UCL
Université catholique de Louvain

# Cours de Biométrie
## LBIRA2101

# Recueil de formules

B. Goovaerts
A. El Gouch
X. Draye

**LBIRA2101 – Formulaire – Partie GLM (chapitres 1 à 5)**

## Equations des modèles étudiés

Modèle statistique à réponse quantitative continue normale (linéaire ou non linéaire) :

$$Y_i = f(x_{1i}, x_{2i}, \ldots, x_{ki}; \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon_i = f(\mathbf{x}_i; \theta) + \varepsilon_i \quad \text{avec } \varepsilon_i \sim iN(0, \sigma^2) \text{ et } i = 1, \ldots, N$$

Modèle linéaire simple : $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Modèle d'ANOVA 1 : $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ avec $\Sigma \alpha_i = 0 \quad i = 1, \ldots, m \quad$ et $\quad j = 1, \ldots, n_i$

Modèle d'ANOVA 2 : $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

Modèle d'analyse de covariance: $Y_{ijk} = \beta_0 + \alpha_j + \beta_1 x_i + \gamma_j x_i + \varepsilon_{ijk}$

Modèle linéaire par rapport aux paramètres : $g(Y_i) = \sum_{j=1}^{p} \beta_j f_j(X_{1i}, \ldots X_{k_1 i}, Q_{1i}, \ldots, Q_{k_2 i}) + \varepsilon_i$

## Modèle linéaire simple

Estimateurs des paramètres : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ et $\hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}$

$$\hat{\sigma}^2 = S^2 = \frac{1}{N-2} \sum_{i=1}^{N} e_i^2 = \frac{1}{n-2} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = MSE$$

Mesures d'ajustement $R^2 = 1 - \dfrac{SSE}{SST} \qquad R^2_{Ajusté} = 1 - \dfrac{MSE}{MST}$

Distributions d'échantillonnage des estimateurs

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{XX}}\right)\right) \qquad \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{S_{XX}}\right) \qquad \text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -S^2 \frac{\bar{x}^2}{S_{XX}}$$

$$\frac{(N-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

**Prédiction**

Réponse moyenne prédite en X=x* : $\hat{\mu}_{Y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \hat{Y}$

Intervalle de confiance pour la réponse moyenne

$$\left[\hat{Y}^* - t_{N-2;1-\alpha/2} S_{\hat{\mu}_{Y|x^*}}, \hat{Y}^* + t_{N-2;1-\alpha/2} S_{\hat{\mu}_{Y|x^*}}\right] \text{ avec } S^2_{\hat{\mu}_{Y|x^*}} = S^2 \left(\frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$$

Intervalle de prédiction pour une observation

$$\left[\hat{Y}^* - t_{N-2;1-\alpha/2} S_P, \hat{Y}^* + t_{N-2;1-\alpha/2} S_P\right] \text{ avec } S_P^2 = S^2 + S^2_{\hat{\mu}_{Y|x^*}} = S^2 \left(1 + \frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$$

Formule (générale) du test F de signification de la régression appliqué au modèle linéaire simple

$$F_{obs} = \frac{\left(SSE_{\text{Restreint}} - SSE_{Complet}\right)/(dl_R - dl_C)}{SSE_{Complet} / dl_C} = \frac{(SST - SSE)}{MSE} = \frac{MSM}{MSE} \sim F_{dl_R - dl_C, dl_C} = F_{1, N-2} \text{sous } H_0$$

Fonction de vraisemblance

$$L(\theta) = \prod_{i=1}^{N} p(y_i; \theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

## Modèle d'ANOVA I (facteur fixe)

Ecriture matricielle du modèle (exemple avec m=3, $n_j$=4 et N=12)

