Université de Strasbourg
M1 - Magistère
Statistique - études de cas

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

Sujet 2: Indicateurs d'ajustement et de qualité prédictive d'un modèle de régression linéaire 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**:
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

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

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 homoscédastique gaussien 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:

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

en notant les résidus (bruts) pour $i = 1, ..., n$, par

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

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.

Introduisons une hypothèse supplémentaire:
($\mathbf{H_8}$): $\frac{\mathbb{X}^t.\mathbb{X}}{n} \xrightarrow{\mathbb{P}} Q$ 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 $n \to \infty$.

• **Résidus standardisés**:
Considérons les résidus (bruts) que l'on a défini pour $i = 1, ..., n$ par $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Lorsque la valeur $|\widehat{\varepsilon}_i|$ est "anormalement" élevée, cela indique que le $i^{\text{ème}}$ réponse a été mal reconstituée par le modèle. Afin de pouvoir trancher si une observation est "anormalement" élevée, il faut être en mesure de pouvoir comparer des choses comparables. Or, bien que l'on ait travaillé sous l'hypothèse d'homoscédasticité qui stipule que $\text{Var}(\varepsilon_i) = \sigma^2$ pour $i = 1, ..., n$, il n'en va pas de même pour les résidus bruts. On a en effet établi que $\text{Var}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_i)$ pour $i = 1, ..., n$. Pour rendre les amplitudes comparables entre les différents résidus, ie afin d'obtenir des résidus de variances égales, on normalise chacun des résidus (bruts) par son écart-type estimé pour ne plus s'intéresser qu'aux résidus normalisés. Les résidus standardisés sont des résidus normalisés définis de la façon suivante:

$$\widehat{\varepsilon}_i^{\text{Stand}} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_i)}}$$

où $\widehat{\sigma}^2$ est l'estimateur sans biais de σ^2 défini en (2).
Pour n assez grand, la distribution des résidus standardisés est la loi $\mathcal{N}(0, 1)$. Pour n assez grand, lorsque le modèle linéaire gaussien s'ajuste bien aux données, on s'attend donc à ce qu'à peu près 95% des résidus standardisés se trouvent entre les quantiles d'ordre respectif 2.5% et 97.5% de la loi $\mathcal{N}(0, 1)$.

• **Coefficient de détermination (multiple)**: Puisque $\widehat{\beta} = \arg\min_{\beta} \|\mathbb{Y} - \mathbb{X}.\beta\|$, cela signifie que $\widehat{\mathbb{Y}} = \mathbb{X}.\widehat{\beta}$ est la projection orthogonale de Y sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\mathbb{X}$, noté Vect($\mathbb{X}$). On en déduit que $(\mathbb{Y} - \widehat{\mathbb{Y}})$ est orthogonal à tout vecteur de Vect($\mathbb{X}$). Si le modèle comporte une ordonnée à l'origine (ce qui est assez usuel), alors 1_n est dans Vect($\mathbb{X}$). Comme par définition de la projection orthogonale $\widehat{\mathbb{Y}}$ est dans Vect($\mathbb{X}$), il vient que $\widehat{\mathbb{Y}} - \overline{Y}_n 1_n$ est dans Vect($\mathbb{X}$). D'après le théorème de Pythagore, on en déduit que

$$\|\mathbb{Y} - \overline{Y}_n 1_n\|^2 = \|\mathbb{Y} - \widehat{\mathbb{Y}}\|^2 + \|\widehat{\mathbb{Y}} - \overline{Y}_n 1_n\|^2,$$

ce que l'on peut également écrire sous la forme suivante

$$\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_n)^2$$

ou encore, en remplaçant $(Y_i - \widehat{Y}_i)$ par $\widehat{\varepsilon}_i$,

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}_{\text{SCR}} + \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_n)^2}_{\text{SCE}}$$

en notant

- SCT = somme des carrés totale,
  = variance empirique de $\mathbb{Y}$ au coefficient multiplicateur près
- SCRes = somme des carrés des résidus,
- SCEx = somme des carrés expliquée par le modèle de régression
  = variance empirique de $\widehat{\mathbb{Y}}$ au coefficient multiplicateur près.
  En effet, si le modèle comporte une ordonnée à l'origine, alors 1_n est dans Vect($\mathbb{X}$). Comme $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ est dans Vect($\mathbb{X}$)$^\perp$, il vient que $1_n.\varepsilon = \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$ donc $\frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0 \Longleftrightarrow$ puis $\frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}_n$.

