, Presses Agronomiques, 362 p.

HADI A.S. [1992]. A new measure of overall potential influence in linear regression. *Comput. Stat. Data Anal.* 14, 1, 1-27.

HOCKING R.R. [1972]. Criteria for selection of a subset regression: which one should be used? *Technometrics* 14, 967-970.

HOCKING R.R. [1983]. Developments in linear regression methodology: 1959-1982. *Technometrics* 25, 219-249.

KVALSETH T.O. [1985]. Cautionary note about R^2. *Amer. Stat.* 39, 279-285.

LARSEN W.A., McCLEARY S.J. [1972]. The use of partial residual plots in regression analysis. *Technometrics* 14, 781-790.

PALM R. [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform. (Gembloux)* 88/1, 27 p.

PALM R., DAGNELIE P. [1993]. *Tendance générale et effets du climat dans la prévision des rendements agricoles des différents pays des Communautés Européennes*. Luxembourg, Office des publications officielles des Communautés Européennes, 128 p.

PALM R., DE BAST A., LAHLOU M. [1991]. Comparaison de modèles agrométéorologiques de type statistique-empirique construits à partir de différents ensembles de variables météorologiques. *Bull. Rech. Agron. Gembloux* 26, 71-89.

PICARD R.R., BERK K.N. [1990]. Data splitting. *Amer. Stat.* 44, 140-147.

ROECKER E. [1991]. Prediction error and its estimation for subset selected models. *Technometrics* 33, 459-468.

SCOTT A., WILD C. [1991]. Transformations and R^2. *Amer. Stat.* 45, 127-129.

SNEE R.D. [1977]. Validation of regression models: methods and examples. *Technometrics* 19, 415-428.

TOMASSONE R., LESQUOY E. et MILLIER C. [1983]. *La régression: nouveaux regards sur une ancienne méthode statistique*. Paris, Masson, 180 p.

WEISBERG S. [1985]. *Applied linear regression*. New York, Wiley, 324 p.