$$\mathbf{Y}=\begin{bmatrix}Y_{11}\\Y_{12}\\Y_{13}\\Y_{14}\\Y_{22}\\Y_{22}\\Y_{23}\\Y_{24}\\Y_{31}\\Y_{32}\\Y_{33}\end{bmatrix}=\begin{bmatrix}\mu+\alpha_1\\\mu+\alpha_1\\\mu+\alpha_1\\\mu+\alpha_1\\\mu+\alpha_2\\\mu+\alpha_2\\\mu+\alpha_2\\\mu+\alpha_2\\\mu+\alpha_3\\\mu+\alpha_3\\\mu+\alpha_3\end{bmatrix}+\begin{bmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{14}\\\varepsilon_{21}\\\varepsilon_{22}\\\varepsilon_{23}\\\varepsilon_{24}\\\varepsilon_{31}\\\varepsilon_{32}\\\varepsilon_{33}\end{bmatrix}=\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix}\times\begin{bmatrix}\mu\\\alpha_1\\\alpha_2\\\alpha_3\end{bmatrix}+\begin{bmatrix}\varepsilon_{11}\\\varepsilon_{12}\\\varepsilon_{13}\\\varepsilon_{14}\\\varepsilon_{21}\\\varepsilon_{22}\\\varepsilon_{23}\\\varepsilon_{24}\\\varepsilon_{31}\\\varepsilon_{32}\\\varepsilon_{33}\end{bmatrix}=\mathbf{X\beta}+\boldsymbol{\varepsilon}\qquad \mathbf{X}_{SAS}=\begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&0&1\\1&0&1\\1&0&0\\1&0&0\\1&0&0\end{bmatrix}$$

Equation de décomposition de variance

$$SS_{Totale}=SS_{Modèle}+SS_{Erreurs}=SS_{Traitements}+SS_{Erreurs}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

Estimateurs des paramètres

$$\hat{\mu}_i=\bar{Y}_i\quad\text{avec}\quad\bar{Y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{ij}\quad\text{et}\quad\hat{\sigma}^2=S^2=MSE=\frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

distribution d'échantillonnage et intervalle de confiance sur les moyennes

$$\bar{Y}_i\sim N(\mu_i,\frac{\sigma^2}{n_i})\qquad\qquad \bar{Y}_i+t_{N-m;1-\alpha/2}\sqrt{\frac{S^2}{n_i}}$$

## Modèle d'ANOVA 2 – facteurs fixes et plan balancé

Equation de décomposition de variance

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\bar{Y}..)^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(\bar{Y}_{ij}-\bar{Y}..)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\bar{Y}_{ij})^2$$

$$SS_{Totale}=SS_{Modèle}+SS_{Erreurs}$$

$$N-1\qquad ab-1\qquad N-ab\qquad \text{degrés de liberté}$$

$$SS_{Modèle}=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{ij}-\bar{Y}..)^2$$

$$=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{i.}-\bar{Y}..)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{j.}-\bar{Y}..)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{ij}-\bar{Y}_{i.}-\bar{Y}_{.j}+\bar{Y}..)^2$$

$$=SS_A+SS_B+SS_{AB}$$

Espérances des carrés moyens

$$E(MS_A) = \sigma^2 + nb\sum_{i=1}^{a}\alpha_i^2 \Big/ (a-1) \qquad E(MS_B) = \sigma^2 + na\sum_{j=1}^{b}\beta_j^2 \Big/ (b-1)$$

$$E(MS_{AB}) = \sigma^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2 \Big/ (a-1)(b-1) \qquad E(MS_{Erreur}) = \sigma^2$$

## Modèle linéaire général (facteurs fixes quantitatifs ou qualitatifs et erreurs iid)

Equation du modèle sous forme matricielle

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \quad \text{avec}\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_N)$$

$$Y \text{ est } N\times 1 \quad X \text{ est } N\times p \quad \boldsymbol{\beta} \text{ est } p\times 1 \quad \text{et } \boldsymbol{\varepsilon} \text{ est } N\times 1$$

Estimateurs des paramètres

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-}\mathbf{X'y} \qquad \hat{\sigma}^2 = S^2 = \frac{1}{N-p}\sum_{i=1}^{N}(Y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)^2 = \frac{1}{N-p}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

Matrice de variance covariance des paramètres : $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1}$

Distribution d'une combinaison linéaire des paramètres estimés:

$$L\hat{\beta} \sim N(L\beta, \sigma^2 L(X'X)^{-1}L')$$

Prédiction en un nouveau vecteur X=x*

$$\hat{Y}^* = \hat{\mu}_{Y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \mathbf{x}^{*\prime}\hat{\boldsymbol{\beta}}$$

Intervalle de confiance sur la réponse moyenne et intervalle de prédiction

$$\left[\hat{Y}^* - t_{N-p;1-\alpha/2}S(\hat{\mu}_{Y|x^*}), \hat{Y}^* + t_{N-p;1-\alpha/2}S(\hat{\mu}_{Y|x^*})\right] \quad \text{avec} \quad S^2_{\hat{\mu}_{Y|x^*}} = S^2(\mathbf{x}^{*\prime}(\mathbf{X'X})^{-1}\mathbf{x}^*)$$

$$\left[\hat{Y}^* - t_{N-p;1-\alpha/2}S_P, \hat{Y}^* + t_{N-p;1-\alpha/2}S_P\right] \quad \text{avec} \quad S_P^2 = S^2(1 + \mathbf{x}^{*\prime}(\mathbf{X'X})^{-1}\mathbf{x}^*)$$

