Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #10 (diagnostique)

Régression Linéaire

A partir d'un modèle linéaire: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
deux étapes sont importantes

- **d'estimation et d'inférence**: supposé que $\mathbf{X}$ est de plein rang ($\mathcal{H}_1$) et estimé le paramètre $\boldsymbol{\beta}$ par MCO (sous l'hypothèse $\mathcal{H}_2$) ou par MV (sous l'hypothèse $\mathcal{H}_2^{\text{Gauss}}$). Grâce à ces hyp.
  - obtention de propriétés statistiques pour $\hat{\boldsymbol{\beta}}$.
  - IC , RC , tests, intervalle de prédiction
- **de validation**:
  - $\mathcal{H}_1$ fausse ou "pas loin", estimation ridge (ou lasso)
  - il faut valider les autres hypothèses qui permettent de valider l'utilisation des outils d'inférence.
  - Validation d'un individu.
  - Validation du modèle global, linéarité des régresseurs.
  - Choix ou non d'inclure un régresseur.

Régression Linéaire

- Par définition: $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.
- On souhaite utiliser le fait que $\boldsymbol{\varepsilon} \approx \hat{\boldsymbol{\varepsilon}}$ pour vérifier des hypothèses faites sur $\boldsymbol{\varepsilon}$.
- **Rappel**: $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbb{I}_n$ (d'après $\mathcal{H}_2$) et

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \quad \text{et} \quad \mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbb{I}_n - \boldsymbol{\mathcal{P}}_{\mathbf{X}}) = \sigma^2(\mathbb{I}_n - \mathbf{H})$$

- Naturellement, on peut définir une version normalisée:

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n.$$

- **Problème**: même sous $\mathcal{H}_2^{\text{Gauss}}$, la loi de t_i n'est pas connu: $\hat{\varepsilon}_i$ et $\hat{\sigma}$ ne sont pas indépendantes.
- **Solution**: estimer σ^2 indépendandamment de $\hat{\varepsilon}_i$.

Pour $i = 1, \ldots, n$, on définit:

- $\mathbf{X}_{(i)}$: la matrice de design $\mathbf{X}$ sans la $i$ème ligne.
- $\hat{\boldsymbol{\beta}}_{(i)}$ et $\hat{\sigma}^2_{(i)}$: les estimateurs de $\boldsymbol{\beta}$ et σ^2 basés sur toutes les observations hormis celle du $i$ème individu.
- on note $\hat{\varepsilon}_i^{(i)} = Y_i - \hat{Y}_{(i)} = Y_i - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}$ = erreur de prédiction de la i-ème variable Y_i sans utiliser l'information du i-ème individu.
- Une alternative au t_i est de définir $\hat{\varepsilon}_i^{(i)} / \sqrt{\hat{\mathrm{Var}}(\hat{\varepsilon}_i^{(i)})}$

On peut démontrer sous $\mathcal{H}_1$ et si la suppression de la $i$ème ligne ne modifie pas le rang de $\mathbf{X}$, que

$$\hat{\varepsilon}_i^{(i)} = \frac{\hat{\varepsilon}_i}{1 - h_{ii}}.$$

Donc en particulier,

- très intéressant car calculer $\hat{\varepsilon}_i^{(i)}$ n'est pas coûteux.
- $\mathrm{Var}(\hat{\varepsilon}_i^{(i)}) = \sigma^2/(1-h_{ii})$.
- L'idée est de définir un estimateur de σ^2 qui ne dépende pas de ε_i (et donc de $\hat{\varepsilon}_i^{(i)}$). En particulier

$$\hat{\sigma}^2_{(i)} = \left\{ \frac{1}{n-p-1} \sum_{j=1, j\neq i}^{n} \left(Y_j - (\mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)})_j \right)^2 \right\} / (1-h_{ii})$$

Sous $\mathcal{H}_1$ et si la suppression de la $i$ème ligne ne modifie pas le rang de $\mathbf{X}$, on définit

$$t_{(i)} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}.$$

Sous $\mathcal{H}_2$ [Gauss], $t_{(i)} \sim \mathcal{S}td_{n-p-1}$.

Régression Linéaire

- Il est préférable d'utiliser $t_{(i)}$ aux t_i car les variances sont plus homogènes.
- $\hat{\sigma}_{(i)}$ n'est pas influencé par des erreurs grossières sur la $i$ème observation.
- Même si $\mathcal{H}_2^{\text{Gauss}}$ n'est pas vérifiée (ou pas vérifiable), la définition reste pertinente. En revanche, il faut espérer que n soit grand ou pouvoir estimer la loi des t_i par bootstrap.
- On peut montrer que

$$t_{(i)} = t_i \sqrt{\frac{n-p-1}{n-p-t_i^2}}$$

Une donnée (c'est-à-dire un couple $(\mathbf{x}_i', y_i)$) est considérée comme aberrante si $t_{(i)}$ est anormalement élevée, c'est-à-dire sous $\mathcal{H}_2^{\text{Gauss}}$ si $|t_{(i)}| > t_{1-\alpha/2,n-1-p}$.

