Optimisation à finalité statistique

Salim Lardjane

Université de Bretagne-Sud

Cours 9 - Applications Statistiques : Problèmes de régression

Problèmes de régression

De nombreux problèmes rencontrés en sciences expérimentales peuvent se ramener à la détermination d'estimations des paramètres d'une fonction de régression de la forme

$$y = h(x_1, \ldots, x_p; \theta_1, \ldots, \theta_m)$$

où y est la *variable de réponse*, $x_1, \ldots, x_p$ les *variables explicatives* et $\theta_1, \ldots, \theta_m$ les paramètres que l'on souhaite estimer.

Problèmes de régression

Lorsque y est une variable continue et h est linéaire en ses paramètres, le problème d'estimation peut être résolu de façon relativement directe par la méthode des moindres carrés, le critère d'ajustement s'écrivant

$$Q = \sum_{i=1}^{n} [f_i(y_i; \theta)]^2$$

où $f_i(y_i; \theta) = y_i - h(x_{1i}, \ldots, x_{pi}; \theta_1, \ldots, \theta_m) \equiv y_i - h_i(\theta)$ désigne l'erreur d'ajustement.

Problèmes de régression

On peut différentier ce critère par rapport aux paramètres de façon à obtenir un ensemble d'équations d'estimation qui sont *linéaires* et peuvent donc être résolues explicitement.

Il y toutefois de nombreux problèmes importants pour lesquels y est *catégorielle* plutôt que continue ou pour lesquels h est *non-linéaire* en ses paramètres.

Comment évaluer l'effet des variables explicatives sur la réponse dans ce cas ?

De tels problèmes requièrent des algorithmes d'estimation plus complexes que ceux utilisés pour la régression linéaire simple.

Régression non-linéaire

Un modèle utilisé en Biochimie (entre autres) s'écrit

$$y = \frac{\theta_1 x}{\theta_2 + x}$$

y et x sont des concentrations de produits chimiques lorsque la réaction est à l'équilibre.

On souhaite estimer θ_1 et θ_2 à partir d'observations de valeurs prises par y et x.

Régression non-linéaire

Ici, la fonction de régression est *non-linéaire* en θ_1 et θ_2.

De nombreuses méthodes ont été suggérées pour mettre le modèle précédent sous une forme linéaire afin de pouvoir utiliser les techniques usuelles de régression pour estimer θ_1 et θ_2.

On peut écrire par exemple

$$\frac{1}{y} = \frac{1}{\theta_1} + \frac{\theta_1}{\theta_2} \cdot \frac{1}{x}$$

La régression linéaire de $1/y$ sur $1/x$ fournit alors des estimations de θ_1 et θ_2.

Régression non-linéaire

De telles méthodes ont été utilisées jusque dans les années 1980 mais sont aujourd'hui *obsolètes* puisque l'on peut avoir recours aux divers algorithmes d'optimisation vus précédemment pour obtenir des estimations de θ_1 et θ_2 de façon directe.

On peut utiliser par exemple la méthode du simplexe de Nelder-Mead, la méthode de la plus grande pente ou la méthode de Newton-Raphson.

Régression non-linéaire

Notons toutefois que *dans le cas d'un critère de moindre carrés*, les calculs requis par les deux derniers algorithmes peuvent être simplifiés en introduisant la matrice

$$S = \begin{pmatrix} \frac{\partial f_1}{\partial \theta_1} & \cdots & \frac{\partial f_1}{\partial \theta_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial \theta_1} & \cdots & \frac{\partial f_n}{\partial \theta_m} \end{pmatrix}$$

où $f_i = y_i - h_i(\theta)$.

Régression non-linéaire

Comme, en général, $n > m$, S n'est pas une matrice carrée.

Le critère d'ajustement des moindres carrés s'écrit

$$Q = \sum_i (y_i - h_i(\theta))^2 = \sum_i f_i^2$$

et en différentiant par rapport aux paramètres, on trouve que les éléments du vecteur gradient g, requis par exemple par l'algorithme de la plus grande pente, peut s'écrire

$$\frac{\partial Q}{\partial \theta_k} = \sum_i 2 f_i \frac{\partial f_i}{\partial \theta_k}$$

On a donc

$$g = 2S'f$$

où $f = (f_1, \ldots, f_n)'$.