Test F général pour modèles emboités

H0 : modèle restreint $\mathbf{y} = \mathbf{X_1\beta_1} + \boldsymbol{\varepsilon}_R$ H1 : modèle complet $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}_C = \mathbf{X_1\beta_1} + \mathbf{X_2\beta_2} + \boldsymbol{\varepsilon}_C$

Statistique de test

$$F_{obs} = \frac{\left(SSE_{\text{Restreint}} - SSE_{Complet}\right)/(dl_R - dl_C)}{SSE_{Complet}/dl_C} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{L}'\left(\mathbf{L}(\mathbf{X'X})^{-1}\mathbf{L}'\right)^{-1}\mathbf{L}\hat{\boldsymbol{\beta}}/p_2}{SSE_{Complet}/(N-p)} \sim \mathrm{F}_{p_2,N-p} \quad \text{avec}\ \mathbf{L\beta} = \boldsymbol{\beta}_2$$

Estimation et tests généraux sur des combinaisons de paramètres ou contrastes

Combinaison linéaire de paramètres : $\mathbf{l\beta} = l_0\beta_0 + l_1\beta_1 + \ldots + l_{p-1}\beta_{p-1}$

Distribution d'échantillonnage d'une CL : $\mathbf{l}\hat{\boldsymbol{\beta}} \sim N(\mathbf{l\beta}, \sigma^2\mathbf{l}(\mathbf{X'X})^{-}\mathbf{l}')$

Intervalle de confiance sur une CL : $\mathbf{l\beta} : \mathbf{l}\hat{\boldsymbol{\beta}} \pm t_{N-r,1-\alpha/2}\, s\sqrt{\mathbf{l}(\mathbf{X'X})^{-}\mathbf{l}'}$

Test sur une combinaison linéaire simple

$$H_0\ \mathbf{l\beta} = \mathbf{l\beta}_0 \leftrightarrow H_1\ \mathbf{l\beta} \neq \mathbf{l\beta}_0 \qquad t_{obs} = \frac{\mathbf{l}\hat{\boldsymbol{\beta}} - \mathbf{l\beta}_0}{\sqrt{s^2(\mathbf{l}(\mathbf{X'X})^{-}\mathbf{l}')}} \sim t_{N-r} \ \text{sous}\ \mathrm{H}_0$$

Test sur un contraste multiple

$$H_0\ \mathbf{L}\hat{\beta} = 0 \leftrightarrow H_1\mathbf{L}\hat{\beta} \neq 0 \qquad F_{obs} = \frac{\left(\mathbf{L}\hat{\beta}\right)(\mathbf{L}(\mathbf{X'X})^{-}\mathbf{L}')^{-1}\mathbf{L}\hat{\beta}/q}{S^2} \sim F_{q,N-r} \ \text{sous}\ \mathrm{H}_0$$

Fonction de vraisemblance

$$L(\theta) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-\frac{N}{2}}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\beta}'\mathbf{X})'(\mathbf{y} - \boldsymbol{\beta}'\mathbf{X})\right)$$

## Modèles linéaires aléatoires et mixtes, variables quantitatives (chapitres 6 à 9)

### Equation des modèles étudiés

Modèle d'ANOVA 1 aléatoire : $Y_{ij} = \mu + a_i + \varepsilon_{ij}$

avec $i = 1..m \;\; j = 1..n \quad a_i \sim iN(0,\sigma_a^2) \quad \varepsilon_i \sim iN(0,\sigma^2)$

Modèle d'ANOVA 2 aléatoire hiérarchisé : $Y_{ijk} = \mu + a_i + b_{j(i)} + \varepsilon_{ijk}$

avec $i = 1..a \;\; j = 1..b \;\; k = 1..n_{ij} \quad a_i \sim iN(0,\sigma_a^2) \quad b_{j(i)} \sim iN(0,\sigma_b^2) \quad \varepsilon_{ijk} \sim iN(0,\sigma^2)$

