Université de Strasbourg
M1 - Magistère
Statistique

Ségolen Geffray
geffray@math.unistra.fr
Année 2019/2020

Sujet 11

• **Brefs rappels sur la régression linéaire**:
Lorsque $\left(X^{(1)}, ..., X^{(p)}\right)$ est un vecteur de covariables quantitatives, le modèle de régression linéaire correspondant à l'observation de $\left(Y, X^{(1)}, ..., X^{(p)}\right)$ s'écrit:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k X^{(k)} + \varepsilon \tag{1}$$

où $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) < \infty$ et $\varepsilon \perp\!\!\!\perp \left(X^{(1)}, ..., X^{(p)}\right)$.
Le modèle de régression linéaire est dit gaussien lorsqu'on rajoute l'hypothèse $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
Le modèle de régression linéaire est dit homoscédastique lorsqu'on rajoute l'hypothèse que la variance de l'erreur résiduelle est constante d'un individu à l'autre.

• **Cas des covariables qualitatives (appelées aussi facteurs)**:
L'écriture (1) est générique. Il faut prendre garde au fait que l'on ne rentre en réalité sous cette forme **que** les variables quantitatives. Supposons ici que $X^{(1)}$ est ici une variable qualitative à J modalités que l'on peut encoder comme suit:

$$X^{(1)} = \begin{cases} 1 & \text{si } X^{(1)} \text{ prend la modalité n}^{\text{o}}\, 1, \\ \vdots & \vdots \\ J & \text{si } X^{(1)} \text{ prend la modalité n}^{\text{o}}\, \text{J}. \end{cases}$$

Si une modalité se dégage naturellement comme variable de référence, on prend soin de l'encoder avec 1 lorsqu'on utilise le logiciel `R`, qui prend la 1ère modalité comme modalité de référence. Par exemple, considérons la covariable qui encode le statut d'un individu relatif au fait d'avoir des antécédents familiaux d'hypertension artérielle, on l'encode comme suit:

$$X^{(1)} = \begin{cases} 1 & \text{si l'individu n'a pas d'antécédents familiaux d'hypertension artérielle} \\ & \text{(modalité de référence),} \\ 2 & \text{dans le cas contraire.} \end{cases}$$

Lorsque $X^{(1)}$ est ici une variable qualitative à J modalités, l'équation (1) est modifiée comme suit. Ce n'est pas $X^{(1)}$ qui est incluse avec un coefficient multiplicateur dans l'équation (1) mais des indicatrices (*dummy variables*) avec autant de coefficients multiplicateurs. Inclure J indicatrices sans précaution particulière surparamètrerait le modèle et entrainerait un problème d'identifiabilité au sens où l'on n'aurait plus l'injectivité de $\beta \to \mathbb{E}_\beta[\mathbb{Y}|X] = \mathbb{X}.\beta$. Imposer la nullité d'un coefficient de sorte que l'on inclut seulement $(J-1)$ indicatrices est une façon de rétablir l'identifiabilité du modèle. La modalité non-incluse dans l'équation sert alors de modalité de référence dans l'interprétation des coefficients. Le logiciel `R` prend par défaut la 1ère modalité comme modalité de référence. Cela donne la nouvelle équation:

$$Y = \beta_0 + \beta_{1,2} I(X^{(1)} = 2) + ... + \beta_{1,J} I(X^{(1)} = J) + \sum_{k=2}^{p} \beta_k X^{(k)} + \varepsilon.$$

Pour $j = 1, ..., J$, le coefficient $\beta_{1,j}$ s'interprète alors comme l'effet sur la valeur moyenne de Y dû au niveau j de la variable $X^{(1)}$ par rapport à la valeur moyenne pour Y avec la modalité 1 de $X^{(k_0)}$, toutes les autres covariables étant fixées quelconques le temps de la comparaison. On parle d'effet différentiel.
Noter qu'il n'y a pas unicité de la manière de rétablir l'identifiabilité. Selon les logiciels, d'autres choix par défaut sont adoptés.

