MAT 3775
Analyse de la régression

Chapitre 4
Extensions de la régression linéaire

P. Boily (uOttawa)

Session d'hiver – 2023

Aperçu

4 – Extensions de la régression linéaire

Nous avons vu que nous pouvons assez facilement étendre la régression linéaire simple à la régression linéaire multiple avec un minimum de perturbations, simplement en utilisant la notation matricielle appropriée.

Dans la pratique, les hypothèses de la RLG sont rarement respectées ; nous avons également présenté les moyens d'identifier les écarts par rapport aux hypothèses, et la manière de remédier à ces situations.

Dans ce chapitre, nous aborderons des extensions de la régression linéaire plus sophistiquées, des extensions qui se rapprochent des applications de la vie réelle.

4.1 – Mutlicollinéarité et inflation de la variance

Les **équations normales** de la régression linéaire multiple sont

$$(\mathbf{X}^\top \mathbf{X})\mathbf{b} = \mathbf{X}^\top \mathbf{Y}.$$

Lorsque $\mathbf{X}^\top \mathbf{X}$ est **inversible**, la solution $\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$ est **unique**.

Si une variable est combinaison linéaire non triviale des autres variables

$$X_k = \alpha_{j_1} X_{j_1} + \cdots + \alpha_{j_\ell} X_{j_\ell},$$

alors $\mathrm{rank}(\mathbf{X}^\top) = \mathrm{rank}(\mathbf{X}^\top \mathbf{X}) < p$, d'où $\mathbf{X}^\top \mathbf{X}$ est **singulière** (**non inversible**), et la solution n'est pas **unique** (le système est **sous-déterminé**).

Exemple : considérons la matrice de conception et le vecteur réponse

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 3 \\ 1 & 3 & 3 & 6 \end{pmatrix} \quad \text{et} \quad \mathbf{Y} = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix}.$$

Déterminer la RLG $\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Solution : voici les constituants des équations normales

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 3 & 5 & 6 & 11 \\ 5 & 11 & 12 & 23 \\ 6 & 12 & 14 & 26 \\ 11 & 23 & 26 & 49 \end{pmatrix} \quad \text{et} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 5 \\ 13 \\ 14 \\ 27 \end{pmatrix}.$$

La forme échelonée-réduite de $[\mathbf{X}^\top\mathbf{X} \mid \mathbf{X}^\top\mathbf{Y}]$ est

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

ce qui signifie que $\boldsymbol{b} = (-2, 1-s, 1-s, s)$ est une solution $\forall s \in \mathbb{R}$. Nous ne pouvons pas calculer la matrice de variance-covariance correspondante $\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, puisque $\det(\mathbf{X}^\top\mathbf{X}) = 0$.

En pratique, il est possible qu'un prédicteur soit **presqu'**une combinaison linéaire des autres prédicteurs ; dans ce cas, la matrice de conception est presque **singulière** (**mal conditionnée**), ce qui entraîne une **incertitude** dans le calcul de $\mathbf{b}$ (et peut être lié à des "**signes de coefficient erronés**").

Dans une RLG, le **facteur d'inflation de la variance pour** β_k est

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p,$$

où R_k^2 est le coefficient de détermination multiple obtenu lorsque l'on modélise X_k en fonction des $p - 2$ autres variables prédicteurs.

Si X_k est **très près** d'une combinaison linéaire des autres variables, alors $R_k^2 \approx 1$, ce qui donne un VIF_k élevé et influence les estimations des moindres carrés. En pratique, on peut s'attendre à d'important problèmes de **multicollinéarité** lorsque $\max_k \mathrm{VIF}_k > 10$.

On peut réduire le problème à l'aide du **centrage des données**, de la **régression en "crête"**, et de la **régression par les composantes principales**.

