

Doc 4. Modéliser. Régression linéaire.

1 Le principe.....	1
2 Un peu de théorie :	1
2.1 Régression linéaire simple.....	1
2.1.1 Les données.....	1
2.1.2 Le résultat.....	2
2.1.3 Le mécanisme	2
2.1.4 La mesure de la qualité de l'adéquation du modèle aux données.....	2
2.2 Régression linéaire multiple :	3
3 Obtenir une régression linéaire sous Pspp.....	3
4 Les tests d'ajustement : Fisher et Student.....	3
4.1 Le test de Student , sur 1 variable explicative.....	4
4.2 Le test de Fisher, sur l'ensemble des variables explicatives.....	4
4.3 Vérification des hypothèses sur les résidus.	5

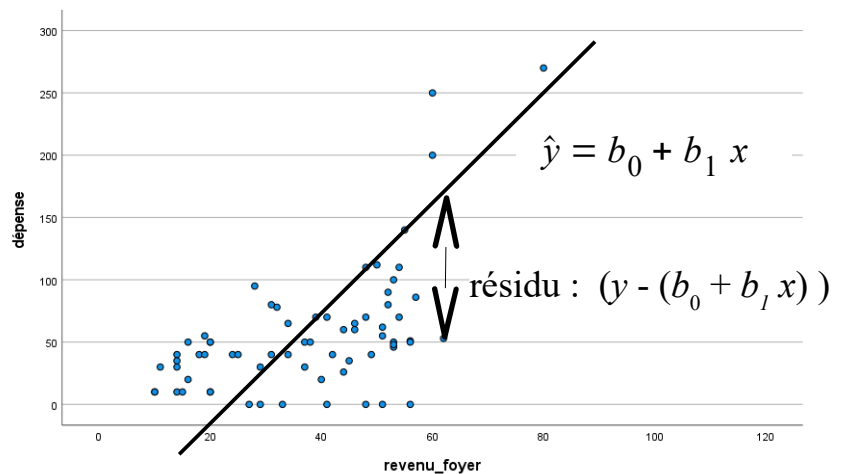
1 Le principe

Voilà un nuage de points représentant deux variables : dépense et revenu du foyer.

On cherche à **résumer l'ensemble de ces données par une fonction affine.**

$$\widehat{Dépense} = b_0 + b_1 \text{revenu_foyer}$$

Le graphe de cette fonction affine est une droite, graphiquement on cherche donc une droite qui résume au mieux un ensemble de points.



On étendra ensuite ce procédé pour ajuster une fonction linéaire de plusieurs variables, par exemple

$$\widehat{Dépense} = b_0 + b_1 \text{revenu_foyer} + b_2 \text{taille_foyer} + b_3 \text{âge}$$

2 Un peu de théorie :

2.1 Régression linéaire simple.

2.1.1 Les données

On dispose de n observations

- d'une variable Y numérique – appelée "variable endogène" , ou "variable expliquée" mais ce terme peut être trompeur, ou encore « variable dépendante »
- et d'une variable X numérique – appelée "variable exogène" , ou "variable explicative", ou encore « variable indépendante » – .

2.1.2 Le résultat

On obtient

- Deux coefficients b_0, b_1 qui définissent la fonction affine cherchée : $\hat{y} = b_0 + b_1 x$
- Pour chaque observation n° j, une **valeur ajustée**, ou prévue, ou encore expliquée, par le modèle, de la variable endogène : $\hat{y}_j = b_0 + b_1 x_j$
- Pour chaque observation n° j, un **résidu** $u_j = \underbrace{y_j}_{\text{valeur observée}} - \underbrace{\hat{y}_j}_{\text{valeur ajustée}}$

2.1.3 Le mécanisme

Réaliser une régression linéaire simple de la variable Y endogène sur la variable X exogène consiste à **ajuster** aux observations de X et Y une fonction affine :

$$y = b_0 + b_1 x$$

c'est-à-dire à déterminer les coefficients $\hat{b}_0$ et $\hat{b}_1$ qui minimisent la somme des carrés des écarts entre les observations et les points de cette droite aux mêmes abscisses : $\sum_{1 \leq j \leq n} (b_0 + b_1 x_j - y_j)^2$

Par exemple on cherche les coefficients β_0, β_1 qui permettent que

$\beta_0 + \beta_1 * (\text{revenu_foyer})$ s'approche au mieux de dépenses.