Régression non-linéaire

Les dérivées secondes de Q par rapport aux paramètres s'écrivent, quant à elles,

$$\frac{\partial^2 Q}{\partial\theta_k \partial\theta_j} = 2\sum_i \frac{\partial f_i}{\partial\theta_j}\frac{\partial f_i}{\partial\theta_k} + 2\sum_i f_i \frac{\partial^2 f_i}{\partial\theta_j \partial\theta_k}$$

Si on fait l'hypothèse que le deuxième terme du deuxième membre de l'équation ci-dessus est négligeable, on obtient

$$\frac{\partial^2 Q}{\partial\theta_k \partial\theta_j} \approx 2\sum_i \frac{\partial f_i}{\partial\theta_j}\frac{\partial f_i}{\partial\theta_k}$$

ce qui signifie qu'on peut mettre la hessienne de Q sous la forme *approximative*

$$H \approx 2S'S \quad (m \times m)$$

Régression non-linéaire

La formule de mise à jour de l'algorithme de Newton-Raphson prend alors la forme

$$\theta^{i+1} = \theta^i - (2S'S)^{-1}(2S'f)$$

c'est-à-dire

$$\theta^{i+1} = \theta^i - (S'S)^{-1}S'f$$

En général, $S'S$ est *définie positive*, donc l'algorithme doit converger et la formule ainsi obtenue est appelée *formule de Gauss-Newton*.

Une modification de cette formule suggérée par Levenberg est d'utiliser plutôt la formule de mise à jour

$$\theta^{i+1} = \theta^i - (\lambda I_m + S'S)^{-1} S'f$$

où λ est un scalaire qui peut être ajusté pour contrôler la suite des itérations et où I_m désigne la matrice identité de dimensions $m \times m$.

Lorsque $\lambda >> 1$, la formule de mise à jour équivaut à

$$\theta^{i+1} = \theta^i - \frac{1}{\lambda} S'f$$

ce qui correspond à la méthode de la plus grande pente.

Régression non-linéaire

Lorsque $\lambda << 1$, la formule de mise à jour de Levenberg équivaut plus ou moins à la formule de mise à jour de Newton-Raphson.

L'utilisation de λ pour contrôler la procédure itérative permet de tirer profit des avantages de la méthode de la plus grande pente *loin du minimum* et de la convergence rapide de la méthode de Newton-Raphson lorsqu'on est *proche* de celui-ci.

Régression non-linéaire

Marquardt a proposé un procédé de détermination de λ à chaque itération qui est généralement considéré comme efficace.

L'algorithme est connu globalement sous le nom d'*algorithme de Levenberg-Marquardt*.

Comme on l'a vu celui-ci est spécifiquement conçu pour traiter des *problèmes de moindres carrés*.

Il converge en général plus rapidement que les algorithmes vus précédemment.

Sous R, l'algorithme de Gauss-Newton est implémenté par la fonction **nls()** et la méthode de Levenberg-Marquardt par **nls.lm()** (package **minpack.lm**).

Régression non-linéaire

Exercice 10. Afin d'illustrer l'utilisation des techniques d'optimisation pour des problèmes de moindres carrés non linéaires, considérons les données de réaction biochimique suivantes :

x	y
84.6	12
83.9	12.5
148.2	17.2
147.8	16.7
463.9	28.3
463.8	26.9
964.1	37.6
967.6	35.8
1925.0	38.5
1900.0	39.9

Régression non-linéaire

Ajuster le modèle $y = \theta_1 x/(\theta_2 + x)$ d'abord par l'algorithme de Gauss-Newton puis par l'algorithme de Levenberg-Marquardt, en partant des conditions initiales $\theta_1 = 10$, $\theta_2 = 50$. Comparer les résultats obtenus.

Régression non-linéaire

Notons que si les paramètres de régression linéaire ou non-linéaire sont le plus souvent estimés par moindres carrés, il est parfois intéressant de les estimer en utilisant d'autres critères d'ajustement, en particulier la somme des valeurs absolues et la valeur absolue maximale (critère de Tchebycheff). Avec les notations précédentes, ces deux critères s'écrivent respectivement

$$L_1 = \sum_i |f_i|$$

$$L_\infty = \max_i |f_i|$$

L'intérêt de ces critère est dû à plusieurs raisons, parmi lesquelles :

Régression non-linéaire

1. De nombreuses données ne remplissent pas la condition de normalité des résidus qui justifie pleinement l'utilisation du critère des moindres carrés.

2. Les estimations par moindres carrés sont sensibles aux valeurs aberrantes.

Notons que les estimations correspondant aux deux critères précédents peuvent être obtenues à l'aide d'algorithmes d'*optimisation linéaire* (non discutés dans ce cours).

Régression logistique

De nombreuses études, notamment en Médecine, font intervenir une *variable de réponse dichotomique* et on souhaite pouvoir modéliser la dépendance entre une telle variable et un ensemble de variables explicatives.

Par exemple, le tableau suivant fournit les résultats d'un essai clinique pour un médicament donné et on souhaite évaluer la relation entre l'amélioration de l'état du patient et l'âge du patient et son poids.