Modèle mixte : $Y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk}$

avec $i = 1..a \;\; j = 1..b \;\; k = 1..n \quad b_j \sim iN(0,\sigma_b^2) \quad (\alpha b)_{ij} \sim iN(0,\sigma_{\alpha b}^2) \quad \varepsilon_{ijk} \sim iN(0,\sigma^2)$

Modèle de plan en bloc aléatoire complet : $Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$

avec $i = 1..a, \quad j = 1..b \quad b_j \sim iN(0,\sigma_b^2) \quad \varepsilon_{ij} \sim iN(0,\sigma^2)$

### Modèle d'ANOVA I aléatoire – méthode GLM

Ecriture matricielle du modèle : $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ (cf ANOVA I fixe)

Equation de décomposition de la variance :

$SST = SSA + SSE$ avec $N-1, m-1$ et $N-m$ degrés de liberté $(N = n.m)$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_i-\bar{Y})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2$$

Espérance des carrés moyens : $E(MSE) = \sigma^2 \quad E(MSA) = \sigma^2 + n\sigma_a^2$

Test d'hypothèse du modèle :

$H_0 : \sigma_a^2 = 0 \quad vs \quad H_1 : \sigma_a^2 > 0$

$F_{obs} = \dfrac{MSA}{MSE} \sim F(m-1, n-1)$ sous $H_0$

Estimateurs des paramètres :

$$\hat{\mu} = \bar{Y} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}Y_{ij} \quad \hat{\sigma}^2 = S^2 = MSE = \frac{1}{N-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2 \quad \hat{\sigma}_a^2 = (MSA - MSE)/n$$

Distribution d'échantillonnage et intervalle de confiance sur la moyenne :

$$\bar{Y} \sim N(\mu, \frac{n\sigma_a^2 + \sigma^2}{N}) \qquad \bar{Y} \pm t_{m-1;1-\alpha/2}\sqrt{\frac{MSA}{N}}$$

Différence avec le modèle fixe :

$\sigma_Y^2 = \sigma_a^2 + \sigma^2$ et dès lors $\mathrm{cov}(Y_{ij}, Y_{ij'}) = \sigma_a^2$ et $\mathrm{cov}(Y_{ij}, Y_{i'j'}) = 0$

## Modèle d'ANOVA I aléatoire – méthode MIXED

Ecriture matricielle du modèle : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ , avec

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0},\mathbf{G}) \quad \mathbf{G} = \sigma_a^2 \mathbf{I}_m$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0},\mathbf{R}) \quad \mathbf{R} = \sigma^2 \mathbf{I}_N$$

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{Xb},\mathbf{V}) \quad \text{où} \quad \mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{V}_m \end{bmatrix} \quad \text{où} \quad \mathbf{V}_i = \begin{bmatrix} \sigma_a^2 + \sigma^2 & & \sigma_a^2 \\ & \ddots & \\ \sigma_a^2 & & \sigma_a^2 + \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n + \sigma_a^2 \mathbf{1}_{n \times n}$$

Fonction de vraisemblance :

$$l = \log L = -\tfrac{N}{2} \log 2\pi - \tfrac{1}{2} \log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Estimateurs des paramètres par maximum de vraisemblance :

$$\hat{\mu} = \overline{Y} \qquad \hat{\sigma}_a^2 = \tfrac{1}{n}\left[(1 - \tfrac{1}{m})MSA - MSE\right] \qquad \hat{\sigma}^2 = MSE$$

Fonction de vraisemblance retreinte :

avec $\mathbf{K}$ tel que $\mathbf{K}'\mathbf{X} = 0$

$$l = \log L = -\tfrac{1}{2} r_{\mathbf{K}} \log 2\pi - \tfrac{1}{2} \log|\mathbf{K}'\mathbf{V}\mathbf{K}| - \tfrac{1}{2} \mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}$$

Estimateurs des paramètres par maximum de vraisemblance restreint :

$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ , ce qui donne pour une ANOVA 1 :

$$\hat{\mu} = \overline{Y} \qquad \hat{\sigma}_a^2 = (MSA - MSE)/n \qquad \hat{\sigma}^2 = MSE$$

Test d'hypothèse par modèles emboîtés :

$H_0$ : Modèle restreint : $Y_{ij} = \mu + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0,\sigma^2)$

