Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

OLS #11 (régression sur des variables qualitatives)

Loi Multinomiale

$\mathbf{Y} = (Y_1, \cdots, Y_d) \sim \mathcal{M}(\mathbf{p})$ où $\mathbf{p} = (p_1, \cdots, p_d)$ si

$$Y_1 + \cdots + Y_d = 1 \text{ et } Y_j \sim \mathcal{B}(p_j),\ \forall j \in \{1, \cdots, d\}$$

i.e. $\mathbf{Y} = (\mathbf{1}_{C_1}, \mathbf{1}_{C_2}, \cdots, \mathbf{1}_{C_d})$

$\mathbf{Y} = (Y_1, \cdots, Y_d) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_1, \cdots, p_d)$ si

$$Y_1 + \cdots + Y_d = n \text{ et } Y_j \sim \mathcal{B}(n, p_j),\ \forall j \in \{1, \cdots, d\}$$

cf loi multinomiale. Pour
$(y_1, \cdots, y_d) \in \mathcal{S}_{d,n} = \{(y_1, \cdots, y_d) \in \mathbb{N}^d : (y_1 + \cdots + y_d = n\}$

$$\mathbb{P}[(Y_1, \cdots, Y_d) = (y_1, \cdots, y_d)] = \frac{n!}{y_1! \cdots y_d!} p_1^{y_1} \cdots p_d^{y_d}$$

Example: $\mathbf{Y} = (Y_0, Y_1) \sim \mathcal{M}(n, \mathbf{p})$ où $\mathbf{p} = (p_0, p_1)$.

On considère ici un variable qualitative, ou catégorielle, ou factorielle, x, prenant J modalités.
Pour $j = 1, \cdots, J$, on note n_j le nombre d'observations de la $j$ème modalité du facteur, $n = n_1 + \cdots + n_J$.
La variable d'intérêt est (toujours) Y. Notons $y_{i,j}$ la $i$ème observation du facteur j.

Example :

y	182	161	161	177	157	170	167	186	178	171	175
x	M	F	F	M	F	M	M	M	M	M	M

devient

y	182	177	170	167	186	178	171	175	161	161	157
x	M	M	M	M	M	M	M	M	F	F	F
$\mathbf{1}_M$	1	1	1	1	1	1	1	1	0	0	0
$\mathbf{1}_F$	0	0	0	0	0	0	0	0	1	1	1

ANOVA

On suppose que $y_{i,j} = \beta_0 + \beta_j + \varepsilon_{i,j}$, où $\mathbb{E}[\varepsilon_{i,j}] = 0$, $\text{cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = 0$ et $\text{Var}(\varepsilon_{i,j}) = 0$

On va disjoncter la variable de groupe $x \in \{1, 2, \cdots, J\}$ en J indicatrices, $(\mathbf{1}_1, \mathbf{1}_2, \cdots, \mathbf{1}_J)$

On peut noter $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ avec

- $\mathbf{y} = (y_{1,1}, \cdots, y_{n_1,1}, y_{1,2}, \cdots, y_{n_2,2}, \cdots, y_{1,J}, \cdots, y_{n_J,J})^\top$
- $\boldsymbol{\varepsilon} = (\varepsilon_{1,1}, \cdots, \varepsilon_{n_1,1}, \varepsilon_{1,2}, \cdots, y_{n_2,2}, \cdots, \varepsilon_{1,J}, \cdots, \varepsilon_{n_J,J})^\top$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_J)^\top$
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{A}]$ où $\mathbf{A}$ est une matrice $n \times J$

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_J} \end{pmatrix}, \quad \text{et } \mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{1}_{n_J} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_J} \end{pmatrix}$$

Ce modèle n'est pas identifiable ($\mathbf{X}$ n'est pas de plein rang)

On va alors imposer une contrainte linéaire, (au choix)

- imposer $\beta_0 = 0$
- imposer $\beta_{j_\star} = 0$ pour un certain $j_\star$ (référence)
- imposer $\beta_1 + \beta_2 + \cdots + \beta_J = 0$
- imposer $n_1\beta_1 + n_2\beta_2 + \cdots + n_J\beta_J = 0$

ANOVA

- imposer $\beta_0 = 0$

Alors $\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top \mathbf{y} = (\overline{y}_{\cdot 1}, \cdots, \overline{y}_{\cdot J}) \in \mathbb{R}^J$$

avec $\overline{y}_{\cdot j} = \dfrac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}$

