# Démarche statistique
# Premiers pas avec R

David Causeur
*Agrocampus Ouest*
*IRMAR CNRS UMR 6625*
*http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/david.causeur*

# Plan

# Plan

# Plan

# Plan

# Intervalle de confiance pour les paramètres

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des combinaisons linéaires des $Y_i$ :

$$\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\overline{x}}{s_x^2}(Y_i-\overline{Y}), \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\overline{x}}{s_x^2}Y_i. \\
&= \sum_{i=1}^{n}\omega_i(x)Y_i, \text{ avec } \omega_i(x) = \frac{1}{n-1}\frac{x_i-\overline{x}}{s_x^2}
\end{aligned}$$

# Intervalle de confiance pour les paramètres

$\hat{\beta}_0$ et $\hat{\beta}_1$ sont des combinaisons linéaires des $Y_i$ :

$$\hat{\beta}_1 = \sum_{i=1}^{n} \omega_i(x) Y_i.$$

Comme combinaison linéaire des $Y_i$, indépendants et suivant une loi normale, $\hat{\beta}_1$ suit lui-même une loi normale :

$$\mathbb{E}(\hat{\beta}_1 \mid X = x) = \beta_1 \qquad [\hat{\beta}_1 \text{ est non-biaisé.}]$$

$$\mathrm{Var}(\hat{\beta}_1 \mid X = x) = \frac{\sigma^2}{n-1} \frac{1}{s_x^2}.$$

# Intervalle de confiance pour les paramètres

Comme combinaison linéaire des $Y_i$, indépendants et suivant une loi normale, $\hat{\beta}_1$ suit lui-même une loi normale :

$$\text{Sachant } X = x,\ \hat{\beta}_1 - \beta_1 \sim \mathcal{N}(0; \frac{\sigma}{\sqrt{n-1}} \frac{1}{s_x})$$

La **précision de l'estimation** est favorisée par :

- un faible écart-type résiduel $\sigma$ (**bonne adéquation du modèle aux données**),
- une **grande taille d'échantillon** $n$,
- une **grande dispersion des valeurs de** $x$.

# Intervalle de confiance pour les paramètres

**Résumé** : $\hat{\beta}_0$ et $\hat{\beta}_1$ suivent une loi normale de moyennes $\beta_0$ et $\beta_1$ respectivement et d'écarts-types $\sigma_{\hat{\beta}_0}$ et $\sigma_{\hat{\beta}_1}$ respectivement :

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2}{n-1}\left[\frac{n-1}{n} + \frac{1}{s_x^2}\right], \; \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{n-1}\frac{1}{s_x^2}.$$

Par conséquent,

$$\mathrm{CI}_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j - t^{(n-2)}_{1-\alpha/2}\hat{\sigma}_{\hat{\beta}_j} ; \hat{\beta}_j + t^{(n-2)}_{1-\alpha/2}\hat{\sigma}_{\hat{\beta}_j}\right], \; j = 0 \text{ ou } j = 1$$

où $\hat{\sigma}_{\hat{\beta}_j}$ est obtenu à l'aide de l'estimateur $\hat{\sigma}^2$ de $\sigma^2$

- Intervalles de confiance des coefficients dans R

# Bande de confiance pour la droite de régression

On appelle **bande de confiance** pour la droite de régression, de niveau de confiance $1 - \alpha$, et on note $\mathrm{CB}_{1-\alpha}(\beta)$ la famille suivante d'intervalles de confiance :

$$\mathrm{CB}_{1-\alpha}(\beta) = \{\mathrm{CI}_{1-\alpha}(\beta_0 + \beta_1 x^\star);\ \text{pour tout } x^\star\},$$

Estimation de $\beta_0 + \beta_1 x^\star$ :

$$\begin{aligned}\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 x^\star = \sum_{i=1}^{n} h_i(x, x^\star) Y_i, \\ &= \text{combinaison linaire des } Y_i.\end{aligned}$$