Exemple : considérons les données suivantes

X_1	X_2	X_3	X_4	Y
1	1	2.063	1	2.995
2	1	3.184	1	3.773
1	1	2.131	2	2.846
2	1	2.867	2	3.963
1	2	3.104	1	5.291
2	2	3.876	1	6.070
1	2	2.999	2	5.034
2	2	3.865	2	6.014

Comparer les modèles de RLG

$$\mathrm{E}\{Y \mid (X_1, X_2, X_3)\} \quad \text{et} \quad \mathrm{E}\{Y \mid (X_1, X_2, X_4)\}.$$

Solution : la sortie R pour le premier modèle est

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08808    0.25637  -0.344   0.7485
X1           1.15062    0.43523   2.644   0.0574 .
X2           2.45248    0.44756   5.480   0.0054 **
X3          -0.27147    0.48792  -0.556   0.6076
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1236 on 4 degrees of freedom
Multiple R-squared:  0.9947,Adjusted R-squared:  0.9907
F-statistic: 249.1 on 3 and 4 DF,  p-value: 5.303e-05
```

Celle du second est

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.08134    0.22254  -0.365 0.733254
X1           0.91339    0.08411  10.859 0.000408 ***
X2           2.20826    0.08411  26.253 1.25e-05 ***
X4          -0.06812    0.08411  -0.810 0.463473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.119 on 4 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9914
F-statistic: 269.3 on 3 and 4 DF,  p-value: 4.545e-05
```

Les paramètres estimés b_0, b_1, et b_2 sont **assez semblables** dans les deux modèles, mais les erreurs-type sont **bien différentes** ; les intervalles de confiance dans le second modèle sont **bien plus serrés** pour β_1 et β_2 qu'ils ne le sont dans le premier modèle.

Pourquoi est-ce le cas ? Notez que $\mathrm{VIF}_1 \approx \mathrm{VIF}_2 \approx \mathrm{VIF}_4 \approx 1$ dans le 2e cas (les variables prédicteurs sont **linéairement indépendantes**), tandis que $\mathrm{VIF}_1 \approx \mathrm{VIF}_2 \approx \mathrm{VIF}_3 \approx 25$ dans le 1er cas.

Cela ne devrait pas être une surprise, puisque X_3 **est presqu'une combinaison linéaire de** X_1 **et** X_2 :

$$\|X_3 - X_1 - X_2\|_2^2 \approx 0.324,$$

tandis que $\|X_1\|_2^2 \approx 4.47$, $\|X_2\|_2^2 \approx 4.47$, et $\|X_3\|_2^2 \approx 8.70$.

4.2 – Régression polynomiale

Dans un ensemble de données comportant un prédicteur X et une réponse Y, tous deux numériques, si la relation est **non linéaire**, on peut envisager de transformer les données de sorte que la relation entre X' et Y' le **devienne**, d'ajuster un modèle **linéaire** à X' et Y', et d'inverser les résultats pour obtenir une relation entre les X et Y originaux.

On peut aussi créer une suite de prédicteurs

$$X_1 = X, \; X_2 = X^2, \quad \ldots, \quad X_k = X^k$$

et traiter l'ensemble de la situation comme un modèle de RLG

$$\mathrm{E}\{Y|(X_1, \ldots, X_k)\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = \beta_0 + \beta_1 X + \cdots + \beta_k X^k.$$

