

Question de Damien Steyer

"Quel est le calcul exact du lack of fit et l'obtention de sa probabilité lors d'une Anova ? »

Réponse de Jacques Goupy

Principe du calcul

Le « lack of fit » (en anglais) ou « Manque d'ajustement » (en français) mesure la différence qui existe entre le modèle a priori (modèle postulé) qui a été choisi par l'expérimentateur et le modèle réel qui régit le phénomène étudié.

Le modèle mathématique a priori choisi par l'expérimentateur est le plus souvent, dans le cas des plans d'expériences, un modèle polynomial de la forme :

$$y = a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + \sum a_{ii} x_i^2 + \dots \quad \{1\}$$

- y est la grandeur à laquelle s'intéresse l'expérimentateur. C'est la réponse ou la grandeur d'intérêt.
- x_i représente un niveau du facteur i ,
- x_j représente un niveau du facteur j .
- a_0, a_i, a_{ij}, a_{ii} sont les coefficients du polynôme.

Mais la réponse est une grandeur aléatoire et le modèle postulé n'est pas forcément le bon modèle, on écrit donc l'équation {1} en introduisant l'écart e .

$$y = a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + \sum a_{ii} x_i^2 + \dots + e \quad \{2\}$$

Il faut décomposer l'écart en deux compléments :

1. Une variable aléatoire, ε , qui tient compte du fait que la réponse est elle-même une grandeur aléatoire. Cette variable possède, comme toutes variables aléatoires, une distribution, une variance et une moyenne. Il n'est pas nécessaire dans un premier temps de préciser ces propriétés. Cette variable s'appelle l'erreur pure.
2. Une variable non aléatoire, Δ , qui représente la différence entre le modèle réel et le modèle a priori. Cette variable s'appelle le manque d'ajustement ou lack of fit.

Le modèle complet est donc :

$$y = a_0 + \sum a_i x_i + \sum a_{ij} x_i x_j + \sum a_{ii} x_i^2 + \dots + \Delta + \varepsilon \quad \{3\}$$

Dans la plupart des manuels de statistique, on suppose que le modèle mathématique représente parfaitement le phénomène étudié et on suppose que $\Delta = 0$. Mais il n'est pas possible de faire de même quand on réalise une expérimentation. En effet, ce que l'on cherche c'est le modèle mathématique le plus proche possible du modèle réel que l'on ne

connaît pas. Il serait donc maladroit de supposer au départ que le modèle mathématique est connu.

Dans le modèle {3} on connaît les réponses, y_i , et les niveaux des facteurs, x_i . En revanche, on ignore la valeurs des coefficients a_0, a_i, a_{ij}, a_{ii} , etc., la valeur de Δ et la valeur de ε .

Pour trouver les inconnues on exécute plusieurs expériences en faisant varier les niveaux des facteurs et en mesurant les réponses. On obtient ainsi un système d'équations. Ce système à n équations s'il y a n réponses mesurées. Il possède $n + p + 2$ inconnues s'il y a p coefficients. Il faut donc trouver $p + 2$ équations supplémentaires si l'on veut résoudre le système. La solution classique de ce problème est basée sur le critère des moindres carrés.

On calcule le carré de $\Delta + \varepsilon$ et l'on cherche le jeu des coefficients qui minimise ce carré. On obtient ainsi p équations supplémentaires. Si l'on s'arrête là, on peut calculer les coefficients mais pas $\Delta + \varepsilon$. C'est le cas notamment quand réalise un plan complet sans répétitions ni essais supplémentaires.

Il faut donc faire de nouveaux essais pour obtenir de nouvelles réponses et de nouvelles équations. C'est la raison des expériences réalisées au centre du domaine d'étude. Ces expériences apportent une équation supplémentaire qui permet de calculer le manque de d'ajustement et elles permettent d'obtenir une évaluation de l'erreur pure.

Décomposition en sommes de carrés

L'analyse de la variance est d'abord une décomposition de la somme des carrés des écarts à la moyenne des réponses mesurées. Les réponses sont donc centrées par rapport à leur moyenne. Dans le cadre des moindres carrés la moyenne, \bar{y} , des réponses mesurées, y , est la même que celle des réponses calculées, \hat{y} . Les écarts e prennent alors des valeurs particulières, ils sont alors dénommés "résidus" et notés \hat{e} . La somme des résidus est nulle. On a

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{i=n} (\hat{e})^2$$

Maintenant si des répétitions ont été effectuées on peut décomposer la somme des carrés des résidus en deux carrés, l'un correspondant à l'erreur pure, l'autre au manque d'ajustement. Le produit $\Delta\varepsilon$ étant nul dans le cadre des moindres carrés, on a

$$\sum_{i=1}^{i=n} (\hat{e})^2 = \sum_{i=1}^{i=n} (\Delta)^2 + \sum_{i=1}^{i=n} (\varepsilon)^2$$

La décomposition en sommes de carrés est donc :

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = \sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{i=n} (\Delta)^2 + \sum_{i=1}^{i=n} (\varepsilon)^2$$

L'analyse de la variance est basée sur cette dernière relation en faisant intervenir des hypothèses statistiques supplémentaires dont la normalité, l'homoscédasticité, la moyenne nulle des variables aléatoires, l'indépendance statistique des écarts, les degrés de liberté des sommes de carrés, etc.

