Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 3: Etude du comportement de l'estimateur des coefficients de régression en fonction de la distribution de l'erreur

---

• **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 à n observations indépendantes 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 à n observations indépendantes** est gaussien lorsqu'on effectue l'hypothèse 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

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

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

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

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

• **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:

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

en notant

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

et

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

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

• **Comportement asymptotique**:
Soit l'hypothèse:
($\mathbf{H_8}$): $\dfrac{\mathbb{X}^t.\mathbb{X}}{n} \xrightarrow{\mathbb{P}} Q$ 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 $n \to \infty$.

Soit l'hypothèse:
$(\mathbf{H_9})$: les erreurs ε_i sont mutuellement indépendantes pour $i = 1, ..., n$, ou de manière équivalente, les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont mutuellement indépendants pour $i = 1, ..., n$.

Sous les hypothèses (H_1)-(H_6), (H_8) et (H_9), la convergence $\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$ a lieu lorsque $n \to \infty$.

Introduisons les conditions de Grenander pour des données à bon comportement:
(G_1): pour chaque colonne $\mathbf{X}^{(k)}$ de $\mathbb{X}$, $(\mathbf{X}^{(k)})^t.\mathbf{X}^{(k)} \xrightarrow{\mathbb{P}} +\infty$ lorsque $n \to \infty$.
(G_2): pour $i = 1, ..., n$, $\dfrac{\left(X_i^{(k)}\right)^2}{(\mathbf{X}^{(k)})^t.\mathbf{X}^{(k)}} \xrightarrow{\mathbb{P}} 0$ lorsque $n \to \infty$.
(G_3): $R_n \xrightarrow{\mathbb{P}} C$ lorsque $n \to \infty$, où C est une matrice définie positive et où R_n est la matrice de corrélation empirique des colonnes de $\mathbb{X}$ à l'exception de la colonne de 1 correspondant à l'inclusion du terme constant.
Les conditions de Grenander sont des conditions assez faibles, invérifiables en pratique, mais susceptibles d'être satisfaites par de nombreux jeux de données.

Sous les hypothèses (G_1)-(G_3), (H_1)-(H_6), (H_8) et (H_9), lorsque $n \to \infty$,

$$\left(\sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}\right)^{-1/2}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_{p+1}(0_{p+1}, I_{p+1}),$$

et

$$\left(\widehat{\sigma}^2(\mathbb{X}^t.\mathbb{X})^{-1}\right)^{-1/2}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}_{p+1}(0_{p+1}, I_{p+1}).$$

• **Comportement à distance finie dans le cas gaussien**:
Sous les hypothèses $(H_1) - (H_7)$,

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

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

$$\left(\widehat{\sigma}^2(\mathbb{X}^t.\mathbb{X})^{-1}\right)^{-1/2}\left(\widehat{\beta}^{(j)} - \beta^{(j)}\right) \sim T(n - (p+1)).$$

• **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.
Etudier de manière empirique à partir de données simulées la distribution de l'estimateur des coefficients de régression en fonction de

- la taille de l'échantillon n simulé,
- la distribution de l'erreur résiduelle simulée.