Démarche statistique
Premiers pas avec R

David Causeur
L'Institut Agro
IRMAR CNRS UMR 6625
https://dcauseur.netlify.app

Plan

Plan

Plan

Données appariées

Etude de cas 'fictive' : 2 produits alimentaires évalués par 3 juges sur une échelle de préférence allant de 1 ('je n'aime pas') à 10 ('j'aime').

	Juges		
Produits	J_1	J_2	J_3
A	2	3	6
B	4	6	8

- Données et première analyse dans R

> Soit Y_{ij} la variable réponse mesurée par le $j$ème individu, $1 \leq j \leq J$, dans le $i$ème groupe, $1 \leq i \leq I$.
> Si le $j$ème individu est le même dans tous les groupes, alors on dit que les données sont **appariées**.

Données appariées

Peut-on sérieusement conclure que l'effet 'produit' n'est pas significatif ?

Produits	Juges		
	J_1	J_2	J_3
A	2	3	6
B	4	6	8

Analyse de la variance pour données appariées

Soit Y_{ij} la variable réponse mesurée par le *j*ème individu, $1 \leq j \leq J$, dans le *i*ème groupe, $1 \leq 1 \leq I$:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

où

- α_i, $i = 2, \ldots, I$, paramètres de l'effet *groupe* ($\alpha_1 = 0$),
- β_j, $j = 2, \ldots, J$ paramètres de l'effet *individu* ($\beta_1 = 0$).

et l'erreur résiduelle $e_{ij} \sim \mathcal{N}(0; \sigma)$.

Remarque. Le modèle à un facteur est un sous-modèle du modèle à deux facteurs : $\varepsilon_{ij} = \beta_j + e_{ij}$.

Ajustement du modèle pour données appariées

Minimisation du critère des moindres carrés :

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 = \min_{\mu,\alpha_i,\beta_j}\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \mu - \alpha_i - \beta_j)^2.$$

- Ajustement du modèle dans `R`

Ajustement du modèle pour données appariées

Estimateurs des paramètres :

$$
\begin{aligned}
\hat{\mu} &= \overline{Y}_{\bullet\bullet} + (\overline{Y}_{1\bullet} - \overline{Y}_{\bullet\bullet}) + (\overline{Y}_{\bullet 1} - \overline{Y}_{\bullet\bullet}), \\
\hat{\alpha}_i &= \overline{Y}_{i\bullet} - \overline{Y}_{1\bullet},\ i = 2, \dots, I, \\
\hat{\beta}_j &= \overline{Y}_{\bullet j} - \overline{Y}_{\bullet 1},\ j = 2, \dots, J.
\end{aligned}
$$

	Juges			
Produits	J_1	J_2	J_3	$\overline{Y}_{i\bullet}$
A	2	3	6	3.67
B	4	6	8	6.00
$\overline{Y}_{\bullet j}$	3.00	4.50	7.00	4.83

Ajustement du modèle pour données appariées

Résidus : $\hat{e}_{ij} = Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet}$.

Estimation de la variance résiduelle σ^2 :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet})^2}{(I-1)(J-1)}.$$

- Comparaison des écarts-types résiduels dans `R`

Test de Fisher pour données appariées

Equation d'analyse de la variance :

$\sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \overline{Y}_{\bullet\bullet})^2$

$$= \sum_{i=1}^{I} J(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \overline{Y}_{i\bullet})^2,$$

$$= \sum_{i=1}^{I} J(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 + \sum_{j=1}^{J} I(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet}\right)^2.$$

- Table d'analyse de la variance dans `R`

Plan

Corrélation

- Description graphique du lien entre LMP et BFAT dans `R`

Corrélation

Un indicateur de la linéarité de l'effet de X sur Y ne dépend

- ni de la position,
- ni de la dispersion

des distributions marginales de X et Y.

