Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2016/2017

Sujet 1: Etude de l'estimateur des coefficients de régression en cas d'omission d'une variable explicative influente ou en cas de rajout d'une variable explicative non influente

---

• **Quelques rappels sur la régression linéaire**:
De manière générale, le but de tout modèle de régression est de spécifier une relation (de nature stochastique) entre une variable Y appelée variable à expliquer et un certain nombre de variables explicatives $X^{(1)}, ..., X^{(p)}$. Dans toute la suite, on supposera que les variables explicatives sont de loi continue.
Notons $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$, les observations correspondant à n individus que l'on suppose distribuées comme $(Y, X^{(1)}, ..., X^{(p)})$. Le modèle de régression linéaire gaussien standard s'écrit sous la forme générique suivante:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)} + \varepsilon_i, \qquad \text{où} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Le modèle de régression linéaire standard repose ainsi sur les hypothèses suivantes:

($\mathbf{H_1}$) **linéarité**: l'espérance conditionnelle de la variable réponse vaut

$$\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)}.$$

Attention, bien que l'espérance conditionnelle de la variable réponse soit une combinaison affine des covariables, la linéarité s'apprécie relativement à la linéarité de $\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}]$ en les paramètres.
On supposera toujours que le modèle est identifiable, à savoir

$$\beta := (\beta_0, ..., \beta_p)^t = (\beta'_0, ..., \beta'_p)^t := \beta' \quad \Longrightarrow \quad \mathbb{E}_\beta[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \mathbb{E}_{\beta'}[Y_i | X_i^{(1)}, ..., X_i^{(p)}].$$

($\mathbf{H_2}$) **centrage des erreurs**: $\mathbb{E}[\varepsilon_i] = 0$ pour $i = 1, ..., n$.

($\mathbf{H_3}$) **exogénéité = non-endogénéité**: $(X_i^{(1)}, ..., X_i^{(p)})$ et ε_i sont indépendants pour $i = 1, ..., n$.

($\mathbf{H_4}$) **non-colinéarité des covariables**: les covariables sont non-corrélées entre elles, ce qui se traduit en pratique par le fait que les vecteurs $(X_1^{(k)}, ..., X_n^{(k)})$ sont non-colinéaires pour $k = 1, ..., p$. On écartera également le cas trivial où l'un de ces vecteurs serait colinéaire au vecteur de longueur n dont toutes les composantes sont égales à 1.

($\mathbf{H_5}$) **non-corrélation**: les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont non-corrélés pour $i = 1, ..., n$.

($\mathbf{H_6}$) **homoscédasticité**: les lois conditionnelles de Y_i sachant $X_i^{(1)}, ..., X_i^{(p)}$ ont même variance donc on a $\sigma_i^2 = \sigma^2$ pour $i = 1, ..., n$.

Le modèle de régression linéaire standard est gaussien lorsqu'on effectue l'hypothèse supplémentaire suivante:

($\mathbf{H_7}$) **normalité**: conditionnellement à $X_i^{(1)}, ..., X_i^{(p)}$, la variable de réponse Y_i suit la loi normale.

Notons I_n=matrice identité de taille $n \times n$ et avec

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(p)} \end{pmatrix} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Le modèle de régression linéaire homoscédastique se réécrit matriciellement sous la forme:

$$\mathbb{Y} = \mathbb{X}.\beta + \boldsymbol{\varepsilon}, \qquad \text{où} \quad \boldsymbol{\varepsilon} \sim (0_n, \sigma^2 I_n) \text{ et } \boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbb{X}.$$

Le modèle de régression linéaire gaussien homoscédastique s'obtient sous forme matricielle lorsque l'on ajoute l'hypothèse $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$.

On dit que le modèle est bien spécifié lorsqu'il correspond effectivement au mécanisme ayant servi à généré les données. Dans le cas contraire, on dit qu'il est mal spécifié.

