

Question de Dominique Lambert

"Pourquoi la RMSE est-elle considérée comme une estimation de l'écart-type des réponses calculées et utilisée pour obtenir une estimation de l'écart-type des coefficients du modèle ?»

Réponse de Jacques Goupy

Définitions (rappel)

Définition de la variance

La variance d'un ensemble de valeurs purement aléatoires est la somme des carrés des écarts à la moyenne divisée par le nombre de degrés de liberté de l'ensemble des écarts.

Définition de l'écart-type

L'écart-type d'un ensemble de valeurs purement aléatoires est la racine carrée de la variance.

Modèle polynomial

Soit une réponse y_i modélisable par un polynôme de p coefficients.

$$y_i = a_0 + a_1x_1 + a_2x_2 + \dots + e_i \quad \{1\}$$

Relation dans laquelle :

- y_i est la réponse mesurée de l'essai i .
- a_i sont les coefficients inconnus du modèle.
- x_i sont les niveaux connus des facteurs.
- e_i est l'écart, composé d'un terme aléatoire et d'un terme traduisant la différence entre le vrai le modèle mathématique et celui choisi par l'expérimentateur. (Attention. Ne pas confondre "écart" et "écart à la moyenne" !).

Si l'on effectue n essais en choisissant des niveaux différents, on obtient un système de n équations indépendantes que l'on peut écrire sous forme matricielle :

$$\mathbf{y} = \mathbf{X} \mathbf{a} + \mathbf{e} \quad \{2\}$$

où

- \mathbf{y} est le vecteur $(n,1)$ des réponses mesurées ou observées.
- \mathbf{X} est la matrice (n,p) des niveaux des facteurs et du modèle mathématique. Elle est supposée de plein rang.

- \mathbf{a} est le vecteur $(p,1)$ des coefficients. C'est le vecteur à déterminer.
- \mathbf{e} est le vecteur $(n,1)$ des écarts. Ce vecteur est inconnu.

Le problème est de trouver la valeur des coefficients à partir des réponses mesurées et des niveaux des facteurs. Le critère des moindres carrés permet ce calcul. On obtient une estimation des coefficients (estimation notée $\hat{\mathbf{a}}$) et les écarts deviennent les résidus et sont notés $\hat{\mathbf{e}}$. La solution des moindres carrés est :

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad \{3\}$$

Il est possible, connaissant les coefficients, de calculer les réponses quels que soient les niveaux des facteurs :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}} \quad \{4\}$$

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}} = \mathbf{X}\hat{\mathbf{a}} + \hat{\mathbf{e}} \quad \{5\}$$

Somme des carrés des résidus

Les réponses mesurées, \mathbf{y} , sont la somme d'une partie modélisée, $\mathbf{X}\hat{\mathbf{a}}$, et d'une erreur $\hat{\mathbf{e}}$ (relation {5}). Pour calculer la variance des \mathbf{y} , il faut calculer la somme des carrés de cette erreur, soit $\hat{\mathbf{e}}'\hat{\mathbf{e}}$.

D'après la relation {5}, on peut écrire :

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{a}}$$

La somme des carrés des résidus est égale à $\hat{\mathbf{e}}'\hat{\mathbf{e}}$.

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{a}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}) = (\mathbf{y}' - \hat{\mathbf{a}}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}),$$

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\mathbf{a}} - \hat{\mathbf{a}}'\mathbf{X}'\mathbf{y} + \hat{\mathbf{a}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{a}},$$

On simplifie cette expression en remarquant que $\mathbf{y}'\mathbf{X}\hat{\mathbf{a}} = \hat{\mathbf{a}}'\mathbf{X}'\mathbf{y}$ (un scalaire est égal à son transposé) et que $\hat{\mathbf{a}}'(\mathbf{X}'\mathbf{X})\mathbf{X}^{-1} = \mathbf{y}'$ (relation {3}).

D'où

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{y} - 2\hat{\mathbf{a}}'\mathbf{X}'\mathbf{y} + \hat{\mathbf{a}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{a}},$$

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{a}}'\mathbf{X}'\mathbf{y},$$

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} \quad \{6\}$$

La somme des carrés des résidus est égale à la somme des carrés des réponses mesurées moins la somme des carrés des réponses calculées avec le modèle.

On montre que les calculs effectués avec le critère des moindres carrés conduisent à une somme des résidus nulles. Ces résidus sont donc bien des écarts à la moyenne qui vaut zéro. On peut les utiliser pour calculer la variance des réponses.

Degrés de liberté

Pour calculer la variance il faut encore connaître le nombre de degrés de liberté des résidus. Nous sommes partis d'un système de n équations indépendantes. Les réponses

mesurées ont donc n degrés de liberté. Le modèle a utilisé p degrés de liberté pour établir les p coefficients du modèle. Il reste donc $n-p$ degrés de liberté pour les résidus.

Variance des réponses

Appliquons la définition

$$V(y_i) = \frac{{}'\hat{\mathbf{e}}\hat{\mathbf{e}}}{n-p} = \frac{{}'\mathbf{yy} - {}'\hat{\mathbf{y}}\hat{\mathbf{y}}}{n-p} \quad \{7\}$$

Écart-type des réponses

Appliquons la définition

$$\sigma_{y_i} = \sqrt{V(y_i)} = \sqrt{\frac{{}'\hat{\mathbf{e}}\hat{\mathbf{e}}}{n-p}} = \sqrt{\frac{{}'\mathbf{yy} - {}'\hat{\mathbf{y}}\hat{\mathbf{y}}}{n-p}} \quad \{8\}$$

Variance et écart-type des coefficients

Les statisticiens ont démontré, sous l'adoption d'hypothèses simplificatrices, que la variance des coefficients était égale à

$$V(\hat{\mathbf{a}}) = V(y_i) ({}'\mathbf{X}\mathbf{X})^{-1} \quad \{9\}$$

Les écarts-types des coefficients sont égaux aux racines carrées des éléments diagonaux de la matrice $\sigma_{y_i}^2 ({}'\mathbf{X}\mathbf{X})^{-1}$.

Pourquoi RMSE et non écart-type

Dans la définition de la variance, il est précisé que les valeurs doivent être purement aléatoires. Or les résidus sont composés de deux grandeurs : une grandeur aléatoire qui tient compte du fait que les réponses sont elles-mêmes aléatoires et d'un terme traduisant le fait que le modèle mathématique choisi n'est peut être pas le bon. Ce dernier terme, encore appelé "manque d'ajustement" (lack of fit) n'est pas aléatoire. En toute rigueur, il n'est pas possible de considérer la variance que nous venons de calculer comme une variance au sens de la définition. Dans beaucoup de livres de statistique, le ou les auteurs admettent que le modèle mathématique choisi est le bon et qu'il n'y a pas de terme de manque d'ajustement. Avec cette hypothèse, la variance des résidus est la variance des réponses mesurées. Mais, dans le domaine des plans d'expériences, il n'est pas possible d'admettre cette hypothèse puisque l'on cherche justement le modèle qui explique le mieux les résultats expérimentaux. On suppose que le modèle est faux et on cherche le moins mauvais. On cherche donc à diminuer la RMSE (Root Mean Square Error ou racine carrée du carré moyen des résidus) pour tendre vers l'estimation du plus faible écart-type et vers le meilleur modèle.