« au mieux » reste subjectif : dans le plan représentant les données (x, y) il y a plusieurs points et une droite d'équation $\beta_0 + \beta_1 * (\text{revenu_foyer})$ ne peut pas les réunir tous.

Pour choisir une droite parmi toutes, un critère s'est dégagé et a été adopté unanimement : on choisit la droite qui minimise $\sum_{1 \leq j \leq n} (\hat{y}_j - y_j)^2 = \sum_{1 \leq j \leq n} (b_0 + b_1 x_j - y_j)^2$.

Il y a une, et une seule, solution, la droite définie par :

$$b_0 = \overline{y} - b_1 \bar{x} \quad \text{et} \quad b_1 = \frac{\overline{xy} - (\bar{x})(\bar{y})}{\bar{x}^2 - (\bar{x})^2}$$

On appelle résidu n° j la quantité $u_j = y_j - \hat{y}_j = y_j - \beta_1 x_j - \beta_0$.

La droite de régression linéaire est celle qui minimise la somme des termes $(u_j)^2$.

2.1.4 La mesure de la qualité de l'adéquation du modèle aux données

On appelle **Coefficient de détermination** le nombre $R^2 = \frac{\sum_{1 \leq j \leq n} (\hat{y}_j - \bar{y})^2}{\sum_{1 \leq j \leq n} (y_j - \bar{y})^2}$.

On peut montrer que $\underbrace{\bar{\hat{y}}}_{\text{moyenne des valeurs ajustées}} = \underbrace{\bar{y}}_{\text{moyenne des observations}}$, d'où $\bar{u} = 0$, et que

$$\underbrace{\frac{1}{n} \sum_{1 \leq j \leq n} (y_j - \bar{y})^2}_{\text{variance des observations}} = \underbrace{\frac{1}{n} \sum_{1 \leq j \leq n} (\hat{y}_j - \bar{y})^2}_{\text{variance expliquée}} + \underbrace{\frac{1}{n} \sum_{1 \leq j \leq n} (u_j^2)}_{\text{variance des résidus}}$$

$$\text{d'où } R^2 = 1 - \frac{\sum_{1 \leq j \leq n} u_j^2}{\sum_{1 \leq j \leq n} (y_j - \bar{y})^2}, \text{ donc}$$

- R^2 est toujours compris entre 0 et 1.
- Le modèle linéaire est d'autant mieux ajusté aux données que R^2 est proche de 1.
- R^2 apparaît comme la proportion de la variance des observations contenue dans la variance des valeurs ajustées $\hat{y}_j$ on exprime couramment que R^2 est **la proportion de la variance des observations qui est expliquée par le modèle**.

2.2 Régression linéaire multiple :

Il y a plusieurs variables exogènes $x_1, x_2, \dots, x_k$, on cherche à ajuster aux observations $y_1, y_2, \dots, y_n$ une fonction affine des $x_1, x_2, \dots, x_k$:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

On définit à nouveau, pour toute observation n°i, le résidu $y_i - \hat{y}_i$ et on choisit à nouveau les coefficients $\beta_1, \beta_2, \dots, \beta_k$ qui minimisent la somme des carrés des résidus, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$. On définit de la même façon le

coefficient de détermination
$$R^2 = \frac{\sum_{j=1}^n (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2} = 1 - \frac{\sum_{j=1}^n u_j^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$$

3 Obtenir une régression linéaire sous Jamovi

On peut utiliser Analyses → **Régression** → **Régression linéaire**. Une fenêtre s'ouvre et demande de choisir la variable dépendante y et les « covariables » $x_1, x_2, \dots, x_k$. Dans cette fenêtre, on peut cliquer Coefficients du modèle, « intervalles de confiance ». Cela permet d'afficher un intervalle de confiance à un niveau de confiance de 95% sur les coefficients.