Estimateur sans biais de $\boldsymbol{\beta}$, de variance $\sigma^2\text{diag}(n_1^{-1}, \cdots, n_J^{-1})$, minimale parmi les estimateurs linéaires sans biais de $\boldsymbol{\beta}$

$$\widehat{\sigma}^2 = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{i,j} - \overline{y}_{\cdot j})^2$$

ANOVA

- imposer $\beta_{j_\star}=0$
- $\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_{j_\star-1},\beta_{j_\star+1},\cdots,\beta_J)^\top\in\mathbb{R}^J$
- $\mathbf{X}=[\mathbf{1}_n,\mathbf{A}_{-j}]$ où où $\mathbf{A}_{-j}$ est une matrice $n\times(J-1)$

$$\mathbf{A}_{-j}=\begin{pmatrix}\mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{j_\star-1}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{n_{j_\star+1}} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_J}\end{pmatrix}$$

Alors $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$,

$$\widehat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}=(\widehat{\beta}_0,\widehat{\beta}_1,\cdots,\widehat{\beta}_J)$$

avec $\widehat{\beta}_{j_\star}=\overline{y}_{\cdot j_\star}=\dfrac{1}{n_{j_\star}}\displaystyle\sum_{i=1}^{n_{j_\star}}y_{i,j_\star}$ et $\widehat{\beta}_j=\overline{y}_{\cdot j}-\overline{y}_{\cdot j_\star}$

- imposer $n_1\beta_1 + n_2\beta_2 + \cdots + n_J\beta_J = 0$

Alors

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{J-1})^\top \in \mathbb{R}^J$
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{A}_\star]$ où $\mathbf{A}_\star$ est une matrice $n \times (J-1)$

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{n_1} - \frac{n_1}{n_J}\mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} - \frac{n_2}{n_J}\mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_{J-1}} - \frac{n_{J-1}}{n_J}\mathbf{1}_{n_{J-1}} \end{pmatrix}$$

Alors

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_{J-1})$$

avec $\widehat{\beta}_0 = \overline{y} = \dfrac{1}{n}\sum_{j=1}^{J} n_j \overline{y}_{\cdot j}$ et $\widehat{\beta}_j = \overline{y}_{\cdot j} - \overline{y}$

Modèle équivalent au précédent: même résidus, mêmes prévision, estimateur sans biais, etc

ANOVA

- imposer $\beta_1 + \beta_2 + \cdots + \beta_J = 0$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{J-1})^\top \in \mathbb{R}^J$

Alors

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_{J-1})$$

avec $\widehat{\beta}_0 = \widetilde{y} = \dfrac{1}{J}\displaystyle\sum_{j=1}^{J} \overline{y}_{\cdot j}$ et $\widehat{\beta}_j = \overline{y}_{\cdot j} - \widetilde{y}$

Modèle équivalent au précédent: même résidus, mêmes prévision, estimateur sans biais, etc.

Tous ces modèles sont équivalents !

ANOVA

Pour rappel, la formule de décomposition de la variance s'écrit

$$TSS = RSS + ESS$$

Soit

$$\|\mathbf{y} - \overline{y}\mathbf{1}\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \overline{y}\mathbf{1}\|^2$$

On suppose ε Gaussien. On veut tester
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0.$
Test de Fisher: sous H_0

$$F = \frac{ESS/(J-1)}{RSS/(n-J)} \sim \mathcal{F}(J-1, n-J)$$

ANOVA à deux facteurs

On suppose ici qu'il est possible d'appartenir aux groupes $j \in \{1, 2, \cdots, J\}$ et $k \in \{1, 2, \cdots, K\}$. Soit $n_{j,k}$ le nombre d'observations dans ce cas.

$$\overline{y}_{\cdot jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk},\ \overline{y}_{\cdot j\cdot} = \frac{1}{n_{j\cdot}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} y_{ijk},\ \overline{y}_{\cdot\cdot k} = \frac{1}{n_{\cdot k}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} y_{ijk}$$

Le modèle s'écrit ici

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

avec $\mathbb{E}[\varepsilon_{ijk}] = 0$, $\text{cov}(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0$ et $\text{Var}(\varepsilon_{i,j,k}) = 0$