• **ANOVA(1) = analyse de la variance à un facteur présentant J modalités**:
Lorsque le modèle de régression inclut une seule covariable qualitative, on parle d'analyse de la variance à un facteur à J modalités. Supposons que l'on dispose d'un échantillon (Y_i, X_i) pour $i = 1, ..., n$ où X_i représente un facteur à J modalités. L'équation du modèle de régression linéaire ou d'ANOVA(1) correspondant est alors:

$$Y_i = \beta_0 + \beta_2 I(X_i^{(1)} = 2) + ... + \beta_J I(X_i^{(1)} = J) + \varepsilon_i, \quad i = 1, ..., n.$$

Comme son nom ne l'indique pas, ce modèle permet de comparer les moyennes des J différents sous-groupes d'une population identifiés par les J modalités du facteur en question. Au lieu de numéroter les individus de 1 à n, on peut les renuméroter de manière équivalente pour indiquer la modalité exprimée par chacun. Notons $n_j \in \mathbb{N}^*$ le nombre d'individus exprimant la modalité j du facteur en question de sorte que l'on a forcément $\sum_{j=1}^{J} n_j = n$. Notons $Y_{i,j}$ l'observation réalisée sur le $i^{\text{ème}}$ individu présentant la $j^{\text{ème}}$ modalité du facteur $X^{(1)}$. Le modèle d'analyse de la variance à un facteur est souvent ré-écrit de la façon suivante:

$$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}, \quad \begin{cases} j = 1, ..., J, \\ i = 1, ..., n_j, \end{cases} \qquad \text{avec } \alpha_1 = 0\,.$$

L'écriture matricielle de la forme $Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$ pour $j = 1, ..., J$ et $i = 1, ..., n_j$ avec $\alpha_1 = 0$ est

$$\begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{1,2} \\ \vdots \\ Y_{n_2,2} \\ \hline \vdots \\ \hline Y_{1,J} \\ \vdots \\ Y_{n_J,J} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & \dots & 0 \\ \hline 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots \\ \hline 1 & \dots & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \dots & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \hline \alpha_2 \\ \vdots \\ \alpha_J \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{n_1,1} \\ \hline \varepsilon_{1,2} \\ \vdots \\ \varepsilon_{n_2,2} \\ \hline \vdots \\ \hline \varepsilon_{1,J} \\ \vdots \\ \varepsilon_{n_J,J} \end{pmatrix}$$

Lorsque $n_j = n_1$ pour tout $j = 2, ..., J$, on dit que le *design* est équilibré. Dans le cas contraire, on dit que le *design* est déséquilibré.

• **Exercice**:
Notons β le vecteur des paramètres de régression à estimer dans le cadre d'un modèle d'analyse de la variance à un facteur et $\widehat{\beta}$ son estimateur.

1. Proposer des simulations de Monte-Carlo permettant d'évaluer le biais empirique de $\widehat{\beta}$ et son écart-type estimé.
   Le biais et l'écart-type estimés sont-ils sensibles à l'hypothèse de normalité? à l'hypothèse d'homoscédasticité? à la taille de l'échantillon? au fait que le *design* soit équilibré ou non? au nombre de coefficients du modèle?

2. Proposer des simulations de Monte-Carlo permettant d'illustrer la convergence asymptotique de $\widehat{\beta}$ vers β. La convergence est-elle sensible à l'hypothèse de normalité? à l'hypothèse d'homoscédasticité? au fait que le *design* soit équilibré ou non? au nombre de coefficients du modèle?

3. Le test de Fisher permet de tester l'effet du facteur sur le niveau moyen de la réponse. Compte tenu de la contrainte d'identifiabilité $\beta_1 = 0$, l'hypothèse nulle est H_0: $\beta_2 = ... = \beta_J = 0$ tandis que l'hypothèse alternative est H_1: $\exists j \in \{2, ..., J\}$ tel que $\beta_j \neq 0$. Ce test est implémenté dans le logiciel `R` au moyen de la fonction `anova`.
   Présenter ce test. Proposer des simulations de Monte-Carlo permettant d'évaluer l'erreur empirique de $1^{\text{ère}}$ espèce et la puissance empirique de ce test. Les résultats obtenus sont-ils sensibles à l'hypothèse de normalité? à la taille de l'échantillon? au fait que le *design* soit équilibré ou non? au nombre de coefficients du modèle?