MAT 3775
Analyse de la régression

Chapitre 6
Observations influentes et valeurs aberrantes

P. Boily (uOttawa)

Session d'hiver – 2023

Aperçu

6 – Observations influentes et valeurs aberrantes

Lorsque nous travaillons avec un seul prédicteur, nous pouvons généralement déterminer rapidement si une prédiction ou une réponse est inhabituelle, dans un certain sens.

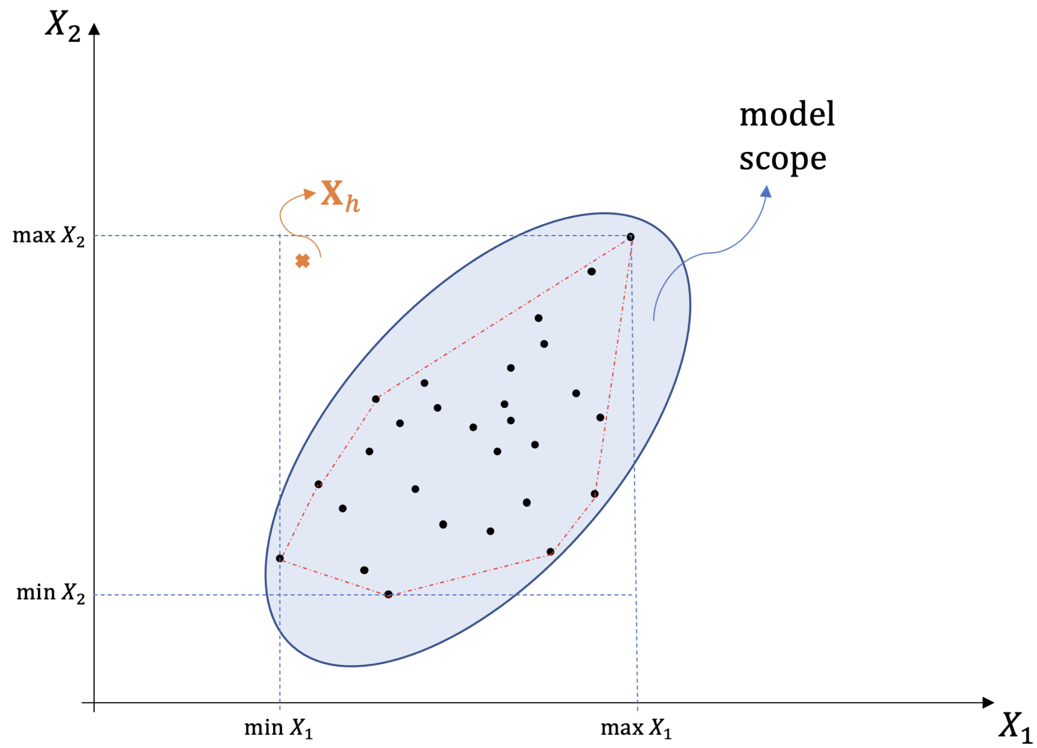
Si un prédicteur est beaucoup plus petite/grande que les autres valeurs de prédiction, nous hésitons à utiliser le modèle de régression pour prédire la réponse car aucune valeur similaire n'a été utilisée pour "former" le modèle.

Lorsque $p > 1$, trouver les observations anormales (prédicteurs et/ou réponses) n'est pas aussi évident.

Dans ce chapitre, nous présentons un petit nombre de méthodes pour y parvenir (il en existe beaucoup plus, voir DUDADS).

6.1 – Effet de levier et extrapolation cachée

Considérons un ensemble de données avec deux prédicteurs X_1, X_2.



Les modèles de régression ne sont généralement utiles que lorsque nous travaillons dans la **portée du modèle** ; la régression est une tentative d'**interpolation** et nous devons éviter les situations d'**extrapolation**.

Nous ne pouvons pas toujours facilement voir si un prédicteur $\mathbf{X}_h$ est dans la portée du modèle ou non ; dans l'image précédente, chaque composante de $\mathbf{X}_h$ est dans la portée des prédicteurs utilisés pour construire le modèle, mais $\mathbf{X}_h$ dans son ensemble **ne l'est pas**. Lorsque p est grand, cette approche **visuelle** échoue.