$H_1$ : Modèle complet : $Y_{ij} = \mu + a_i + \varepsilon_{ij} \quad a_i \sim N(0,\sigma_a^2) \quad \varepsilon_{ij} \sim N(0,\sigma^2)$

avec des vraisemblances *L(R)* et *L(F)*

$LRT = -2\left[\log_e L(R) - \log_e L(F)\right] \sim \chi_p^2$ avec $p$=1

## Modèle d'ANOVA 2 aléatoire, facteurs hiérarchisés

Equation de décomposition de variance :

$$SST = SSM + SSE \qquad \text{avec } N-1,\ ab-1 \text{ et } N-ab \text{ degrés de liberté} \quad N = abn$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y})^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij}-\bar{Y})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{ij})^2$$

$$SSM = SSA + SSB(A) \qquad \text{avec } ab-1,\ a-1 \text{ et } ab-a \text{ degrés de liberté}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ij}-\bar{Y})^2 = bn\sum_{i=1}^{a}(\bar{Y}_{i.}-\bar{Y})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij}-\bar{Y}_{i.})^2$$

Espérance des carrés moyens :

$$E(MSA) = \sigma^2 + nb\sigma_a^2 + n\sigma_b^2 \qquad E(MSB(A)) = \sigma^2 + n\sigma_b^2 \quad \text{et} \quad E(MSE) = \sigma^2$$

Estimation des paramètres :

$$\hat{\sigma}_a^2 = [MSA - MSB(A)]/nb \qquad \hat{\sigma}_b^2 = [MSB(A) - MSE]/n \quad \text{et} \quad \hat{\sigma}^2 = MSE$$

Tests d'hypothèse basé sur les carrés moyens :

$$H_0 : \sigma_a^2 = 0 \quad H_1 : \sigma_a^2 > 0 \qquad MSA / MSB(A) \sim F(a-1, a(b-1)) \quad sous\ H_0$$

$$H_0 : \sigma_b^2 = 0 \quad H_1 : \sigma_b^2 > 0 \qquad MSB(A) / MSE \sim F(a(b-1), N-ab)) \quad sous\ H_0$$

Formulation matricielle du modèle (méthode MIXED) :

$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$ , avec

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0},\mathbf{G}) \quad \mathbf{G} = \sigma_a^2\mathbf{I}_m$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0},\mathbf{R}) \quad \mathbf{R} = \sigma^2\mathbf{I}_N$$

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{Xb},\mathbf{V}) \quad \text{où} \quad \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

Structure de la matrice $\mathbf{G}$ :

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_b \end{bmatrix} \quad \text{où} \quad \mathbf{G}_a = \sigma_a^2\mathbf{I}_a \quad \text{et} \quad \mathbf{G}_b = \sigma_b^2\mathbf{I}_b$$

Structure de la matrice $\mathbf{V}$ (pour un example où $a = b = n = 2$)

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \quad \text{où} \quad \mathbf{V}_i = \begin{bmatrix} \sigma_a^2+\sigma_b^2+\sigma^2 & \sigma_a^2+\sigma_b^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2+\sigma_b^2 & \sigma_a^2+\sigma_b^2+\sigma^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_b^2+\sigma^2 & \sigma_a^2+\sigma_b^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2+\sigma_b^2 & \sigma_a^2+\sigma_b^2+\sigma^2 \end{bmatrix}$$

Tests d'hypothèse selon la méthode MIXED:

Appliquer la méthode des modèles emboîtés séparément pour chaque facteur

Distribution d'échantillonnage et intervalle de confiance sur la moyenne :

$$\bar{Y} \sim N\left(\mu, \frac{nb\sigma_a^2 + n\sigma_b^2 + \sigma^2}{N}\right) \qquad \bar{Y} \pm t_{a-1;1-\alpha/2}\sqrt{\frac{MSA}{N}}$$

## Modèle mixte, facteurs croisés

Ecriture matricielle du modèle : $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ , avec

$$\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{G}) \quad \mathbf{G} = \sigma_a^2 \mathbf{I}_m$$

$$\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{R}) \quad \mathbf{R} = \sigma^2 \mathbf{I}_N$$

$$\mathbf{Y} \sim \mathbf{N}(\mathbf{Xb}, \mathbf{V}) \quad \text{où} \quad \mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$$

$$\mathbf{Y} = \begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \\ 1 & 1 & \\ 1 & & 1 \\ 1 & & 1 \\ 1 & & 1 \\ 1 & & 1 \end{bmatrix} \times \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} 1 & & 1 & & & \\ 1 & & 1 & & & \\ & 1 & & 1 & & \\ & 1 & & 1 & & \\ 1 & & & & 1 & \\ 1 & & & & 1 & \\ & 1 & & & & 1 \\ & 1 & & & & 1 \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ \alpha b_{11} \\ \alpha b_{21} \\ \alpha b_{12} \\ \alpha b_{22} \end{bmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Structure de la matrice $\mathbf{G}$ :