ANOVA

Pour simplifier les notations, supposons $n_{jk} = n/(JK)$ (constant), $\forall j, k$.
On va disjoncter le couple de lois multinomiales $(X_1, X_2) \in \{1, 2, \cdots, J\} \times \{1, 2, \cdots, K\}$ en $(J + K + JK)$ indicatrices

$$(\underbrace{\mathbf{1}_1, \mathbf{1}_2, \cdots, \mathbf{1}_J}_{\mathbf{A}}, \underbrace{\mathbf{1}_1, \mathbf{1}_2, \cdots, \mathbf{1}_K}_{\mathbf{B}}, \underbrace{\mathbf{1}_{11}, \mathbf{1}_{12}, \cdots, \mathbf{1}_{JK}}_{\mathbf{C}})$$

de telle sorte que

$$y = (1, \mathbf{A}, \mathbf{B}, \mathbf{C}) \begin{pmatrix} \mu \\ \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + \varepsilon$$

Là encore, le modèle n'est pas identifiable car $\mathbf{X}$ est $n \times (1 + J + K + JK)$ mais de rang JK: il faut un ensemble de $1 + J + K$ contraintes (linéaires).

- imposer $\mu = 0$, $\alpha_j = 0 \ \forall j$ et $\beta_k = 0 \ \forall k$
- imposer $\alpha_{j_\star} = 0$, $\beta_{k_\star} = 0$ et $\gamma_{j_\star k} = \gamma_{j_\star k} = 0$ pour $j_\star$ et $k_\star$ (références)
- imposer $\alpha_1 + \alpha_2 + \cdots + \alpha_J = 0$, $\beta_1 + \beta_2 + \cdots + \beta_K = 0$ et

$$\gamma_{j1} + \gamma_{j2} + \cdots + \gamma_{jK} = 0, \ \forall j, \gamma_{1k} + \gamma_{2k} + \cdots + \gamma_{Jk} = 0, \ \forall k$$

- $\mu = 0$, $\alpha_j = 0\ \forall j$ et $\beta_k = 0\ \forall k$

alors $y_{ijk} = \gamma_{jk} + \varepsilon_{ijk}$. L'estimateur par moindres carrés est

$$\widehat{\gamma}_{jk} = \overline{y}_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}$$

Et

$$\widehat{\sigma}^2 = \frac{1}{n - JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{jk})^2$$

- $\boldsymbol{\alpha}^\top \mathbf{1} = 0$, $\boldsymbol{\beta}^\top \mathbf{1} = 0$, $\boldsymbol{\gamma}_{j\cdot}^\top \mathbf{1} = 0$, $\boldsymbol{\gamma}_{\cdot k}^\top \mathbf{1} = 0$

Alors

$$\widehat{\mu} = \overline{y},\ \widehat{\alpha}_j = \overline{y}_{j\cdot} - \overline{y},\ \widehat{\beta}_k = \overline{y}_{\cdot,k} - \overline{y}$$

$$\widehat{\gamma}_{jk} = \overline{y}_{jk} - \overline{y}_{j\cdot} - \overline{y}_{\cdot k} + \overline{y}$$

La décomposition de la variance s'écrit ici

$$TSS = RSS + ESS_A + ESS_B + ESS_C$$

où

$$ESS_A = \frac{n}{J}\sum_{j=1}^{J}\widehat{\alpha}_j^2,\; ESS_B = \frac{n}{K}\sum_{k=K}^{J}\widehat{\beta}_k^2 \text{ and } ESS_C = \frac{n}{JK}\sum_{j=1}^{J}\sum_{k=K}^{J}\widehat{\gamma}_{jk}^2$$

Le test multiple $H_0 : \gamma_{jk} = 0,\ \forall j, k$ est appelé test de l'interaction, et

$$F = \frac{ESS_C/((J-1)(K-1))}{RSS/(n-JK)} \sim \mathcal{F}((J-1)(K-1), n-JK)$$

sous H_0.

Le test multiple $H_0 : \alpha_j = 0, \forall j$ est appelé test de l'effet du facteur A, et

$$F = \frac{ESS_A/(J-1)}{RSS/(n-JK)} \sim \mathcal{F}(J-1, n-JK)$$

sous H_0.

Le test multiple $H_0 : \beta_j = 0, \forall j$ est appelé test de l'effet du facteur B, et

$$F = \frac{ESS_B/(K-1)}{RSS/(n-JK)} \sim \mathcal{F}(K-1, n-JK)$$

sous H_0.