Le **levier** de la $i$ème observation est :

$$h_{ii} = \mathbf{X}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i, \quad \mathbf{X}_i \text{ est la } i\text{ème rangée de } \mathbf{X};$$

c-à-d que h_{ii} est le $i$ème élément diagonal de $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

Le **levier** détermine si un niveau de prédiction $\mathbf{X}_h$ est dans le **champ d'application du modèle** : si

$$\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h > \max\{h_{ii} \mid i = 1, \ldots, n\},$$

il ne l'est pas et $\hat{Y}_h = \mathbf{X}_h\mathbf{b}$ est une **extrapolation cachée**.

Remarquons que $0 \leq h_{ii} \leq 1$, for $i = 1, \ldots, n$. En effet, comme :

1. $\mathbf{0} \leq \sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\{\mathbf{HY}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}^\top = \sigma^2\mathbf{H} \implies h_{ii} \geq 0$, $\forall i$

2. $\mathbf{0} \leq \sigma^2\{\mathbf{e}\} = \sigma^2\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}) \implies 1 - h_{ii} \geq 0$, $\forall i$

En générale, la surface de $\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h = c$ est un ellipsoïde de centre $\overline{\mathbf{X}} = (1, \overline{X}_1, \ldots, \overline{X}_p)$ (plus c est grand, plus la "distance" à $\overline{\mathbf{X}}$ est grande).

Une **valeur aberrante en X** est une observation qui est **atypique** par rapport aux **niveaux du prédicteur**.

On note que

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\mathrm{tr}(\mathbf{H}) = \frac{p}{n} \quad (p \leq n);$$

1. si $h_{ii} \leq 0.2$, l'effet de levier du $i$ème cas est **faible** (près de $\overline{\mathbf{X}}$) ;
2. si $0.2 < h_{ii} < 0.5$, l'effet de levier du $i$ème cas est **modéré** ;
3. si $h_{ii} \geq 0.5$, l'effet de levier du $i$ème cas est **élevé** (anomalie potentielle) ;
4. si n est élevé et $h_{ii} > 3\overline{h} = \frac{3p}{n}$, le $i$ème cas est une valeur aberrante en X.

Exemple : on souhaite ajuster le modèle de RLG

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

à un ensemble de données comportant n observations, avec

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 1.17991 & -0.00731 & 0.00073 \\ -0.00731 & 0.00008 & -0.00012 \\ 0.00073 & -0.00012 & 0.00046 \end{pmatrix} \quad \text{et} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 220 \\ 36768 \\ 9965 \end{pmatrix}$$

Quelles sont les estimations ponctuelles des coefficients de régression $\boldsymbol{\beta}$? Nous aimerions prédire la valeur de Y_h lorsque $X_1 = 200$ et $X_2 = 50$, c-à-d au point $\mathbf{X}_h = (1, 200, 50)^\top$. Quel est l'effet de levier de $\mathbf{X}_h$? S'agit-il d'un cas d'extrapolation cachée ? Sinon, quelle est la valeur prédite Y_h?

Solution : les estimation de moindres carrés des coefficients de régression sont

$$\mathbf{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} -1.91943 \\ 0.13744 \\ 0.33234 \end{pmatrix}.$$

L'effet de levier de $\mathbf{X}_h$ est

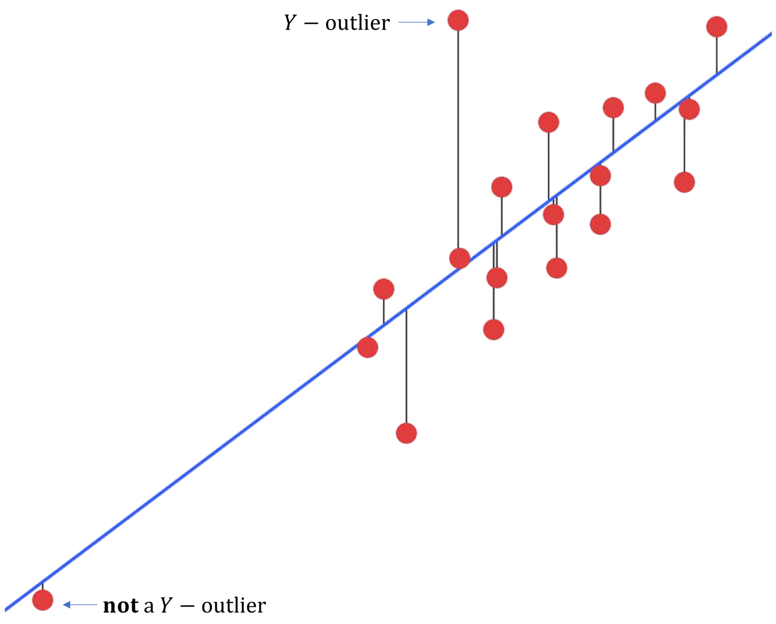
$$\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h = 0.27891;$$

il est suffisamment faible pour suggérer que nous ne sommes pas dans une situation d'extrapolation cachée (bien que n soit inconnu, donc nous ne pouvons pas le comparer à $\frac{3p}{n}$).