Régression Linéaire

- La matrice $\mathcal{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ joue un rôle important, e.g.

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathcal{P}_{\mathbf{X}}\mathbf{Y}; \quad \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{P}_{\mathbf{X}}\boldsymbol{\varepsilon}.$$

- En général, on note $\mathbf{H} = \mathcal{P}_{\mathbf{X}}$, et on a:

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

- Le terme h_{ii} est appéle "poids" de l'observation i sur sa propre estimation.
- Rappels: $h_{ii} \in [0, 1]$, $h_{ij} \in [-1/2, 1/2]$ pour $j \neq i$ et si $h_{ii} \in \{0, 1\}$ alors $h_{ij} = 0$ pour $j \neq i$; $\mathrm{tr}(\mathbf{H}) = \sum h_{ii} = p$
- $\Rightarrow$
  - si $h_{ii} = 1$, $\hat{Y}_i$ est entièrement déterminé par Y_i.
  - si $h_{ii} = 0$, Y_i n'a pas d'influence sur $\hat{Y}_i$.
  - plus h_{ii} est grand et plus Y_i participe à sa propre prédiction.

Un point i est un point levier si la valeur de la matrice de projection dépasse les valeurs suivantes

- $h_{ii} > \frac{2p}{n}$ selon Hoaglin and Welsch, 1978.
- $h_{ii} > \frac{3p}{n}$ pour $p > 6$ et $n - p > 12$ selon Velleman and Welsch, 1981.
- $h_{ii} > 0.5$ selon Huber, 1981.

La distance de Cook, calculée pour tout individu, est définie pour $i = 1, \ldots, n$ par

$$C_i = \frac{1}{p\hat{\sigma}^2} \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)^\top (\mathbf{X}^\top \mathbf{X}) \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)$$

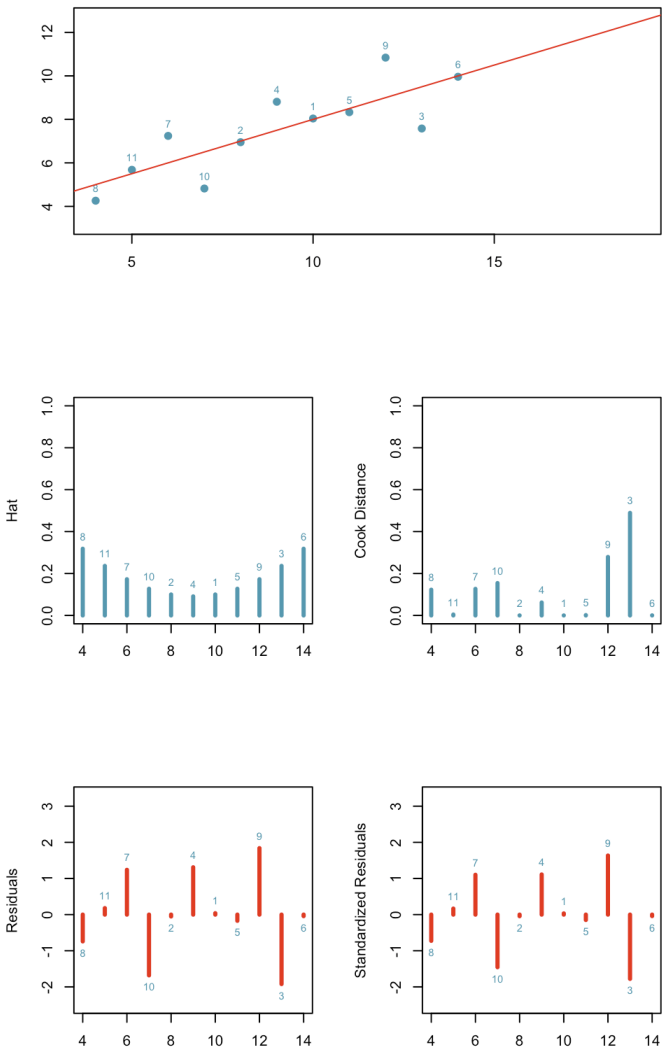
où $\hat{\boldsymbol{\beta}}_{(i)}$ est l'estimation par MCO obtenue sur le jeu de données privé du $i$ème individu. On peut montrer que

$$C_i = \frac{h_{ii}}{p(1-h_{ii})} t_i^2 = \frac{h_{ii}}{p(1-h_{ii})^2} \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2}, \qquad \text{où } t_i \approx t_{(i)}.$$

compromis entre point levier (fort h_{ii}) et point aberrant (fort $t_{(i)}$).