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\alpha b} \end{bmatrix} \quad \text{où} \quad \mathbf{G}_a = \sigma_a^2 \mathbf{I}_a \quad \text{et} \quad \mathbf{G}_{\alpha b} = \sigma_{\alpha b}^2 \mathbf{I}_{a.b}$$

Structure de la matrice $\mathbf{V}$ (pour un example où $a = b = n = 2$)

$$\mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}_3 = \begin{bmatrix} \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2 & \sigma_b^2 + \sigma_{\alpha b}^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 + \sigma_{\alpha b}^2 & \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2 & \sigma_b^2 + \sigma_{\alpha b}^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_{\alpha b}^2 & \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2 \end{bmatrix}$$

Conséquences de la définition du modèle :

$$Y_{ijk} \sim N[\mu + \alpha_i, \sigma_b^2 + \sigma_{\alpha b}^2 + \sigma^2]$$

Equation de décomposition de la variance : cf ANOVA 2 (facteurs fixes et croisés)

Espérance des carrés moyens :

$$E(MSA) = \sigma^2 + nb\frac{\sum \alpha_i^2}{a-1} + n\sigma_{\alpha b}^2 \qquad E(MSB) = \sigma^2 + na\sigma_b^2$$

$$E(MSAB) = \sigma^2 + n\sigma_{\alpha b}^2 \qquad E(MSE) = \sigma^2$$

Test d'hypothèse sur le facteur fixe basé sur les carrés moyens :

$$H_0 : \alpha_1 = .. = \alpha_a = 0 \qquad \frac{MSA}{MSAB} \sim F(a-1, (a-1)(b-1)) \quad \text{sous } \mathrm{H}_0$$

Méthode générale d'inférence sur une combinaison des effets fixes et aléatoires :

Combinaison linéaire $\mathbf{L}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}$ avec $\mathbf{L} = \left[\ \mu \mid \alpha_1 \ \ \alpha_2 \mid b_1 \ \ b_2 \mid \alpha b_{11} \ \ \alpha b_{12} \ \ \alpha b_{21} \ \ \alpha b_{22}\ \right]$

Intervalle de confiance pour $\mathbf{L}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}$: $\qquad \mathbf{L}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{u} \end{bmatrix} \pm t_{\hat{\upsilon}, \alpha/2}\sqrt{\mathbf{L}\hat{\mathbf{C}}\mathbf{L}'}$

où $\mathbf{C}$, la matrice variance-covariance de $\boldsymbol{\beta}$ et $\mathbf{u}$

Test d'hypothèse général sur une combinaison des effets fixes et aléatoires :

$$H_0 : \mathbf{L}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \mathbf{0} \quad H_1 : \mathbf{L}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} \neq \mathbf{0}$$

Contraste simple : $\dfrac{\mathbf{L}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{u} \end{bmatrix}}{\sqrt{\mathbf{L}\hat{\mathbf{C}}\mathbf{L}'}} \underset{approx.}{\sim} t_{\hat{\upsilon}}$

Contraste multiple et test général : $\dfrac{\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{u} \end{bmatrix}' \mathbf{L}'(\mathbf{L}'\hat{\mathbf{C}}\mathbf{L})^{-1}\mathbf{L}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{u} \end{bmatrix}}{rank(\mathbf{L})} \underset{approx.}{\sim} F(rank(\mathbf{L}), \hat{\upsilon})$

## Modèle mixte, cas du plan en blocs aléatoires complets

Particularité par rapport au modèle mixte croisé :

$\varepsilon_{ij} = Y_{ij} - \hat{Y}_{ij}$ est égal à $(\alpha b)_{ij} = Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..}$ du modèle mixte croisé

Espérance des carrés moyens :

$$E(MSA) = \sigma^2 + b\frac{\sum \alpha_i^2}{a-1} \qquad E(MSB) = \sigma^2 + a\sigma_b^2 \qquad E(MSAB) = \sigma^2$$

Structure de la matrice $\mathbf{V}$ :

$$\mathbf{V} = \begin{bmatrix} \mathbf{V_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V_2} \end{bmatrix} \quad \text{avec} \quad \mathbf{V}_1 = \mathbf{V}_2 = \begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 \end{bmatrix}$$

# RÉGRESSION LOGISTIQUE

$Y$ une variable de Bernoulli, i.e. $Y \in \{0,1\} \equiv \{\text{échec, succès}\}$.