Exemple : ajustez les données suivantes

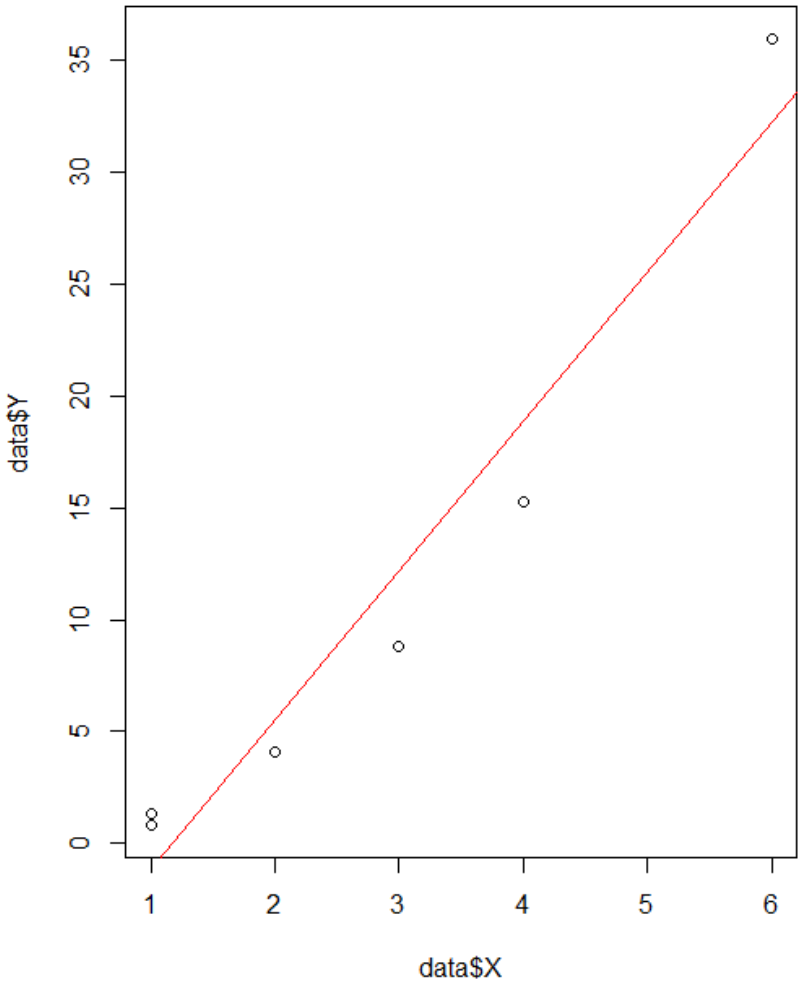
X	1	1	2	4	3	6
Y	0.8	1.3	4.1	15.3	8.8	36

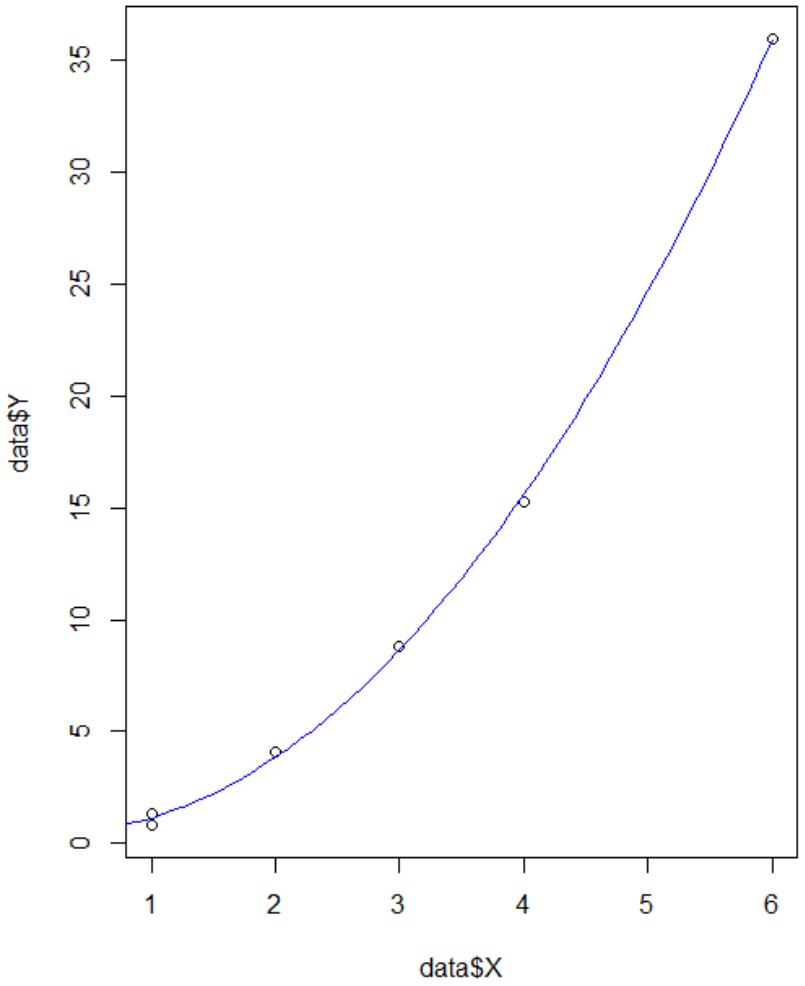
Solution : nous pouvons ajuster un modèle linéaire aux données