Le coefficient de détermination (multiple) est défini par

$$\begin{aligned} R^2 &= \frac{\text{SCEx}}{\text{SCT}} = \frac{\|\widehat{\mathbb{Y}} - \overline{\mathbb{Y}}_n 1_n\|^2}{\|\mathbb{Y} - \overline{\mathbb{Y}}_n 1_n\|^2} \\ &= \frac{\text{SCT} - \text{SCRes}}{\text{SCT}} = 1 - \frac{\text{SCRes}}{\text{SCT}} = 1 - \frac{\|\mathbb{Y} - \widehat{\mathbb{Y}}\|^2}{\|\mathbb{Y} - \overline{\mathbb{Y}}_n 1_n\|^2} \end{aligned}$$

Ceci montre que l'on a toujours $0 \leq R^2 \leq 1$. Le coefficient de détermination (multiple) évalue l'ajustement global du modèle aux données en ce qu'il exprime la proportion de variabilité de Y expliquée par le modèle. Dans le cas extrême où $R^2 = 1$, on a $\mathbb{Y} \in$ Vect($\mathbb{X}$) ce qui signifie que le modèle proposé explique parfaitement $\mathbb{Y}$. Dans le cas extrême où $R^2 = 0$, on a $\widehat{Y}_i = \overline{\mathbb{Y}}_n$ pour $i = 1, ..., n$ ce qui signifie que le modèle proposé n'apporte aucune information sur $\mathbb{Y}$.
Lors d'une procédure de sélection de variables, on pourrait penser à trouver la combinaison de variables qui maximise le R^2. En réalité, ce critère ne convient pas. En effet, le R^2 augmente de manière mécanique avec le nombre de variables: plus on ajoute de variables, plus il est proche de 1 même si les variables ajoutées ne sont pas pertinentes. Ainsi dans un processus de sélection de variables, le R^2 convient uniquement pour comparer des solutions comportant le même nombre de variables, ce qui est bien trop réducteur.

• **Coefficient de détermination partiel (ou R^2 ajusté ou corrigé):** Ce coefficient est une modification du R^2 qui tient compte du nombre de degrés de liberté à savoir du nombre de coefficients non liés à estimer introduites dans le modèle pour l'espérance conditionnelle de $\mathbb{Y}$.

$$\widetilde{R}^2 = 1 - \frac{\text{SCRes}/(n-(p+1))}{\text{SCT}/(n-1)} = 1 - \frac{\|\mathbb{Y} - \widehat{\mathbb{Y}}\|^2/(n-(p+1))}{\|\mathbb{Y} - \overline{\mathbb{Y}}_n 1_n\|^2/(n-1)} = \frac{(n-1)R^2 - p}{n-(p+1)}.$$

Le critère $\widetilde{R}^2$ puisse prendre des valeurs négatives et satisfait toujours $\widetilde{R}^2 < R^2$.
Puisqu'on a toujours $\widetilde{R}^2 < R^2$, dans le cadre d'une procédure de sélection de variables, on choisit le modèle le plus simple ayant un $\widetilde{R}^2$ le plus proche de la limite supérieure R^2.
L'introduction du R^2 ajusté avait pour objectif de rendre comparable des régressions comportant un nombre différent de covariables. En réalité, l'effet contraignant du nombre de degrés de liberté (qui est $(p+1)$ de manière générique ici) n'est pas assez fort et ce nouveau critère favorise encore trop les solutions comportant un grand nombre de variables.

• **Levier d'une observation**:
On appelle matrice des leviers (*Hat matrix*) la matrice H donnée par

$$H = \mathbb{X}.(\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t\,.$$

Le levier de l'observation i est le $i^{\text{ème}}$ coefficient diagonal de H.
On a obtenu la relation $\widehat{\mathbb{Y}} = H.\mathbb{Y}$. On en déduit que, pour $i = 1, ..., n$, la valeur ajustée $\widehat{Y_i}$ est donnée par la relation:

$$\widehat{Y_i} = \sum_{j=1}^{n} H_{ij}Y_j = h_{ii}Y_i + \sum_{j\neq i} H_{ij}Y_j.$$

Ainsi, H_{ii}, le $i^{\text{ème}}$ coefficient diagonal de H noté plus simplement h_i, représente le poids de Y_i sur sa reconstruction par le modèle à savoir $\widehat{Y_i}$. Or, H étant une matrice de projection, on en déduit que

- pour tout $i \in \{1, ..., n\}$, $0 \leq H_{ii} \leq 1$,
- pour tout $(i, j) \in \{1, ..., n\}^2$ avec $i \neq j$, $-\dfrac{1}{2} \leq H_{ij} \leq \dfrac{1}{2}$,
- $H_{ii} = 0 \Longrightarrow H_{ij} = 0$ pour tout $j \neq i$,
- $H_{ii} = 1 \Longrightarrow H_{ij} = 0$ pour tout $j \neq i$.