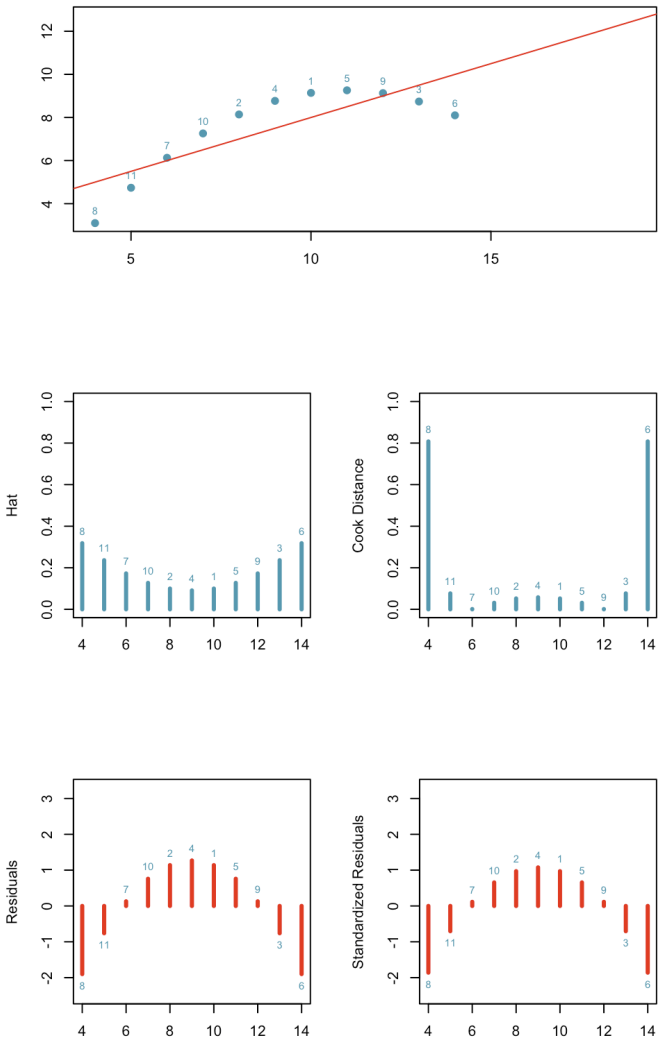
Anscombe

```
> data(anscombe)
> summary(lm(y1~x1,data=anscombe))

Coefficients:
   Estimate    Std  t value Pr(>|t|)
(Int)  3.00    1.1247    2.667  0.025
x1     0.50    0.1179    4.241  0.002

Residual standard error: 1.237
Multiple R-squared: 0.6665
Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF
p-value: 0.00217
```



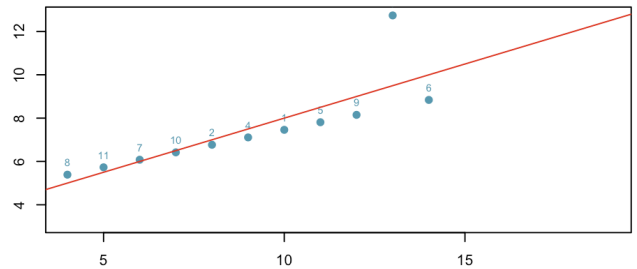
Anscombe

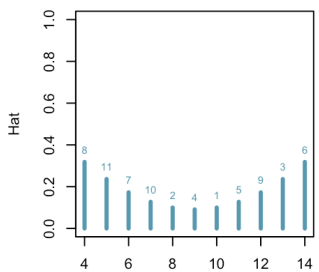
```
> data(anscombe)
> summary(lm(y2~x2,data=anscombe))

Coefficients:
   Estimate    Std  t value Pr(>|t|)
(Int)  3.00    1.125    2.667  0.025
x2     0.50    0.118    4.239  0.002

Residual standard error: 1.237
Multiple R-squared: 0.6662
Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF
p-value: 0.00218
```

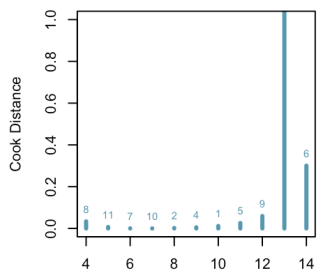


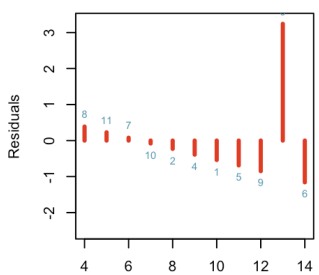
Anscombe

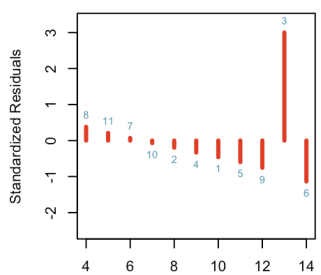
```
> data(anscombe)
> summary(lm(y3~x3,data=anscombe))

Coefficients:
   Estimate    Std  t value Pr(>|t|)
(Int)  3.00    1.1245     2.667
    0.025
x3      0.50    0.1179     4.239
    0.002

Residual standard error: 1.236
Multiple R-squared: 0.6663
Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF
p-value: 0.00218
```











Anscombe

```
> data(anscombe)
> summary(lm(y4~x4,data=anscombe))

Coefficients:
   Estimate    Std  t value Pr(>|t|)
(Int)  3.00    1.1239     2.671
    0.025
x3      0.49    0.1178     4.243
    0.002

Residual standard error: 1.236
Multiple R-squared: 0.6663
Adjusted R-squared: 0.6297
F-statistic: 17.97 on 1 and 9 DF
p-value: 0.00218
```