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.913      2.734  -2.895  0.04435 *
X              6.693      0.818   8.182  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.55 on 4 degrees of freedom
Multiple R-squared:  0.9436, Adjusted R-squared:  0.9295
F-statistic: 66.94 on 1 and 4 DF,  p-value: 0.001215
```

L'ajustement semble raisonnable ($R_a^2 = 0.9295$), mais un tracé des données suggère que quelque chose tourne pas rond : visuellement, l'ajustement quadratique semble meilleur ($R_a^2 = 0.9994$).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56635    0.47768   1.186 0.321128
X           -0.49591    0.34935  -1.420 0.250809
X2           1.06466    0.05046  21.101 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```


data$Y
data$X
0
5
10
15
20
25
30
35
1
2
3
4
5
6


data$Y
data$X
0
5
10
15
20
25
30
35
1
2
3
4
5
6

À noter: des trois coefficients, seul le coefficient quadratique b_2 est significatif à $\alpha = 0.05$, même si l'ajustement semblait **assez serré**, visuellement. Bien que la relation entre X et X^2 soit **non linéaire**, les puissances sont toujours **corrélées**, ce qui conduit à un VIF assez élevé :

$$\text{VIF}_1 = \frac{1}{1 - R_1^2} = \frac{1}{1 - 0.9510685} = 20.43673.$$

C'est typique de la régression polynomiale : la mesure corrective suggérée est d'utiliser des **prédicteurs centrés** $x_i = X_i - \overline{X}$.

L'ajustement quadratique de l'exemple précédent prendrait la même forme :

$$\begin{aligned}\mathrm{E}\{Y\} &= \beta_0 + \beta_1(X - \overline{X}) + \beta_2(X - \overline{X})^2 \\ &= \left\{\beta_0 - \beta_1\overline{X} + \beta_2\overline{X}^2\right\} + \left\{\beta_1 - 2\beta_2\overline{X}\right\}X + \beta_2X^2 = \beta_0' + \beta_1'X + \beta_2'X^2\end{aligned}$$

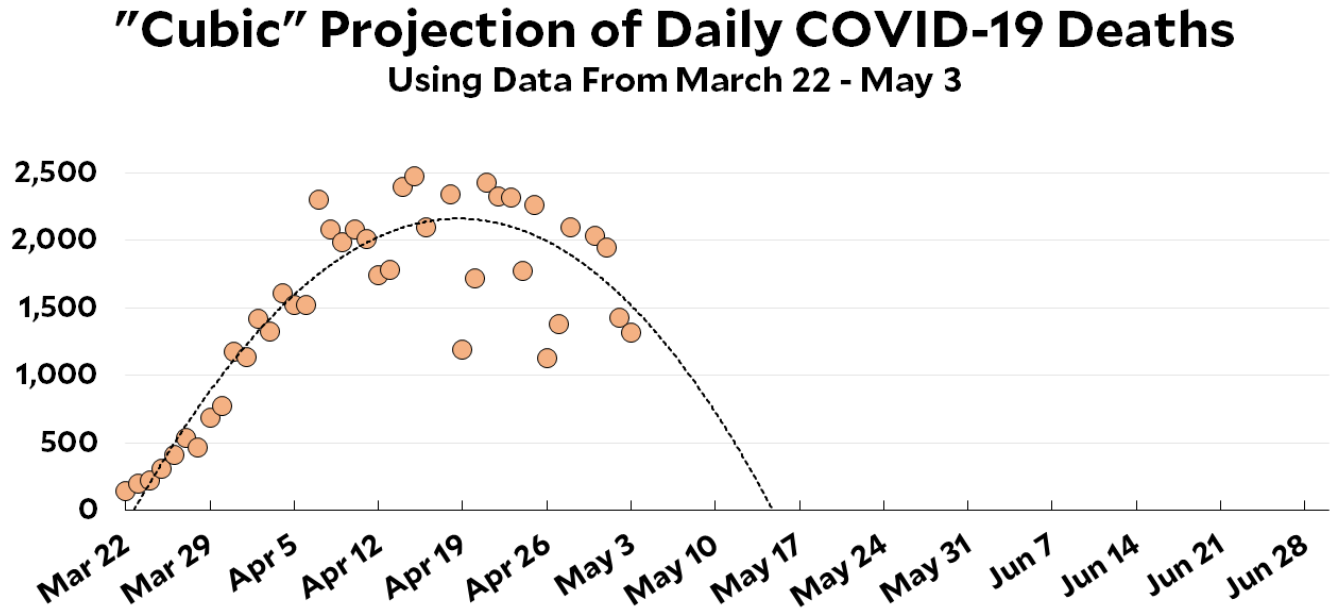
mais maintenant **tous** les coefficients sont significatifs à $\alpha = 0.05$:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.70814    0.20935   36.82 4.41e-05 ***
Xm           5.53718    0.09472   58.46 1.10e-05 ***
X2m          1.06466    0.05046   21.10 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3354 on 3 degrees of freedom
Multiple R-squared:  0.9996,Adjusted R-squared:  0.9994
F-statistic:  3973 on 2 and 3 DF,  p-value: 7.331e-06
```