Exemple

Reprenons l'exemple 2 du livre "Introduction aux plans d'expériences. Il s'agissait d'un plan complet sans expériences au centre du domaine d'étude. On obtenait les coefficients du modèle mais il était impossible de calculer l'erreur expérimentale et le lack of fit.

Supposons que nous puissions faire quatre expériences au centre du domaine d'étude comme l'indique le tableau ci-dessous.

On adopte comme modèle mathématique a priori

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_{12}x_1x_2 + a_{13}x_1x_3 + a_{23}x_2x_3$$

Tableau 1 : Plans d'expériences : Les bijoux

Or (1)	Densité de courant (2)	Cobalt (3)	Vitesse observée	Vitesse calculée
-1	-1	-1	53	54,75
1	-1	-1	122	121,75
-1	1	-1	20	19,75
1	1	-1	125	126,75
-1	-1	1	48	47,75
1	-1	1	70	71,75
-1	1	1	68	69,75
1	1	1	134	133,75
0	0	0	84	80,75
0	0	0	83	80,75
0	0	0	80	80,75
0	0	0	82	80,75

SOMME DES CARRÉS DES RÉPONSES MESURÉES

La moyenne vaut :

$$\bar{y} = \frac{1}{12} (53 + 122 + 20 + 125 + 48 + 70 + 68 + 134 + 84 + 83 + 80 + 82) = 80,75$$

D'où le calcul de la somme des carrés des réponses mesurées :

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = (53 - 80,75)^2 + (122 - 80,75)^2 + \dots + (83 - 80,75)^2 + (82 - 80,75)^2 = 12324,25$$

$$\sum_{i=1}^{i=n} (y_i - \bar{y})^2 = 12324,25$$

SOMME DES CARRÉS DES RÉPONSES CALCULÉES

On calcule les réponses avec le modèle a priori

$$\sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 = (54,75 - 80,75)^2 + (121,75 - 80,75)^2 + \dots + (80,75 - 80,75)^2 + (80,75 - 80,75)^2 = 12294$$

$$\sum_{i=1}^{i=n} (\hat{y}_i - \bar{y})^2 = 12294$$

SOMME DES CARRÉS DES RÉSIDUS (ERREUR PURE + MANQUE D'AJUSTEMENT)

La somme des carrés des résidus est égale à la différence entre la somme des carrés des réponses mesurées et la somme des carrés des réponses calculées :

$$\sum_{i=1}^{i=n} (\hat{\varepsilon})^2 = 12324,25 - 12294 = 30,25$$

SOMME DES CARRÉS DE L'ERREUR PURE

La moyenne des réponses mesurées au centre est

$$\bar{y} = \frac{1}{4} (84 + 86 + 83 + 82) = \frac{335}{4} = 83,75$$

D'où la somme des carrés correspondante :

$$\sum_{i=1}^{i=n} (\varepsilon)^2 = (84 - 82,25)^2 + (83 - 82,25)^2 + (80 - 82,25)^2 + (82 - 82,25)^2$$

$$\sum_{i=1}^{i=n} (\varepsilon)^2 = 3,0625 + 0,5625 + 5,0625 + 0,0625$$

$$\sum_{i=1}^{i=n} (\varepsilon)^2 = 8,75$$

CALCUL DU MANQUE D'AJUSTEMENT

La somme des carrés du manque d'ajustement est égale à la différence entre la somme des carrés des résidus et la somme des carrés de l'erreur pure.

$$\sum_{i=1}^{i=n} (\Delta)^2 = \sum_{i=1}^{i=n} (\hat{\varepsilon})^2 - \sum_{i=1}^{i=n} (\varepsilon)^2$$

$$\sum_{i=1}^{i=n} (\Delta)^2 = 30,25 - 8,75 = 21,5$$

Maintenant si l'on veut aller plus loin dans l'analyse de ces résultats, il faut introduire les hypothèses statistiques classiques de normalité, d'homoscédasticité, d'indépendance, de degrés de liberté, etc. des variables statistiques. Le rapport de Fisher mesure alors la probabilité que le manque d'ajustement et l'erreur pure soit du même ordre de grandeur. Un rapport de Fisher faible correspond à une forte probabilité que ces deux variables soient du même ordre de grandeur.

Ici le rapport de Fisher est égal à

$$F = \frac{21,5/2}{8,75/3} = 3,68$$

Ce qui correspond à une probabilité de 0,155. Si l'on refait 100 fois le plan d'expériences, l'erreur pure et le manque d'ajustement seront équivalents environ 15 fois.