La réponse prédite à $\mathbf{X}_h$ est ainsi $\hat{Y}_h = \mathbf{X}_h^\top\mathbf{b} = 42.18557$.

6.2 – Résidus supprimés studentisés

Les valeurs aberrantes en X sont déterminées sans référence à une **surface de régression** $\hat{Y}(\mathbf{x}) = \mathbf{xb}$; nous pouvons également identifier des observations dont les réponses sont **inattendument distantes** de $\hat{Y}(\mathbf{x})$.



Une **valeur aberrante en Y** est une observation qui produit un **large** résidu de régression.

1. Si le **résidu studentisé (interne)** est suffisamment grand,

$$|r_i| = \left|\frac{e_i}{\mathrm{s}\{e_i\}}\right| = \left|\frac{e_i}{\sqrt{\mathrm{MSE}}\sqrt{1-h_{ii}}}\right| \geq 3,$$

   disons, alors la $i$ème observation est aberrante en Y ;

2. Une autre approche : supprimer le $i$ième cas du modèle et ré-ajuster

$$\mathbf{b}_{(i)} = \left(\mathbf{X}_{(i)}^{\top}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^{\top}\mathbf{Y}_{(i)},$$

   pour obtenir une valeur prédite pour la $i$ème observation, $\hat{Y}_{i(i)}$.

Pour $i = 1, \ldots, n$, le **résidu supprimé** est $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}$ et la **studentisation externe** est

$$t_i = \frac{d_i}{\mathrm{s}\{d_i\}} = e_i \sqrt{\frac{n-p-1}{\mathrm{SSE}(1-h_{ii}) - e_i^2}} \sim t(n-p-1),$$

où

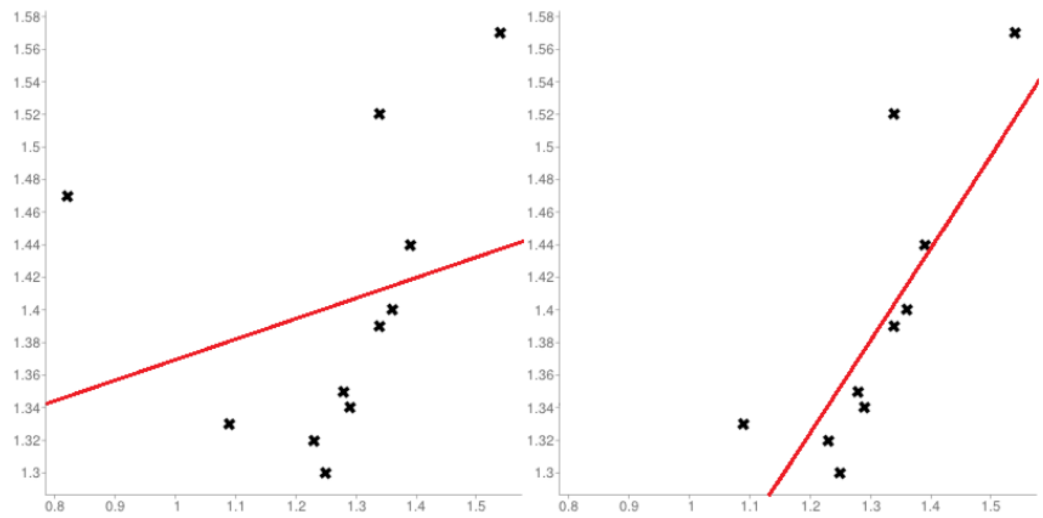
$$\mathrm{s}^2\{d_i\} = \mathrm{MSE}_{(i)} \left[1 + \mathbf{X}_i \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_i^\top\right].$$

Règle de décision : si $|t_i| > t(1 - \frac{\alpha/n}{2}; n-p-1)$, la $i$ème observation est aberrante en Y à un niveau de confiance α.

