Que font (en gros) les modélisations numériques ?

Antoine Bérut et Michel Fruchart

29 octobre 2013

N.B. : Ce document a uniquement pour but de décrire un peu ce que font les logiciels comme `Regressi` *lorsque l'on veut ajuster un modèle théorique (de type $Y = f(X)$) sur des données expérimentales, afin d'avoir les idées un peu claires sur le sujet. Il ne prétend pas donner une approche exhaustive ni optimale des méthodes permettant d'ajuster un modèle sur des données.*

1 Obtenir les meilleurs paramètres pour un modèle donné

On considère que l'on a deux jeux de données expérimentaux X et Y, qui comportent chacun N points (numérotés de 1 à N), ainsi que les incertitudes sur les points $Y(i)$, composées de N valeurs $\sigma_Y(i)$ qui ne sont pas nécessairement les mêmes pour chaque point. On cherche à ajuster un modèle théorique quelconque avec p paramètres libres (nommés a_1, a_2, …, a_p) du type :

$$Y = f(X, a_1, a_2, \ldots, a_p) \tag{1}$$

Par exemple pour un modèle affine on cherchera a_1 et a_2 tels que :

$$Y = a_1 + a_2 \times X \tag{2}$$

En général, la méthode la plus simple pour obtenir les paramètres qui correspondent le mieux aux données consiste à minimiser vis-à-vis des paramètres la fonction erreur suivante :

$$\chi^2(a_1, a_2, \ldots, a_p) = \sum_{i=1}^{N} \frac{\left[Y(i) - f(X(i), a_1, a_2, \ldots, a_p)\right]^2}{\sigma_Y(i)^2} \tag{3}$$

N.B. : Si on ne connait pas les barres d'erreur, on remplace simplement chaque valeur de σ_Y par 1 (dans ce cas on fait implicitement l'hypothèse que les barres d'erreur sur les points de Y sont toutes les mêmes). Dans le cas où on a également σ_X les barres d'erreur sur X, on généralise en remplaçant chaque valeur $\sigma_Y(i)^2$ par $\sigma_Y(i)^2 + [f'(X(i)) \times \sigma_X(i)]^2$, grâce à la propagation des erreurs. Pour `Regressi`*, seules ces deux options (pas de barre d'erreur ou barres d'erreur sur X et Y) existent. Pour prendre en compte les barres d'erreur, il faut que toutes les grandeurs présentes dans le tableur en soit munies et avoir coché « Méthode des ellipses (chi2) » dans l'onglet « Options de modélisation ».*

On récupère ainsi les paramètres « optimaux » : a_1^*, a_2^*, …, a_p^* qui minimisent la fonction χ^2. Il faut noter que si les barres d'erreurs sont modifiées par un facteur multiplicatif global (par exemple si on prend un jeu d'incertitudes σ'_Y telles que $\sigma'_Y(i) = 10 \times \sigma_Y(i)$ pour tout i), cela ne modifie pas les valeurs a_i^* obtenues.

2 Estimer la validité de la modélisation

Pour estimer la validité de la modélisation, on regarde la grandeur suivante (appelée ici χ^2 réduit) :

$$\chi^2_{\mathrm{r}} = \frac{1}{N-p}\sum_{i=1}^{N}\frac{\left[Y(i) - f(X(i), a_1^*, a_2^*, \ldots, a_p^*)\right]^2}{\sigma_Y(i)^2} \tag{4}$$

qui n'est rien d'autre que la valeur de la fonction χ^2 pour les paramètres optimaux, divisée par $(N-p)$, le nombre de degrés de liberté.

Dans le cas où on ne connait pas les incertitudes sur Y, on peut montrer que cette valeur est un estimateur de la variance de Y. Elle doit donc être comparée au carré de l'incertitude que l'on peut estimer sur Y(ou à la valeur moyenne de Y au carré) afin de voir si cette estimation est raisonnable ou non.

Dans le cas où on a utilisé les valeurs σ_Y (ou σ_Y et σ_X) comme incertitudes sur Y (ou X et Y), cette valeur devrait idéalement être égale à 1, car σ_Y et σ_X devraient être les écarts types des grandeurs Y et X. Si χ^2_{r} est très grand devant 1, c'est que les barres d'erreurs que l'on a estimées sont trop petites pour que le modèle soit compatible avec les données. Cela signifie soit que les barres d'erreurs ont été sous-estimées (ce qui est relativement rare), soit que le modèle n'est pas valide. Si χ^2_{r} est très petit devant 1, c'est que les barres d'erreur que l'on a estimées sont trop grandes et donc que la modélisation est « trop bonne ». Cela signifie soit que les barres d'erreur ont été surestimées, soit que l'on pourrait sans doute également faire passer un autre modèle par les barres d'erreurs et donc qu'on ne peut pas vraiment conclure sur sa validité. Par exemple, si on modélise un ensemble de 5 points par un polynôme d'ordre 6, on arrivera sans problème à avoir une valeur de χ^2_{r} très petite, ce qui ne veut pas dire que le modèle est bon pour autant.