• **Estimation de l'effet des variables explicatives**:
Dans le cadre du modèle linéaire gaussien standard présenté ci-dessus, on se demande si les variables explicatives introduites dans le modèle ont un effet ou non sur la variable à expliquer. Cet effet est quantifié par le coefficient de régression correspondant.
L'estimateur des moindres carrés ordinaires de β est $\widehat{\beta} = (\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t.\mathbb{Y}$. Sous les hypothèses $(H_1) - (H_4)$, cet estimateur est sans biais

$$\mathbb{E}[\widehat{\beta}|\mathbb{X}] = \beta\,.$$

Sous les hypothèses $(H_1) - (H_6)$, la variance de $\widehat{\beta}$ est

$$\mathrm{Var}(\widehat{\beta}|\mathbb{X}) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}\,.$$

Notons $\widehat{\sigma}^2$ l'estimateur de la variance de σ^2 défini par:

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^t.\widehat{\boldsymbol{\varepsilon}}}{n-(p+1)}$$

en notant

$$\widehat{\varepsilon}_i = Y_i - \mathbb{X}_i.\widehat{\beta}$$

et

$$\widehat{\boldsymbol{\varepsilon}} = \mathbb{Y} - \mathbb{X}.\widehat{\beta}\,.$$

Sous les hypothèses (H_1)-(H_6), l'estimateur $\widehat{\sigma}^2$ est sans biais pour σ^2.

Introduisons une hypothèse supplémentaire:
($\mathbf{H_8}$): $\dfrac{\mathbb{X}^t.\mathbb{X}}{n} \xrightarrow{\mathbb{P}} Q$ lorsque $n \to \infty$ où Q est une matrice symétrique définie positive.

Sous les hypothèses (H_1)-(H_6) et (H_8), la convergence $\widehat{\beta} \xrightarrow{\mathbb{P}} \beta$ a lieu lorsque $n \to \infty$.

Sous les hypothèses $(H_1) - (H_7)$, la loi de l'estimateur $\widehat{\beta}$ de β est

$$\widehat{\beta}_{EMV} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}),$$

et, pour $j = 0, 1, ..., p$, on a

$$\frac{\left(\widehat{\beta}^{(j)} - \beta^{(j)}\right)}{\sqrt{\widehat{\sigma}^2[(\mathbb{X}^t.\mathbb{X})^{-1}]_{j,j}}} \sim T(n - (p+1)).$$

• **Effet de l'omission d'une variable influente sur le biais de $\widehat{\beta}$:**
Supposons que le "vrai" modèle ayant effectivement généré les données est

$$\begin{aligned}\mathbb{Y} &= \mathbb{X}.\beta + \varepsilon \\ &= \mathbb{X}_{(1)}.\beta_{(1)} + \mathbb{X}_{(2)}.\beta_{(2)} + \varepsilon, \quad \varepsilon \sim (0_n, \sigma^2 I_n)\end{aligned}$$

où l'on a partitionné la matrice $\mathbb{X}$ en deux sous-matrices $\mathbb{X}_{(1)}$ et $\mathbb{X}_{(2)}$ de taille respective $n \times q_1$ et $n \times q_2$ avec $q_1 + q_2 = p + 1$ de sorte que $\mathbb{X} = [\mathbb{X}_{(1)}, \mathbb{X}_{(2)}]$. On suppose que $\mathbb{X}_{(2)}$ a une influence sur $\mathbb{Y}$ (ie que $\beta_{(2)} \neq 0$) mais que l'on régresse $\mathbb{Y}$ sur $\mathbb{X}_{(1)}$ sans inclure (à tort) $\mathbb{X}_{(2)}$. L'estimateur obtenu pour $\beta_{(1)}$ est alors

$$\widehat{\beta}_{(1)} = (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{Y}.$$