Il est possible qu'une observation soit aberrante en X sans l'être en Y, et *vice-versa*.

6.3 – Observations influentes

Nous pouvons également nous intéresser aux observations **influentes** – des observations dont l'absence (ou la présence) dans les données modifie de manière significative (qualitiative) la **nature de l'ajustement**.



Les observations influentes peuvent ne pas être aberrante, et *vice-versa*.

DFFITS_i est une mesure de l'**influence** du $i$ème cas sur $\hat{Y}$ dans un voisinage de $\mathbf{X}_i$. La **différence par rapport à la valeur ajustée** est

$$\mathrm{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\mathrm{MSE}_{(i)}\, h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

Pour les échantillons de petite et moyenne taille, si $|\,\mathrm{DFFITS}_i\,| > 2$, alors le $i$ème cas est **probablement influent**. Pour les échantillons plus grands, si $|\,\mathrm{DFFITS}_i\,| > 2\sqrt{\frac{p}{n}}$, alors le $i$ème cas est **influent**.

Une mesure similaire peut être déterminée pour savoir si le cas i a beaucoup d'influence sur la valeur du **paramètre ajusté** b_k :

$$\mathrm{DFBETAS}_i^k = \frac{b_k - b_{k(i)}}{\sqrt{\mathrm{MSE}_{(i)}\,[(\mathbf{X}^\top\mathbf{X})^{-1}]_{k,k}}}.$$

6.4 – Distance de Cook

La **distance de Cook** mesure également l'influence du $i$ème cas :

$$D_i = \frac{1}{p \cdot \mathrm{MSE}} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2 = \frac{e_i^2}{p \cdot \mathrm{MSE}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right] \sim F(p, n-p).$$

Règle de décision :

- si $D_i < F(0.2; p; n-p)$, le $i$ème cas a **peu d'influence**;
- if $D_i > F(0.5; p; n-p)$, le $i$ème cas est **très influentiel**.

L'approche des moindres carrés est pratique, mais elle n'est pas **robuste** contre la présence de cas aberrants/influents (médiane, valeur absolue).

Exemple : soient

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 4 & 2 \end{pmatrix} \quad \text{et} \quad \mathbf{Y} = \begin{pmatrix} 2.1 \\ 24.2 \\ 29.5 \\ 27.6 \\ 30.5 \\ 27.5 \end{pmatrix}.$$

Y a-t-il des observations aberrantes en X et/ou en Y, ou encore influentes?

Solution : comme $n = 6$, l'échantillon est petit. Le vecteur coefficient est

$$\mathbf{b} = \begin{pmatrix} -7.3 \\ 5.51 \\ 5.70 \end{pmatrix},$$

pour lequel

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b} = (-1.8, 3.2, -2.7, 1.28, -1.32, 1.37)^\top.$$

Les résidus externes sont $(-18.47, 2.40, -1.99, 0.41, -0.5, 0.57)^\top$. Puisque

$$t\left(1 - \frac{\alpha/n}{2}; n-p-1\right) = t\left(1 - \frac{0.1/6}{2}; 6-3-1\right) = 7.65,$$

seul la première observation est aberrante en Y lorsque $\alpha = 0.1$; de façon conservatrice, lorsque $|t_i|$ est grand, nous devrions étudier davantage l'influence du cas i ; nous ne manquerons pas d'examiner le **cas** 1 en détail.

(Notez le terme de correction de Bonferroni).

Pour les valeur aberrantes en X, nous cherchons $h_{ii} > 0.5$:

$$\mathbf{h} = (0.87, 0.45, 0.58, 0.19, 0.41, 0.48)^\top.$$

Les cas 1,3 ont des effets de levier **élevés**, suggérant qu'ils sont possiblement aberrant en X ; les cas 2,5,6 ont des effets de levier **modérés** (peu probable qu'ils soient aberrant en X, à moins que 5/6 observations le soient).

Nous avons également

$$\text{DFFITS} = (-48.7, 2.29, -2.33, 0.2, -0.42, 0.54)^\top,$$

ce qui suggère que seuls les 3 premiers cas sont influents. Les **distances de Cook** sont $\mathbf{D} = (6.9, 0.67, 0.91, 0.02, 0.08, 0.13)^\top$**;** comme D_1 est la seule distance plus grande que $F(0.5; p, n-p) = 1$, seul le **premier** cas est susceptible d'être influent.