N.B. : `Regressi` *indique une valeur qu'il appelle « Ecart quad. Y » ou « Chi2/(N-p) », selon que l'on a ou non pris en compte les barres d'erreur, et qui est en fait* $\sqrt{\chi^2_r}$.

3 Estimer l'erreur sur les paramètres obtenus

La dernière étape consiste à déterminer un intervalle de confiance pour les paramètres obtenus. *A priori*, cette étape n'est justifiée que si la validité du modèle est raisonnable.

L'idée générale est de regarder l'évolution de la fonction $\chi^2(a_1, a_2, \ldots, a_p)$ en fonction des paramètres, autour de son minimum (atteint pour les paramètres optimaux a_1^*, a_2^*, …, a_p^*). Sous les bonnes conditions, cette évolution sera bien prédite par un développement limité à l'ordre 2 de la fonction autour du minimum (les dérivées premières étant nulles puisqu'on est autour d'un minimum). Par exemple pour une fonction à deux paramètres :

$$\Delta\chi^2 = \chi^2(a_1, a_2) - \chi^2(a_1^*, a_2^*) \approx \frac{1}{2}\left(\frac{\partial^2\chi^2}{\partial a_1^2}\Delta a_1^2 + \frac{\partial^2\chi^2}{\partial a_2^2}\Delta a_2^2\right) + \frac{\partial^2\chi^2}{\partial a_1 \partial a_2}\Delta a_1 \Delta a_2 \tag{5}$$

où on a posé $\Delta a_i = (a_i - a_i^*)$.

Il faut ensuite choisir un critère pour définir l'intervalle qui détermine la validité de la modélisation. C'est un choix arbitraire, qui dépend en particulier de la taille de l'intervalle de confiance que l'on souhaite prendre et du nombre de paramètres pour lesquels on chercher à estimer l'erreur conjointement. En général, on s'intéresse aux paramètres pris un à un, et on prend $\Delta\chi^2 = 1$. Il ne reste plus qu'à trouver les variations des paramètres Δa_i qui suivent ce critère pour obtenir les barres d'erreur σ_{a_i} qui leur sont associées. Pour `Regressi`, la notice indique que l'intervalle de confiance à $\pm n\sigma$ est obtenu à partir des valeurs des paramètres qui donnent :

$$\chi^2(a_1, a_2, \ldots, a_p) = \frac{\chi^2(a_1^*, a_2^*, \ldots, a_p^*)}{1 + (\frac{n}{N})^2}. \tag{6}$$

N.B. : C'est là qu'on voit qu'il est inutile de prendre les barres d'erreur sur les paramètres pour une modélisation dont la validité est douteuse. Si on a par exemple χ^2_r qui est très grand devant 1, il suffira a priori d'une petite variation des paramètres pour atteindre $\Delta\chi^2 = 1$. On aura donc des barres d'erreur sur les paramètres qui seront faibles alors que le modèle n'est en réalité pas adapté aux données.

De manière générale, les algorithmes de modélisation renvoient souvent C, la matrice estimée des covariances des paramètres, qui est l'inverse de la matrice Hessienne H de la fonction χ^2 divisée par 2 :

$$H_{ij} = \frac{\partial^2 \chi^2}{\partial a_i \partial a_j} \quad \text{et} \quad C = \left[\frac{1}{2}H\right]^{-1} \tag{7}$$

Dans le cas général ces valeurs n'ont pas de signification précise : ce sont les valeurs estimées des covariance des paramètres de la modélisation, mais elles ne sont pas directement reliées à l'intervalle de confiance des paramètres. Mais sous l'hypothèse que les erreurs suivent une loi normale (cette hypothèse est implicitement faite par `Regressi`), et que le modèle est linéaire en les paramètres (ou bien que l'ensemble des points se trouvent dans une zone où le modèle peut être linéarisé en les paramètres), la fonction $\chi^2(a_1, a_2, \ldots, a_p)$ suit une loi de probabilité du χ^2 à $N - p$ degrés de liberté et l'intervalle de confiance sur les paramètres est directement relié aux valeurs de C. On peut alors montrer que le Δa_i qui correspond à $\Delta\chi^2 = 1$ est donné par $\sigma_{a_i} = \sqrt{C_{ii}}$ et qu'il correspond à un intervalle de confiance de 68 %. On peut alors dire que $a_i = a_i^* \pm \sigma_{a_i}$ pour un intervalle de confiance à 68 %. De la même façon, on peut montrer que $a_i = a_i^* \pm 2\sigma_{a_i}$ pour un intervalle de confiance à 95 %, $a_i = a_i^* \pm 3\sigma_{a_i}$ pour un intervalle de confiance à 99 %, etc.

4 Bibliographie

Vous pouvez trouver des informations utiles (et bien plus complètes) dans les références suivantes :

- Les *Numerical Recipes in C* (disponibles gratuitement en ligne sur le site `http://www.nr.com/` pour les versions obsolètes), chapitre *Modeling of Data* (page 656 pour la seconde édition parue en 1992).
- L'article « Analyse de données, méthodes numériques et sciences physiques » de Trigeassou et Beaufils dans le BUP 731 (page 297).
- L'article « Régression linéaire et incertitudes expérimentales » par Beaufils et Richoux dans le BUP 796 (page 1361).