- Si on dispose d'un échantillon iid $Y_1, \ldots, Y_n$ de $Y$, alors l'EMV de $p = P(Y=1) \equiv P(S)$ est

$$\hat{p} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{\text{nombre de S}}{\text{nombre total}} \equiv \frac{s}{n} \sim_a N(p, p(1-p)/n).$$

- On appelle **cote ou chance** de succès le rapport

$$o(S) = \frac{P(S)}{P(E)} = \frac{p}{1-p},$$

- Soit $Z$ une v.a. qui prend deux valeurs : "$1 \equiv$ G1" et "$2 \equiv$ G2". Soit $p_1 = P(Y=1|Z=1)$ et $p_2 = P(Y=1|Z=2)$. Le rapport

$$or(S) = \frac{o_1}{o_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

  est appelé **rapport des cotes** de succès. L'EMV de $or$ est

$$\hat{or} = \frac{s_1 e_2}{e_1 s_2},$$

  Pour une taille de l'échantillon grande,

$$\ln(\hat{or}) \sim_a N\left(\ln(or), \sigma^2\right), \text{ avec } \hat{\sigma}^2 = 1/s_1 + 1/e_1 + 1/s_2 + 1/e_2.$$

- Soit $X$ une variable continue. On note par $p(x) = P(Y=1|X=x)$. et $o(x) = \frac{p(x)}{1-p(x)}$. Le modèle logistique stipule que

$$logit(p(x)) \equiv \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$
$$\Leftrightarrow p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}} \equiv logistic(\beta_0 + \beta_1 x).$$

  *Interprétation des paramètres :* $e^{\beta_0} = o(0)$ et $e^{\beta_1} = \frac{o(x+1)}{o(x)}$.

- L'EMV $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ de $\boldsymbol{\beta} = (\beta_0, \beta_1)$ satisfait

$$\hat{\boldsymbol{\beta}} \sim_a N_2(\boldsymbol{\beta}, V_{\boldsymbol{\beta}}),$$

  où $V_{\boldsymbol{\beta}}$ est une matrice de variance covariance.

- Au niveau $\alpha$, pour tester l'hypothèse $H_0 : \beta_1 = \beta_1^0$ contre $H_1 : \beta_1 \neq \beta_1^0$, la p-valeur (asymptotique) est

$$P(\chi_1^2 > z^2),$$

où $Z = \frac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma}_1} \sim_a N(0,1)$.
- Un intervalle de confiance (asymptotique) à 95% pour $\beta_1$ est donné par

$$\hat{\beta}_1 \pm 1.96\hat{\sigma}_1.$$

- Pour une valeur de $x$ donnée, le modèle estime que

$$\hat{p}(x) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x)}} \equiv logistic(\hat{\beta}_0 + \hat{\beta}_1 x)$$

Soit $\hat{\sigma}^2(x) = \hat{\sigma}_0^2 + x^2\hat{\sigma}_1^2 + 2x\hat{\sigma}_{01}$. Un intervalle de confiance (asymptotique) à 95% pour $p(x_1)$ est donné par

$$logistic((\hat{\beta}_0 + \hat{\beta}_1 x) \pm 1.96\hat{\sigma}(x)).$$

- Si le modèle s'ajuste bien, alors les valeurs prédites seront proches des valeurs observées. Dans ce cas,

$$e_j = \frac{\hat{s}_j - s_j}{\sqrt{t_j \hat{p}_j (1 - \hat{p}_j)}} \approx N(0,1), \text{ (pearson)}$$

avec $\tilde{p}_j = s_j / t_j$, $\hat{s}_j = t_j \hat{p}_j$ et $\hat{\sigma}_j = \hat{\sigma}(x_j)$.
- Pour obtenir une probabilité de succès $p$ il faut fixer $x$ à $x_p = \frac{logit(p) - \beta_0}{\beta_1}$. L'EMV de $x_p$ est

$$\hat{x}_p = \frac{logit(p) - \hat{\beta}_0}{\hat{\beta}_1} \sim_a N(x_p, \sigma^2(x_p)/\beta_1^2).$$