... il peut être évalué à partir des séries **centrées-réduites.**

Les valeurs $\tilde{x}_i$ de la série $(x_1, \ldots, x_n)$ **centrée-réduite** s'obtiennent de la manière suivante :

$$\tilde{x}_i = \frac{x_i - \overline{x}}{s_x}.$$

▶ Données centrées-réduites dans R

Corrélation

Le **coefficient de corrélation** r_{xy} est la moyenne des produits des valeurs centrées-réduites :

$$r_{xy} = \frac{\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i}{n-1}.$$

De manière équivalente,

$$r_{xy} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{s_x s_y} = \frac{s_{xy}}{s_x s_y},$$

où s_{xy} est la **covariance** des séries x et y :

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}.$$

Corrélation

r_{xy} **s'interprète comme suit** :

- $r_{xy} \approx 1$ peut être un bon indicateur d'une relation linéaire croissante entre X et Y.
- $r_{xy} \approx -1$ peut être un bon indicateur d'une relation linéaire décroissante entre X et Y.
- $r_{xy} \approx 0$ peut être un bon indicateur d'une absence de relation linéaire entre X et Y.

r_{xy} doit **compléter** l'impression visuelle déduite d'un graphique (nuage de points).

▶ Coefficient de corrélation linéaire dans `R`

Régression linéaire

Relation entre LMP (Y) et épaisseur de gras (X) : la manière dont $\mathbb{E}(Y \mid X = x)$ dépend de x est décrite par une fonction linéaire de x.

On suppose que, sachant $X = x$, Y suit une loi normale, de même écart-type pour tout x.
Il y a un **effet linéaire de X sur Y** si l'espérance conditionnelle de Y sachant $X = x$ est une fonction linéaire de x :

$$\mathbb{E}(Y \mid X = x) = \beta_0 + \beta_1 x,$$

où β_0 est le **terme constant** et β_1 le **coefficient directeur**. Le modèle ci-dessus s'appelle **modèle de régression linéaire simple.**

Régression linéaire

Pourquoi *régression*?

Le sens statistique du terme *régression* vient de

Francis Galton (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. **15**. 246–263

qui s'intéresse à l'héritabilité du phénotype *taille* chez l'humain.

Régression linéaire

De manière alternative,

Le **modèle de régression linéaire simple** de l'effet linéaire de X sur Y est le suivant :

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

où $\varepsilon = Y - \mathbb{E}(Y \mid X = x) = Y - \beta_0 - \beta_1 x$ est appelé **erreur résiduelle**.

Sachant $X = x$, ε est distribué selon une loi normale avec

- $\mathbb{E}(\varepsilon \mid X = x) = 0$;
- et $\mathrm{Var}(\varepsilon \mid X = x) = \sigma^2$.

Méhode des moindres carrés

Minimisation du **critère des moindres carrés** $\mathrm{SS}(\beta_0, \beta_1)$:

$$\mathrm{SS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Les **estimateurs des moindres carrés** $\hat{\beta}_0$ et $\hat{\beta}_1$ minimisent $\mathrm{SS}(\beta_0, \beta_1)$.

- Ajustement du modèle dans `R`

Estimateurs des moindres carrés

- **Estimateur de β_0 :**

$$\hat{\beta}_0 \quad = \quad \overline{Y} - \hat{\beta}_1 \overline{x}.$$

... la droite de régression ajustée passe par l'individu 'moyen', de coordonnées $(\overline{x}, \overline{y})$.

- **Estimateur de β_1 :**

$$\hat{\beta}_1 \quad = \quad \frac{s_{xy}}{s_x^2}.$$

Estimateurs des moindres carrés

Pour un individu avec $X = x$, la **valeur ajustée de la réponse** est $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

De la même manière, la **droite de régression ajustée** a pour équation $x \mapsto \hat{\beta}_0 + \hat{\beta}_1 x$: c'est la droite *la plus proche* des données.

▶ Droite de régression dans `R`

Erreur d'ajustement

Qualité de l'ajustement mesurée par :

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimation de σ^2 :

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-2}.$$

- Ecart-type résiduel dans `R`

Effet groupe ou effet linéaire - Résumé

Y a t'il un effet de *ceci* sur *cela* ?

1. Effet groupe : *ceci* est une variable catégorielle
   - Modèle d'analyse de la variance à un facteur
   - Modèle d'analyse de la variance à deux facteurs si données appariées
2. Effet linéaire : *ceci* est une variable quantitative
   - Modèle de régression linéaire (simple)
   - A suivre : effets linéaires par groupe ...

Une seule fonction d'ajustement : `mod=lm(y~x1+x2)`

- Tests des effets : `anova(mod)`
- Analyse des effets : `summary(mod)`