▶ Intervalle de confiance de prédiction dans R

# Bande de confiance pour la droite de régression

Comme combinaison linéaire des $Y_i$, $\hat{Y}$ suit une loi normale :

$$\begin{aligned}
\mathbb{E}(\hat{Y} \mid X = x) &= \beta_0 + \beta_1 x^\star . \\
\mathrm{Var}(\hat{Y} \mid X = x) &= \frac{\sigma^2}{n}\left[1 + \frac{n}{n-1}\left(\frac{x^\star - \overline{x}}{s_x}\right)^2\right].
\end{aligned}$$

Par conséquent, $\mathrm{CI}_{1-\alpha}(\beta_0 + \beta_1 x^\star) =$

$$\left[\hat{Y} - t^{(n-2)}_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{n}{n-1}\left(\frac{x^\star - \overline{x}}{s_x}\right)^2} ; \hat{Y} + t^{(n-2)}_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1 + \frac{n}{n-1}\left(\frac{x^\star - \overline{x}}{s_x}\right)^2}\right],$$

# Bande de confiance pour la droite de régression

Soit $Y^\star$ la valeur non-observée de la variable réponse pour un individu tel que $X^\star = x^\star$ :

$$\hat{Y}^\star = \hat{\beta}_0 + \hat{\beta}_1 x^\star \text{ est la valeur prédite de } Y^\star$$

avec

$$\begin{aligned}
\mathbb{E}(Y^\star - \hat{Y}^\star \mid X = x,\ X^\star = x^\star) &= 0. \\
\mathrm{Var}(Y^\star - \hat{Y}^\star \mid X = x,\ X^\star = x^\star) &= \sigma^2 + \frac{\sigma^2}{n}\left[1 + \frac{n}{n-1}\left(\frac{x^\star - \overline{x}}{s_x}\right)^2\right].
\end{aligned}$$

**La prédiction la plus précise** est donc obtenue pour $x^\star = \overline{x}$.

# Test d'un effet linéaire

Dans quelle mesure

$\mathcal{M}$ : $Y = \beta_0 + \beta_1 x + \varepsilon$ avec RSS $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

s'ajuste-t'il mieux aux données que

$\mathcal{M}_0$ : $Y = \beta_0 + \varepsilon$ avec $\text{RSS}_0 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ ?

**Équation d'analyse de la variance** :

$$\text{RSS}_0 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \text{RSS}.$$

▶ Graphe ajusté versus observé dans R

# Test d'un effet linéaire

**Le coefficient $\mathbf{R^2}$** compare RSS et $RSS_0$ :

$$R^2 = \frac{RSS_0 - RSS}{RSS_0}.$$

- $0 \leq R^2 \leq 1$ ;
- $R^2 = 0$ : absence d'effet de $x$ ;
- $R^2 = 1$ : effet 'total' de $x$.

▶ Coefficient $R^2$ dans `R`

# Test de Fisher

**Test de l'effet linéaire de $X$ sur $Y$ :**

$$\begin{cases} H_0 : & \mathcal{M} \text{ ne s'ajuste pas mieux aux données que } \mathcal{M}_0 \\ H_1 : & \mathcal{M} \text{ s'ajuste mieux aux données que } \mathcal{M}_0 \end{cases}$$

**Statistique de test :**

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}/(n-2)}.$$

**Un degré de liberté pour** $\mathrm{RSS}_0 - \mathrm{RSS} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$ :

$$\hat{Y}_i - \overline{Y} = \hat{\beta}_1 (x_i - \overline{x}) \text{ proportionnel à } x_i - \overline{x}.$$

▶ Statistique de Fisher dans R

# Test de Fisher

$F = 88.369$ doit-il considéré comme une valeur élevée ?

La **loi de $F$ sous l'hypothèse nulle** est la loi de **Fisher** $\mathcal{F}_{1,n-2}$

- Analyse de la variance du modèle dans R

# Test de Fisher

**Table d'Analyse de la Variance** :

- `Df` : degrés de liberté, respectivement 1 et $n-2$ ;
- `Sum Sq` : sommes de carrés, respectivement $RSS_0 - RSS$ et RSS ;
- `Mean Sq` : moyennes des carrés, respectivement $(RSS_0 - RSS)/1$ et $RSS/(n-2)$ ;
- `F value` : Statistique de Fisher, le rapport des moyennes de carrés ;
- `Pr(>F)` : p-value, la probabilité qu'une statistique de Fisher soit plus grande que `F value` sous l'hypothèse nulle.