Déterminons son espérance:

$$\begin{aligned}\mathbb{E}[\widehat{\beta}_{(1)}|\mathbb{X}] &= \mathbb{E}[(\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{Y}|\mathbb{X}] \\ &= (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{E}[\mathbb{Y}|\mathbb{X}] \\ &= (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{E}[\mathbb{X}_{(1)}.\beta_{(1)} + \mathbb{X}_{(2)}.\beta_{(2)} + \varepsilon|\mathbb{X}] \\ &= (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)}.\beta_{(1)} + (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{X}_{(2)}.\beta_{(2)} + (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{E}[\varepsilon|\mathbb{X}] \\ &= \beta_{(1)} + (\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}\mathbb{X}_{(1)}^t.\mathbb{X}_{(2)}.\beta_{(2)}\end{aligned}$$

de sorte que $\widehat{\beta}_{(1)}$ est un estimateur biaisé de $\beta_{(1)}$ sauf si $\mathbb{X}_{(1)}^t.\mathbb{X}_{(2)} = 0_{q_1 \times q_2}$.

• **Effet de l'inclusion d'une variable non-influente sur $\widehat{\beta}$:**
Partitionnons la matrice $\mathbb{X}$ en deux sous-matrices $\mathbb{X}_{(1)}$ et $\mathbb{X}_{(2)}$ de taille respective $n \times q_1$ et $n \times q_2$ avec $q_1 + q_2 = p + 1$ de sorte que $\mathbb{X} = [\mathbb{X}_{(1)}, \mathbb{X}_{(2)}]$. Partitionnons le vecteur β en deux sous-vecteurs $\beta_{(1)}$ et $\beta_{(2)}$ de longueur respective q_1 et q_2 avec $q_1 + q_2 = p + 1$ de sorte que $\beta = \begin{pmatrix}\beta_{(1)} \\ \beta_{(2)}\end{pmatrix}$.
Supposons que $\mathbb{X}_{(2)}$ n'a en réalité aucune influence sur $\mathbb{Y}$. Ainsi, le "vrai" modèle ayant effectivement généré les données est

$$\mathbb{Y} = \mathbb{X}_{(1)}.\beta_{(1)} + \varepsilon, \quad \varepsilon \sim (0_n, \sigma^2 I_n).$$

Supposons que l'on régresse $\mathbb{Y}$ sur $\mathbb{X}$ ie en incluant (à tort) $\mathbb{X}_{(2)}$.

$$\begin{aligned}\mathbb{Y} &= \mathbb{X}.\beta + \varepsilon \\ &= \mathbb{X}_{(1)}.\beta_{(1)} + \mathbb{X}_{(2)}.\beta_{(2)} + \varepsilon, \quad \varepsilon \sim (0_n, \sigma^2 I_n)\end{aligned}$$

Dans ce cas, d'après ce qui précède, on sait que $\widehat{\beta}\begin{pmatrix}\widehat{\beta}_{(1)}\\ \widehat{\beta}_{(2)}\end{pmatrix}$ est sans biais pour β. Evidemment, en pratique, on ne trouvera pas exactement $\widehat{\beta}_{(2)}=0$ mais une valeur souvent assez proche de 0. Notons $\check{\beta}_{(1)}$ l'estimateur que l'on aurait obtenu si l'on avait régressé le "vrai" modèle sur les données ie $\check{\beta}_{(1)}=(\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)})^{-1}.\mathbb{X}_{(1)}^t.\mathbb{Y}$.
Comparons les variances des deux estimateurs sans biais $\widehat{\beta}_{(1)}$ et $\check{\beta}_{(1)}$ de $\beta_{(1)}$.

La variance de $\widehat{\beta}_{(1)}$ est le 1er bloc de la matrice suivante:

$$\text{Var}\left(\widehat{\beta}_{(1)}\right)=\sigma^2\begin{pmatrix}\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)} & \mathbb{X}_{(1)}^t.\mathbb{X}_{(2)}\\ \mathbb{X}_{(2)}^t.\mathbb{X}_{(1)} & \mathbb{X}_{(2)}^t.\mathbb{X}_{(2)}\end{pmatrix}^{-1}.$$

Rappel d'algèbre linéaire: L'inverse d'une matrice peut être calculé par blocs, en utilisant la formule analytique suivante:

$$\begin{pmatrix}A & B\\ C & D\end{pmatrix}^{-1}=\begin{pmatrix}\left(A-BD^{-1}C\right)^{-1} & -\left(A-BD^{-1}C\right)^{-1}BD^{-1}\\ -D^{-1}C\left(A-BD^{-1}C\right)^{-1} & D^{-1}+D^{-1}C\left(A-BD^{-1}C\right)^{-1}BD^{-1}\end{pmatrix},$$

pourvu que D et $A-BD^{-1}C$ soient inversibles.