Ce n'est pas une surprise : le terme VIF_1 du modèle centré est bien plus faible, à 1.502374.

Le reste de la machinerie ordinaire des moindres carrés s'applique facilement.

Graphiquement et/ou mathématiquement, la régression polynomiale peut donc s'avérer très puissante et pratique à utiliser. Mais la commodité n'est pas toujours une raison suffisante pour utiliser un modèle de régression...



Exemple : nous ajustons la réponse à une régression cubique centrée avec prédicteur $x = X - \overline{X}$ en ajoutant une variable à la fois, afin d'obtenir le modèle linéaire simple

$$\mathrm{E}\{Y \mid x\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

Le résumé des sommes de carrés de la régression est présenté ci-dessous :

source de variation	df	SS
x	1	485.71303
$x^2 \mid x$	1	9.11434
$x^3 \mid x, x^2$	1	6.33018
Erreur	23	285.50912

En utilisant $\alpha = 0.05$, testez $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ ou $\beta_3 \neq 0$.

Solution : si H_0 est valide, la statistique

$$F^* = \frac{\mathrm{SSR}(R)/(p-q)}{\mathrm{SSR}(F)/(n-p)} = \frac{\mathrm{SSR}(x^2, x^3|x)/(p-q)}{\mathrm{SSE}(x, x^2, x^3)/(n-p)}$$

suit une loi $F(p-q, n-p)$, où $q = 2$ est le nombre de paramètres dans le **modèle réduit** et $n-p = n-4 = 23$, le # de degrés de liberté de l'erreur, de sorte que $n = 27$.

Si $\alpha = 0.05$, la valeur critique est $F(0.95; 2, 23) = 3.422$. Puisque

$$F^* = \frac{\left[\mathrm{SSR}(x^2|x) + \mathrm{SSR}(x^3|x, x^2)\right]/2}{\mathrm{SSE}(x, x^2, x^3)/23} = \frac{(9.114 + 6.332)/2}{285.509/23} = 0.622,$$

alors $F^* < F(0.95; 2, 23)$ et on **ne rejette pas** H_0 à ce niveau de confiance.