Dans la fenêtre résultats, les coefficients s'affichent dans un tableau nommé Coefficients.

Prédicteur	Estimation	Erreur standard	t	p
Ordonnée à l'origine	15.655	13.152	1.190	0.238
revenu_oyer	0.293	0.400	0.731	0.467
taille_oyer	8.749	4.912	1.781	0.080

Le coefficient de détermination R^2 apparaît dans le tableau « Mesure de l'ajustement ».

Modèle	R	R^2
1	0.349	0.122

Q1 trouver la fonction affine $b_0 + b_1 \times \text{revenu_oyer}$ qui fasse la meilleure approximation possible de *dépense* – on dit qu'on **ajuste** le modèle $\text{dépense} = b_0 + b_1 \text{revenu_oyer} + \text{résidu}$

Q2 Donner le coefficient de détermination.

Q3 Selon le résultat de la question 1, que se passe-t-il, en moyenne, lorsque revenu_oyer augmente d'une unité ? En déduire **quelle est la signification de la valeur de b_1 trouvée à la question précédente**.

Q4 ajuster le modèle $\text{dépense} = b_0 + b_1 \text{revenu_oyer} + b_2 \text{taille_oyer} + \text{résidu}$

Quel est le coefficient de détermination ? Que valent b_1 et b_2 ?

4 Les tests d'ajustement : Fisher et Student

On **suppose** à partir d'ici que les observations sont réellement **issues d'un modèle**

$$y = b_0 + b_1 x + u \quad \text{ou} \quad y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k + u \quad \text{où}$$

les résidus u sont des variables aléatoires

- indépendantes,

- d'espérance 0,
- de même variance,
- de loi Gaussienne.

On peut alors établir les 2 tests statistiques suivants pour juger de l'adéquation du modèle aux données. Dans le cas d'une régression simple, il s'agit du même test sous deux expressions différentes, il est inutile de l'appliquer deux fois.

4.1 Le test de Student , sur 1 variable explicative

On teste l'**hypothèse nulle H0** : $b_i = 0$ pour un entier $i \leq k$,

la variable x_i , toutes choses égales par ailleurs, c'est-à-dire compte tenu des autres variables exogènes, n'a pas d'influence linéaire sur y

contre l'**hypothèse adverse H1** : $b_i \neq 0$ la variable x_i a une influence linéaire sur y.

Statistique utilisée : $t = \frac{\hat{b}_i}{\widehat{\sigma}_{\hat{b}_i}}$

Sous les hypothèses citées et l'hypothèse H0, t suit la loi de Student T(n- 2)

Règle de décision : on fixe un risque d'erreur de première espèce $\alpha = P_{H0 \text{ vraie}}(\text{rejeter } H0)$

A l'aide d'une table on peut trouver une valeur $t_{\max}$ telle que $P(t > t_{\max}) = P(t < -t_{\max}) = \alpha / 2$

Si $|t| > t_{\max}$ on rejette H0, on accepte H1, selon laquelle X_i a une influence linéaire sur Y.

Si $-t_{\max} < t < t_{\max}$ on accepte H0 selon laquelle X_i , toutes choses égales par ailleurs, n'a pas d'influence linéaire sur Y.

Jamovi fournit un seuil de significativité – ou p-valeur – du test de Student dans le tableau « Coefficients » colonne "p".

Prédicteur	Estimation	Erreur standard	t	p
Ordonnée à l'origine	-899.807	865.973	-1.04	0.303
taille_oyer	11.191	3.775	2.96	0.004
naissance	0.465	0.437	1.06	0.292

On **rejette H0**, et on accepte H1, selon laquelle

- **X_i a une influence linéaire significative sur Y**, dans le cas d'une seule variable explicative
- **X_i a une influence linéaire significative sur Y, toutes choses égales par ailleurs**, dans le cas où il y a plusieurs variables explicatives

si le **seuil de significativité** est **inférieur** au risque d'erreur de 1ère espèce α choisi.