Régression logistique

Résultats d'essai clinique :

	y_i	âge	poids en kilos
1	1	27	70
2	1	35	72.5
3	1	33	75.5
4	1	38	69
5	1	23	80
6	1	41	82.5
7	1	36	67.5
8	0	34	71
9	0	38	71.5
10	0	39	85
11	0	37	82.5
12	0	28	82
13	0	42	84.5
14	0	44	75.5
15	0	48	90
16	0	29	83.5
17	0	39	66.5
18	0	39	85.5
19	0	44	85.5
20	0	47	81.5

Régression logistique

Une façon de modéliser la dépendance entre la réponse, l'âge et le poids consiste à supposer que la probabilité que la réponse y_i prenne la valeur 1 est une fonction linéaire fes variables explicatives :

$$p_i = \mathbb{P}(y_i = 1) = \sum_{j=1}^{p} \beta_j \, x_{ji}$$

où y_i est la variable de réponse pour le patient i et x_{ji} $(j = 1, \ldots, p)$ sont les valeurs des p variables explicatives pour ce même patient (dans l'exemple, $p = 2$).

Régression logistique

Le modèle ainsi postulé est un modèle de régression linéaire multiple usuel pour une variable de réponse continue et les paramètres $\beta_1, \ldots, \beta_p$ peuvent donc être estimés par moindres carrés.

Une telle approche est cependant inaptée en pratique et ce pour diverses raisons :

1. Les y_i ne sont pas normalement distribués

2. Les estimations par moindres carrés des paramètres conduisent à des valeurs ajustées de la probabilité qui ne vérifient pas les conditions $0 \leq p_i \leq 1$.

Régression logistique

Un modèle plus approprié est obtenue en exprimant la *transformation logistique* des p_i comme fonction linéaire des variables explicatives, c'est-à-dire

$$\lambda_i \equiv \log \frac{p_i}{1-p_i} = \sum_{j=1}^{p} \beta_j \, x_{ji}$$

Les valeurs de $p_i = 0$ et $p_i = 1$ correspondent respectivement à $\lambda_i = -\infty$ et $\lambda_i = +\infty$, ce qui règle le problème des contraintes sur les p_i ajustées.

L'ajustement de ce modèle à des données telles que celles du tableau précédent implique l'estimation des paramètres $\beta_1, \ldots, \beta_p$. Pour ce faire, la méthode d'estimation la plus utilisée est celle du maximum de vraisemblance.

Régression logistique

La fonction de vraisemblance s'écrit :

$$\mathcal{L}(\beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

De plus, en utilisant l'expression du modèle postulé, on a

$$p_i = \frac{\exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}}{1 + \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}}$$

d'où

$$\mathcal{L}(\beta_1, \ldots, \beta_p) = \frac{\exp\left\{\sum_{j=1}^{p} \beta_j t_j\right\}}{\prod_{i=1}^{n} (1 + \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\})}$$

où

$$t_j = \sum_{i=1}^{n} x_{ji}\, y_i$$

Par conséquent, la log-vraisemblance s'écrit

$$\ell(\beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} \beta_j\, t_j - \sum_{i=1}^{n} \log\left(1 + \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}\right)$$

Régression logistique

La maximisation de ℓ peut être effectuée en utilisant les algorithmes d'optimisation vus plus haut.

L'algorithme le plus fréquemment utilisé est celui de Newton-Raphson. La **proc logistic** de SAS l'utilise notamment.

Elle propose également d'utiliser l'algorithme du score de Fisher (méthode par défaut). Les deux procédures coïncident en fait pour l'exemple considéré.

Régression logistique

L'algorithme de Newton-Raphson requiert les dérivées premières et secondes de ℓ par rapport aux paramètres, qui sont :

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_s} &= t_s - \sum_{i=1}^{n} \frac{x_{si} \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}}{1 + \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}} \\
\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} &= -\sum_{i=1}^{n} \frac{x_{ri} x_{si} \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}}{\left(1 + \exp\left\{\sum_{j=1}^{p} \beta_j x_{ji}\right\}\right)^2}
\end{aligned}$$

Des valeurs initiales des paramètres peuvent être obtenues de diverses manières, mais le choix de celles-ci *n'est pas en général un facteur critique*.

Régression logistique

Exercice 11. Afin d'illustrer l'utilisation des différents algorithmes d'optimisations vus plus haut sur le problème de régression logistique, on utilisera les données du tableau précédent (essai clinique). On utilisera l'algorithme de quasi-Newton BFGS avec approximation numérique du gradient et la méthode du simplexe de Nelder-Mead pour divers jeux de conditions initiales.

Remarque. Un point important concernant les algorithmes de type quasi-Newton est que, bien qu'ils ne requièrent pas d'expression explicite des dérivées secondes, ils *fournissent généralement une très bonne approximation de l'inverse de la hessienne à l'optimum*. Or, dans un contexte statistique, celle-ci est utile comme approximation de la matrice de variance-covariance de l'estimateur du maximum de vraisemblance.