De plus, on la propriété suivante:

$$\sum_{i=1}^{n} h_i = \text{Tr}(H) = \text{Tr}\left(\mathbb{X}.(\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t\right) = \text{Tr}\left((\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t\mathbb{X}.\right) = \text{Tr}(I_{p+1}) = p + 1.$$

Cela a conduit à poser que l'observation i est un point influent, selon les auteurs,

- ou bien, si $h_i > 1/2$,
- ou bien, si $h_i > (p + 1)/n$.

• **critère du PRESS (***Predicted Residual Error Sum of Squares***):** La statistique du PRESS est une mesure du pouvoir prédictif d'un modèle qui est construite d'après le principe de la validation croisée dans sa version *leave-one-out*. L'idée à la source de la validation croisée est qu'en matière de validation, une observation ne peut être à la fois juge et partie. De manière générale, la méthode consiste alors à diviser l'échantillon en un sous-échantillon d'apprentissage (*learning data*) et un sous-échantillon de validation (*test data*). On estime ensuite les paramètres de la régression à partir du sous-échantillon d'apprentissage. Puis, on utilise le modèle ainsi estimé pour prédire les réponses du sous-échantillon de validation. Puis, en mesurant l'écart entre les valeurs observées dans le sous-échantillon de validation et leur valeur ainsi prédite, on évalue la performance prédictive du modèle.
Dans la version *leave-one-out* de la validation croisée, chacune des observations est utilisée tour à tour comme sous-échantillon de validation, les $(n - 1)$ observations restantes étant alors utilisées comme sous-échantillon d'apprentissage. La procédure de validation croisée dans sa version *leave-one-out* s'écrit:
pour chaque observation $i \in \{1, ..., n\}$,

- retirer l'observation i des données,

- calculer l'estimation des paramètres de régression à partir des $(n-1)$ observations restantes, ce que l'on note $\widehat{\beta}_{(-i)}$,
- prédire la réponse de l'individu i à partir du modèle ainsi estimé, $\widehat{Y}_{i,(-i)} = \mathbb{X}_i.\widehat{\beta}_{(-i)}$,
- mesurer l'écart (quadratique ici) entre l'observation Y_i et sa prédiction $\widehat{Y}_{i,(-i)}$,

enfin, sommer les écarts obtenus. Le critère de PRESS mesure l'écart entre les valeurs observées et les valeurs ainsi prédites sans que l'observation correspondante serve à calculer les estimations des paramètres:

$$PRESS = \|\mathbb{Y} - \widehat{\mathbb{Y}}_{-i}\|^2 = \sum_{i=1}^{n}(y_i - \widehat{Y}_{i,-i})^2$$

en notant $\mathbb{Y}_{-i} = \begin{pmatrix} \widehat{Y}_{1,-1} \\ \vdots \\ \widehat{Y}_{n,-n} \end{pmatrix}$ la prédiction de l'endogène où $\widehat{Y}_{i,-i}$ est la prévision effectuée pour l'observation i utilisée en donnée supplémentaire et donc non prise en compte lors de l'ajustement du modèle. La formule suivante permet d'éviter d'ajuster n régressions:

$$PRESS = \sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{1 - h_i}\right)^2.$$

On calcule le critère du PRESS sur un certain nombre de modèles linéaires candidats. Le modèle correspondant au PRESS le plus faible est le modèle le plus performant du point de vue de la prédiction. Les modèles sur-paramétrés, donc sujets aux problèmes de sur-ajustement, ont tendance à produire de petits résidus bruts pour les observations prises en compte lors de l'ajustement mais de grands résidus bruts pour les prédictions d'observations exclues de l'ajustement.

• 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.
Les valeurs des leviers sont disponibles au moyen de l'instruction

```
hatvalues(mylm)
```

La fonction `rstandard` fournit les résidus standardisés à partir d'un modèle ajusté avec la fonction `lm` du package `stats`, chargé par défaut:

```
rstandard(mylm)
```

Exercice 1.
Dans le cadre d'une procédure de choix des variables à inclure dans un modèle de régression linéaire gaussien standard, illustrer de manière empirique à partir de données simulées le comportement des critères du R^2, du $\widetilde{R}^2$, du PRESS et des résidus standardisés, 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.