4.3 – Effets d'interaction

Nous avons vu que nous pouvons augmenter la RLS en X afin d'inclure des termes de puissance supérieure (après avoir centré les données pour minimiser les effets de la multicolinéarité).

Rien ne nous empêche de le faire avec un nombre quelconque de prédicteurs $X_1, \ldots, X_p$, ce qui conduit à un **modèle additif**

$$E\{Y\} = f_1(X_1) + \cdots + f_p(X_p),$$

où les f_i sont des **fonctions polynomiales** à 1 variable (elles pourraient être toute fonction linéaire des coefficients de régression $\beta_{i,j}$).

Dans ce qui suit, on suppose que $p = 2$ pour des raisons de simplicité.

Nous pouvons ajouter un **terme d'interaction** $f_3(X_1, X_2) = \beta_3 X_1 X_2$. En accord avec le **principe hiérarchique**, nous pouvons considérer le modèle

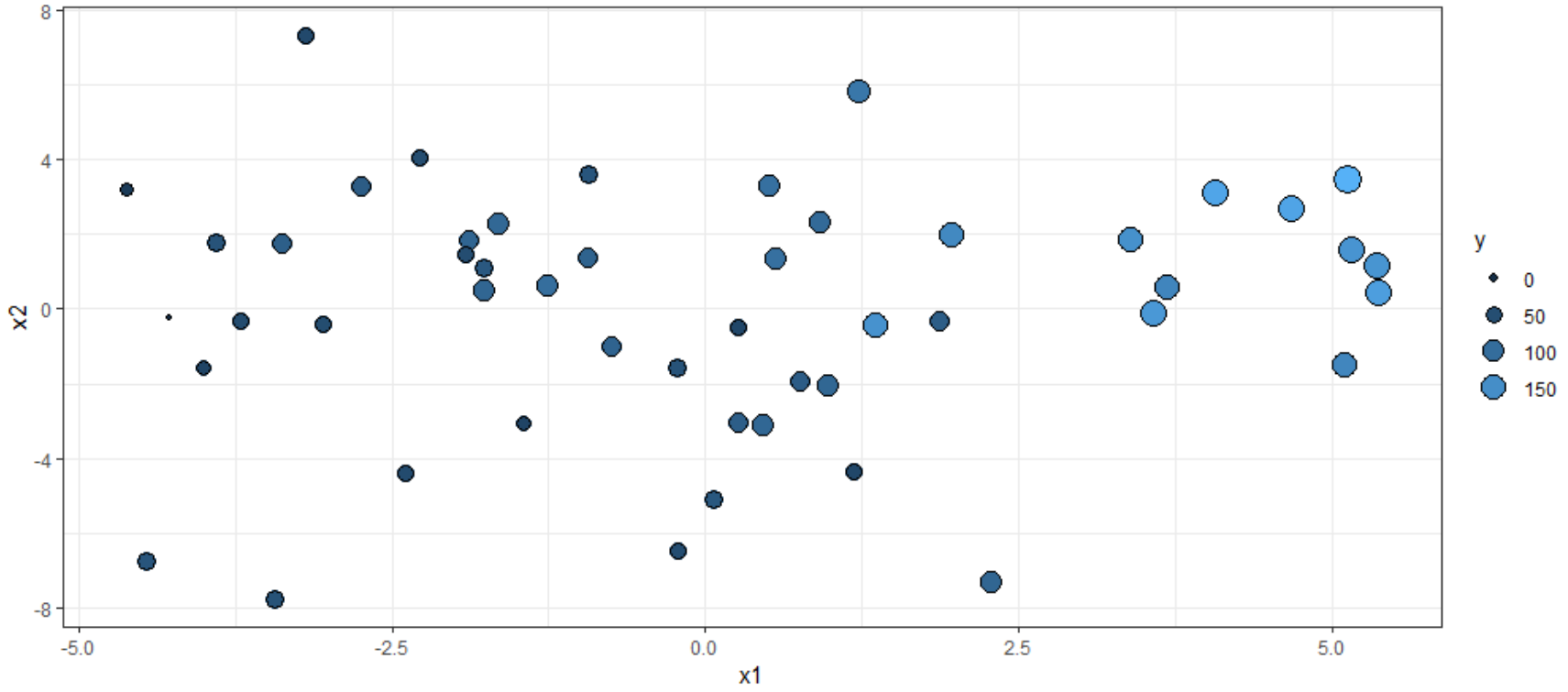
$$\begin{aligned} \mathrm{E}\{Y\} &= f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) \\ &= \beta_0 + \beta_{1,1} X_1 + \beta_{2,1} X_2 + \beta_{1,2} X_1^2 + \beta_3 X_1 X_2 + \beta_{2,2} X_2^2, \end{aligned}$$