# Effets linéaires par groupes

La relation entre *LMP* et *épaisseur de gras* est-elle la même pour tous les types génétiques ?

Deux variables explicatives pour une même problématique :

- l'épaisseur de gras, **variable quantitative** ;
- et le type génétique, **variable catégorielle**.

Si l'effet d'une variable **n'est pas le même** selon la modalité d'une variable catégorielle, on parle d'**effet d'interaction** entre les deux variables explicatives.

# Régression linéaire avec effet 'groupe'

Soit $Y_{ij}$ la variable réponse pour le $j$ème individu, $j = 1, \ldots, n_i$, dans le $i$ème groupe, $i = 1, \ldots, I$ et $x_{ij}$ la valeur correspondante de la variable explicative.

Modèle de régression dans le **1er groupe ('référence')** :

$$Y_{1j} \quad = \quad \mu + \beta x_{1j} + \varepsilon_{1j}, \; \varepsilon_{1j} \sim \mathcal{N}(0; \sigma)$$

Modèle de régression dans le ***i*ème groupe**, avec $i \neq 1$ :

$$Y_{ij} \quad = \quad \mu + \alpha_i + (\beta + \gamma_i) x_{ij} + \varepsilon_{ij}, \; \varepsilon_{1j} \sim \mathcal{N}(0; \sigma)$$

où :

- $\alpha_2, \ \ldots, \ \alpha_I$ sont les **paramètres de l'effet 'groupe'** ;
- $\gamma_2, \ \ldots, \ \gamma_I$ sont les **paramètres d'interaction.**

# Régression linéaire avec effet 'groupe'

**Deux sous-modèles :**

- le **modèle d'analyse de la variance à un facteur** pour l'effet 'groupe', obtenu avec $\beta = 0$ et $\gamma_2 = \ldots = \gamma_I = 0$.

- le **modèle de régression linéaire simple** pour l'effet de $X$ sur $Y$, obtenu avec $\alpha_2 = \ldots = \alpha_I = 0$ et $\gamma_2 = \ldots = \gamma_I = 0$.

▶ Ajustement du modèle dans R

# Régression linéaire avec effet 'groupe'

Correspondance entre les résultats de `lm` et les paramètres :

| Paramètre | Nom dans `R` | Valeur estimée |
|---|---|---|
| $\mu$ | `(Intercept)` | 80.2215328 |
| $\beta$ | `BFAT` | -1.4575042 |
| $\alpha_2$ | `GENETP25` | -9.7600733 |
| $\alpha_3$ | `GENETP50` | -13.7187769 |
| $\gamma_2$ | `BFAT:GENETP25` | 0.6268832 |
| $\gamma_3$ | `BFAT:GENETP50` | 0.9550714 |

L'effet de l'épaisseur de gras sur LMP semble

- le plus évident pour le type génétique `P0`,
- moins clair pour le type génétique `P25`
- et encore moins clair pour le type génétique `P50`.

# Test d'un effet 'groupe' en régression linéaire

Dans quelle mesure

$$\mathcal{M} : \; Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i)x_{ij} + \varepsilon_{ij}$$

s'ajuste-t'il mieux aux données que

$$\mathcal{M}_0 : \; Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij} \, ?$$

- Test de l'effet d'interaction dans R

# Test d'un effet 'groupe' en régression linéaire

Table d'analyse de la variance :

- `Res.Df` : degrés de liberté de RSS ;
- `RSS` : somme des carrés des écarts résiduels ;
- `Df` : degrés de liberté de $\text{RSS}_0 - \text{RSS}$ ;
- `Sum of Sq` : gain d'ajustement $\text{RSS}_0 - \text{RSS}$ de `Model 2` par rapport à `Model 1` ;
- `F` : statistique de Fisher pour la comparaison de `Model 1` et `Model 2` ;
- `Pr(>F)` : p-value du test.