

I. Analyse de régression

1. Modèle et estimation

Le modèle général a la forme : $y_i = b_0 + b_1 x_{1i} + \dots + b_p x_{pi} + e_i$. Les hypothèses sont :

- la valeur de y (la variable dépendante) dépend d'une combinaison linéaire des variables x (les variables indépendantes ou explicatives).
- la seule composante aléatoire du modèle est le terme ε d'erreur.
- les erreurs entre les différentes observations ne sont pas corrélées. Elles sont de moyenne nulle, de variance constante et distribuées normalement.

La représentation matricielle du modèle est la suivante : $y = Xb + e$ où $X = \begin{bmatrix} 1 & x_{10} & x_{p0} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{pn} \end{bmatrix}$

$$\text{et } b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}.$$

Le vecteur estimé des paramètres est noté b . L'estimation des paramètres est obtenue par la minimisation de la somme des carrés des erreurs du modèle (algorithme des moindres carrés). Les valeurs estimées de la variable dépendante sont obtenues par $\hat{y} = Xb$. La somme des carrés des erreurs d'estimation est obtenue par $(y - \hat{y})'(y - \hat{y})$, ou encore $(y - Xb)'(y - Xb)$ qu'il s'agit de minimiser. La solution est $b = (X'X)^{-1}(X'y)$. Il n'existe donc de solution que si $(X'X)$ est inversible, ce qui implique qu'aucune variable explicative ne soit la combinaison linéaire d'autres variables.

2. Test du modèle

On cherche à tester l'hypothèse de base du modèle de régression, à savoir qu'il existe une relation linéaire entre les différentes variables indépendantes et la variable dépendante :

$$H_0 : \beta_0 = \beta_1 = \dots = \beta_p = 0$$

On peut, sur la base de l'estimation du vecteur des coefficients b , obtenir les grandeurs suivantes :

- la somme du carré des erreurs (SSE) : $SSE = e'e$ où $e = y - \hat{y}$.
- la somme du carré de la régression (SSR) : $SSR = (\hat{y} - \bar{y})'(\hat{y} - \bar{y})$.
- la somme totale des carrés (SST) : $SST = SSE + SSR$.
- le nombre de degré de liberté des résidus : nombre d'individus - nombre de paramètres à estimer (DFRES = $n - (p+1)$).
- le nombre de degré de liberté de la régression : DFREG = p (nombre de variables indépendantes).
- le nombre total de degré de liberté : DFTOT = DFRES + DFREG = $n-1 = p+n-(p+1)$

La variance des résidus est donc donnée par $MSE = \frac{SSE}{DFRES}$ et la variance de la régression par

$MSR = \frac{SSR}{DFREG}$. L'hypothèse H_0 est alors testée à l'aide du test de Fisher : $F = \frac{MSR}{MSE}$ avec DFREG et DFRES degrés de liberté.

3. Test des paramètres

Il s'agit de se poser la question de savoir, pour chaque paramètre, quelle est la probabilité qu'il soit nul.

$$H_0 : \beta_i = 0$$

Pour tester cette hypothèse, on utilise la statistique de student : $t = \frac{b_i}{SEb_i}$ où b_i est la valeur estimée

de β_i et SEb_i est l'erreur standard du paramètre b_i . L'estimation de SEb_i se fonde sur le calcul de la matrice de variance/covariance entre les différentes variables et l'estimation de l'erreur moyenne MSE. SEb_i est donc le terme diagonale n° i de la matrice $(X^tX)^{-1}MSE$. Le nombre de degrés de liberté est DFRES.

4. Qualité de l'ajustement

La qualité de l'ajustement peut être évaluée à l'aide du coefficient de détermination $R^2 = \frac{SSR}{SST}$.

Cependant, dans le contexte de la régression multiple, cela pose le problème de la sur-paramétrisation du modèle. Plus l'on ajoute de variables explicatives, plus le R^2 augmente, même si les nouvelles variables explicatives sont très "reliées" à la variable dépendante. Pour éviter ce phénomène, on calcule le coefficient de détermination ajusté :

$$R_{adj}^2 = 1 - \frac{\frac{SSE}{DFRES}}{\frac{SST}{DFTOT}}$$

ce qui permet de tenir compte à la fois de l'accroissement du nombre de variables explicatives et de la réduction de SSE par rapport à SST.

Il faut également noter que le coefficient de détermination est le carré du coefficient de corrélation et représente donc bien un indicateur d'association linéaire.

5. Multi-colinéarité

L'existence de trop fortes dépendances linéaires entre les variables indépendantes fausse l'estimation des paramètres.

Une première étape intéressante est toujours de calculer la matrice de corrélation entre les différentes variables. Cela permet de faire apparaître les trop fortes interdépendances entre les variables explicatives. Malheureusement, seules les interactions directes apparaissent.

Pour tester les interactions entre trois variables, on utilise un modèle de régression (pour tester, par exemple, les interactions entre X_2 et l'effet combiné de X_1 et X_3 , on régresse ces dernières sur X_2). On

prend alors la racine carrée du coefficient de détermination pour estimer le coefficient de corrélation multiple entre les variables. La procédure peut cependant devenir lourde à mettre en oeuvre vu l'aspect combinatoire des interactions possibles.

6. Sélection des variables

Si l'on veut être certain de trouver la meilleure combinaison des variables indépendantes qui explique le comportement de la variable dépendant, l'approche la plus directe et la plus lente de tester toutes les combinaisons possibles. Le nombre de combinaisons est 2^p où p est le nombre de variables explicatives...

Deux autres procédures existent :

- l'approche en arrière : partant du modèle complet, on élimine progressivement les variables les moins significatives. La variable la moins significative est celle dont la suppression provoque la plus petite baisse de la statistique de Fisher.
- l'approche en avant : on part d'un modèle avec une seule variable explicative. La variable retenue initialement est la plus significative (celle qui a le plus grand coefficient de corrélation avec la variable explicative). On ajoute ensuite une seconde variable (celle qui permet le plus fort accroissement du test de Fisher).

Remarque : les Fisher des sous-modèles sont appelés F partiels.