4.2 Le test de Fisher, sur l'ensemble des variables explicatives

Hypothèses H0 et H1 :

H0 : $\beta_1 = \beta_2 = \beta_3 \dots = \beta_k = 0$: l'ensemble des variables explicatives n'a aucune influence dans le modèle.

H1 : *au moins 1 des β_i , $1 \leq i \leq k$, est différent de 0* : Les variables explicatives, ensemble – mais pas forcément chacune d'elles – ont une influence significative dans le modèle.

La **statistique utilisée** est $f = \frac{(n-k-1)}{k} \frac{R^2}{1-R^2}$, dans le cas de la régression simple f est aussi égale à t^2 .

Sous les mêmes hypothèses et H0, la statistique f suit la loi de Fisher de paramètres k et n-k-1, F(k,n-k-1)

Règle de décision

On fixe un risque de 1ère espèce $\alpha = P(\text{rejeter } H0 / H0 \text{ vraie})$.

On rejette H0, et on accepte H1, si f dépasse une valeur maximale donnée par une table de la loi de Fisher.

Mesures de l'ajustement du modèle						
Modèle	R	R ²	Test de modèle général			
			F	ddl1	ddl2	p
1	0.360	0.130	4.85	2	65	0.011

. Pour faire apparaître les résultats du test de Fisher, à gauche de la fenêtre, dans « Ajustement du modèle », cocher « test F ». Jamovi **fournit un seuil de significativité** du test de Fisher dans la cellule la plus à droite du tableau « **p** ». On rejette H_0 , et on accepte H_1 selon laquelle au moins 1 des coefficients n'est pas nul, si ce seuil est inférieur au risque α choisi

Q5 Les modèles ajustés aux questions précédentes étaient-ils adaptés aux données ?

Q6 Comment juger quelle variable, parmi *revenu_ foyer* et *taille_ foyer*, a l'influence la plus visible sur *dépense* ?

Q7 Ajuster le modèle $dépense = b_0 + b_1 \text{revenu_foyer} + b_2 \text{genre} + b_3 \text{sophistiquée} + u$.

1. Le modèle dans son ensemble est-il bien adapté aux données ?

2. Que vaut le coefficient b_2 ?

3. **Comment interpréter ce coefficient?** Rappel : genre vaut 1 pour les hommes, 2 pour les femmes.

Q8 Ajuster le modèle $dépense = b_0 + b_1 \text{revenu_foyer} + b_2 \text{fréquentez_vous} + u$

Q9 Comparer l'estimation de b_1 aux questions 1 et 7. Comment interpréter la différence ?

Q10 Ajuster le modèle

$$\text{sophistiquée} = b_0 + b_1 \text{marques} + b_2 \text{sobre} + b_3 \text{classique} + b_4 \text{rock} + b_5 \text{variétés} + b_6 \text{rap} + u$$

Quelles sont les variables explicatives à conserver ? Que penser des valeurs et des signes des coefficients ?

4.3 Vérification des hypothèses sur les résidus.

Si les hypothèses probabilistes ne sont pas réalisées, les tests ne sont pas valables. Vous allez ici pouvoir vérifier l'une d'elles : les résidus suivent-ils une loi Gaussienne ?

Q11

a) Ouvrir à nouveau la fenêtre de régression pour construire la régression de la question 4.

Dans la section Vérification des hypothèses, cliquer sur « test de normalité ». Il s'agit d'un test de Shapiro-Wilk. L'hypothèse nulle est : les résidus sont issus d'une loi Gaussienne, l'hypothèse adverse : ils sont issus d'une autre loi.

Qu'en pensez-vous ?

4.4 Récupération des calculs :

Dans la section « sauvegarder », cocher « valeurs prédites ». Regardez les données, une nouvelle variable est créée, elle contient les valeurs prédites : $b_0 + b_1 \text{revenu_foyer} + b_2 \text{taille_foyer}$