Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 10: Etude de l'estimateur des coefficients de régression par moindres carrés ordinaires et par régression sur composantes principales lorsque p devient grand devant n

---

• **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

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

• Régression sur Composantes Principales (dite PCR = Principal Components Regression):

Une régression linéaire standard sur p variables possiblement corrélées entre elles ou pour p grand devant n peut être remplacée par une régression utilisant les p' premières composantes principales, lesquelles sont décorrélées, pour un $p' \leq p$.
Considérons le modèle

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

où $\mathbb{Y}$ est le vecteur centré et réduit des réponses (ce qui entraîne que $\sigma^2 = 1$), et où $\mathbb{X}$ est la matrice des exogènes **centrées et réduites** donc sans constante, de taille $n \times p$.

En 1ère étape, effectuons une Analyse en Composantes Principales sur la matrice $\mathbb{X}$ qui a été préalablement centrée et réduite. Pour résumer, l'ACP revient à remplacer les variables $X^{(1)}, ..., X^{(p)}$ (qui sont possiblement corrélées entre elles) par de nouvelles variables $C^{(1)}, ..., C^{(p)}$ appelées composantes principales, qui sont des combinaisons linéaires des colonnes de $\mathbb{X}$. Par construction, ces composantes principales sont non corrélées entre elles.On procède de la façon suivante. La matrice $\mathbb{X}^t.\mathbb{X}$ est symétrique semi-définie positive donc est diagonalisable dans $\mathbb{R}^+$ au moyen d'une matrice de passage P orthogonale. Soient $\lambda_1, ..., \lambda_p$ les valeurs propres de $\mathbb{X}^t.\mathbb{X}$ que l'on suppose rangées par ordre décroissant: $\lambda_1 \geq ... \geq \lambda_p \geq 0$. La matrice P est la matrice orthogonale des vecteurs propres de $\mathbb{X}^t.\mathbb{X}$ qui sont deux à deux orthogonaux. Si l'on norme à 1 ces vecteurs propres, la matrice P est une matrice de rotation.Ces vecteurs propres sont appelés les facteurs principaux. Construisons la matrice des composantes principales $\widetilde{\mathbb{X}}$ en posant $\widetilde{\mathbb{X}} = \mathbb{X}.P$. Posons ensuite $\widetilde{\beta} = P^t.\beta$.
Le modèle $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ se réécrit donc

$$\mathbb{Y} = \mathbb{X}.(P.P^t).\beta + \varepsilon = (\mathbb{X}.P).(P^t.\beta) + \varepsilon = \widetilde{\mathbb{X}}.\widetilde{\beta} + \varepsilon.$$

On peut montrer que, si l'on se contente des p' premières composantes, on obtient la meilleure approximation au sens des moindres carrés de $\mathbb{X}$ par p' composantes.

En 2ème étape, effectuons donc l'approximation

$$\widetilde{\mathbb{X}} \approx [C^{(1)}, ..., C^{(p')}, 0_n, ..., 0_n]$$

ce qui revient à imposer une contrainte linéaire de taille $(p - p')$ à savoir $(0_{p'}, 1_{p-p'}) \begin{pmatrix} \widetilde{\beta}^{(1)} \\ \widetilde{\beta}^{(2)} \end{pmatrix} = 0_p$

en scindant $\widetilde{\beta}$ en deux sous-vecteurs de taille respective p' et $(p - p')$ i.e. $\widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}^{(1)} \\ \widetilde{\beta}^{(2)} \end{pmatrix}$. On

travaille alors dans le modèle approximé de régression approximé:

$$\mathbb{Y} = [C^{(1)}, ..., C^{(p')}].\widetilde{\beta}^{(1)} + \mathfrak{E}.$$

En 3ème étape, déterminons l'estimateur des moindres carrés ordinaires de $\widetilde{\beta}^{(1)}$:

$$\begin{aligned}
\widehat{\widetilde{\beta}^{(1)}} &= ([C^{(1)}, ..., C^{(p')}]^t.[C^{(1)}, ..., C^{(p')}])^{-1}.[C^{(1)}, ..., C^{(p')}]^t.\mathbb{Y} \\
&= \begin{pmatrix} \frac{1}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda_{p'}} \end{pmatrix} . \begin{pmatrix} (C^{(1)})^t \\ \vdots \\ (C^{(p')})^t \end{pmatrix} .\mathbb{Y} \\
&= \begin{pmatrix} \frac{(C^{(1)})^t.\mathbb{Y}}{\lambda_1} \\ \vdots \\ \frac{(C^{(p')})^t.\mathbb{Y}}{\lambda'_p} \end{pmatrix} .
\end{aligned}$$

On peut en déduire une prédiction de $\mathbb{Y}$ par la formule:

$$\widehat{\mathbb{Y}} = [C^{(1)}, ..., C^{(p')}].\widehat{\widetilde{\beta}^{(1)}}.$$

En 4ème étape, on peut revenir aux variables explicatives de départ en exprimant le fait que les composantes principales sont des combinaisons linéaires des colonnes de $\mathbb{X}$, lesquelles sont notées $\mathbf{X}^{(1)}$, ..., $\mathbf{X}^{(p)}$:

$$C_j = \sum_{k=1}^{p} P_{kj}\mathbf{X}^{(k)}.$$

Ainsi, on peut définir l'estimateur PCR de β en utilisant l'identification suivante:

$$\begin{aligned}
[C^{(1)}, ..., C^{(p')}].\widehat{\widetilde{\beta}^{(1)}} &= \sum_{j=1}^{p'} \widehat{\widetilde{\beta}_j^{(1)}} C_j \\
&= \sum_{j=1}^{p'} \widehat{\widetilde{\beta}_j^{(1)}} \sum_{k=1}^{p} P_{kj}\mathbf{X}^{(k)} \\
&= \sum_{k=1}^{p} \underbrace{\left(\sum_{j=1}^{p'} P_{kj} \widehat{\widetilde{\beta}_j^{(1)}} \right)}_{=\widehat{\beta_{k,\mathrm{PCR}}}} \mathbf{X}^{(k)} .
\end{aligned}$$

Comment choisir le nombre optimal de composantes principales à conserver?
Une premier critère consiste à conserver les composantes dont la valeur propre associée est supérieure ou égale à 1. Un deuxième critère consiste à conserver les composantes telles que la proportion de variance empirique des p' premières composantes parmi les p

$$\psi_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

atteigne une valeur cible, comme par exemple 75% ou 90%.

• 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.
L'instruction

```
prcomp(y~x1+x2+x3,scale=TRUE)
```

permet d'effectuer une régression sur composantes principales de la matrice $\mathbb{X}$ centrée et réduite.

Exercice 1.
Dans le cadre du modèle de régression linéaire gaussien standard, lorsque p devient grand devant n, illustrer de manière empirique à partir de données simulées le comportement de l'estimateur des moindres carrés ordinaires et de l'estimateur PCR des coefficients de régression, en termes de biais, variance, écart quadratique moyen, consistance et distribution de l'estimateur $\widehat{\beta}$. Vous veillerez à faire varier également:

- le critère de choix du nombre de composantes principales à retenir dans le modèle,
- l'ampleur de p devant n,
- la taille de l'échantillon n simulé,
- la variance de l'erreur résiduelle simulée.