D'après le rappel, la variance de $\widehat{\beta}_{(1)}$ est donc:

$$\text{Var}\left(\widehat{\beta}_{(1)}\right)=\sigma^2\left(\mathbb{X}_{(1)}^t.\left(I-\mathbb{X}_{(2)}.(\mathbb{X}_{(2)}^t.\mathbb{X}_{(2)})^{-1}.\mathbb{X}_{(2)}^t\right)\mathbb{X}_{(1)}\right)^{-1}.$$

La variance de $\check{\beta}_{(1)}$ est

$$\text{Var}\left(\check{\beta}_{(1)}\right)=\sigma^2\left(\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)}\right)^{-1}.$$

La différence entre les inverses de deux matrices de variance est

$$\begin{aligned}\text{Var}\left(\widehat{\beta}_{(1)}\right)^{-1}-\text{Var}\left(\check{\beta}_{(1)}\right)^{-1}&=\sigma^2\left(\mathbb{X}_{(1)}^t.\left(I-\mathbb{X}_{(2)}.(\mathbb{X}_{(2)}^t.\mathbb{X}_{(2)})^{-1}.\mathbb{X}_{(2)}^t\right)\mathbb{X}_{(1)}-\mathbb{X}_{(1)}^t.\mathbb{X}_{(1)}\right)\\&=-\sigma^2\,\mathbb{X}_{(1)}^t.\mathbb{X}_{(2)}.(\mathbb{X}_{(2)}^t.\mathbb{X}_{(2)})^{-1}.\mathbb{X}_{(2)}^t\mathbb{X}_{(1)}\end{aligned}$$

qui est une matrice semi-définie négative. Ainsi la matrice $\text{Var}\left(\widehat{\beta}_{(1)}\right)-\text{Var}\left(\check{\beta}_{(1)}\right)$ est semi-définie positive.

• Implémentation au moyen du logiciel `R`:
Le logiciel `R` permet d'ajuster simplement un modèle de régression à des données. Pour ajuster un modèle linéaire sur un échantillon $(Y_i,X_i^{(1)},...,X_i^{(p)})_{i=1,...,n}$, on stockera les n valeurs des Y_i dans un vecteur `y`, puis, pour $k=1,\ldots,p$, on stockera les n valeurs des $X_i^{(k)}$ dans un vecteur `xk`. Le modèle linéaire est alors ajusté au moyen de l'instruction suivante (où $p=3$)

```
mylm<- lm(y~x1+x2+x3)
```

et le résultat est stocké dans l'objet `mylm`. L'instruction suivante permet d'accéder à l'ensemble des quantités calculées:

```
summary(mylm)
```

Sont notamment calculées, les estimations des β_j pour $j = 0, \ldots, p$ et les estimations de leurs écarts-types. L'objet `summary(mylm)` est une liste dont on peut récupérer chacune des composantes qui nous intéresse. Pour cela, l'instruction `str(summary(mylm))` permet de voir le nom et la structure des différentes composantes de cette liste.

Exercice 1.
Dans le cadre du modèle de régression linéaire gaussien standard, illustrer de manière empirique à partir de données simulées le comportement de l'estimateur des coefficients de régression, en termes de biais, variance, écart quadratique moyen, consistance et distribution de l'estimateur $\widehat{\beta}$, tour à tour,

1. lorsqu'est/sont incluse(s) lors de l'ajustement une ou plusieurs variable(s) explicative(s) non-influente(s) en réalité,

2. lorsqu'est/sont omise(s) lors de l'ajustement une ou plusieurs variable(s) explicative(s) influente(s) en réalité.

Vous veillerez à faire varier également:

- la taille de l'échantillon n simulé,

- la variance de l'erreur résiduelle simulée.