même s'il peut y avoir de bonnes raisons de considérer un modèle comme

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_1 X_2.$$

Dans ce dernier cas, si $\beta_1 \beta_2 > 0$, alors on fait face à un **renforcement** lorsque $\beta_1 \beta_3 > 0$ et à une **interférence** lorsque $\beta_1 \beta_3 < 0$.

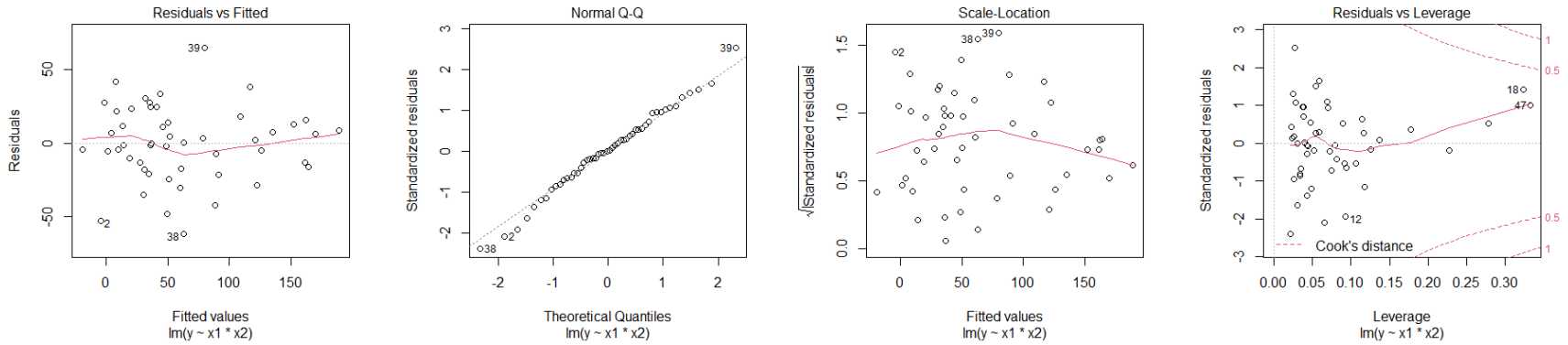
Exemple : considérons un ensemble de données avec $n = 50$ observations (2 prédicteurs centrés X_1, X_2 et une réponse Y, voir ci-dessous).



Nous calculons l'ajustement pour les modèles d'interaction réduit et complet. Le premier présente une interaction de renforcement ($\beta_1\beta_3 > 0$).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.7494     3.7043  16.669  < 2e-16 ***
x1           15.6463     1.3017  12.020 8.55e-16 ***
x2            5.1396     1.2010   4.279 9.40e-05 ***
x1:x2         1.6886     0.4379   3.856 0.000356 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.06 on 46 degrees of freedom
Multiple R-squared:  0.8166,Adjusted R-squared:  0.8047
F-statistic: 68.28 on 3 and 46 DF,  p-value: < 2.2e-16
```

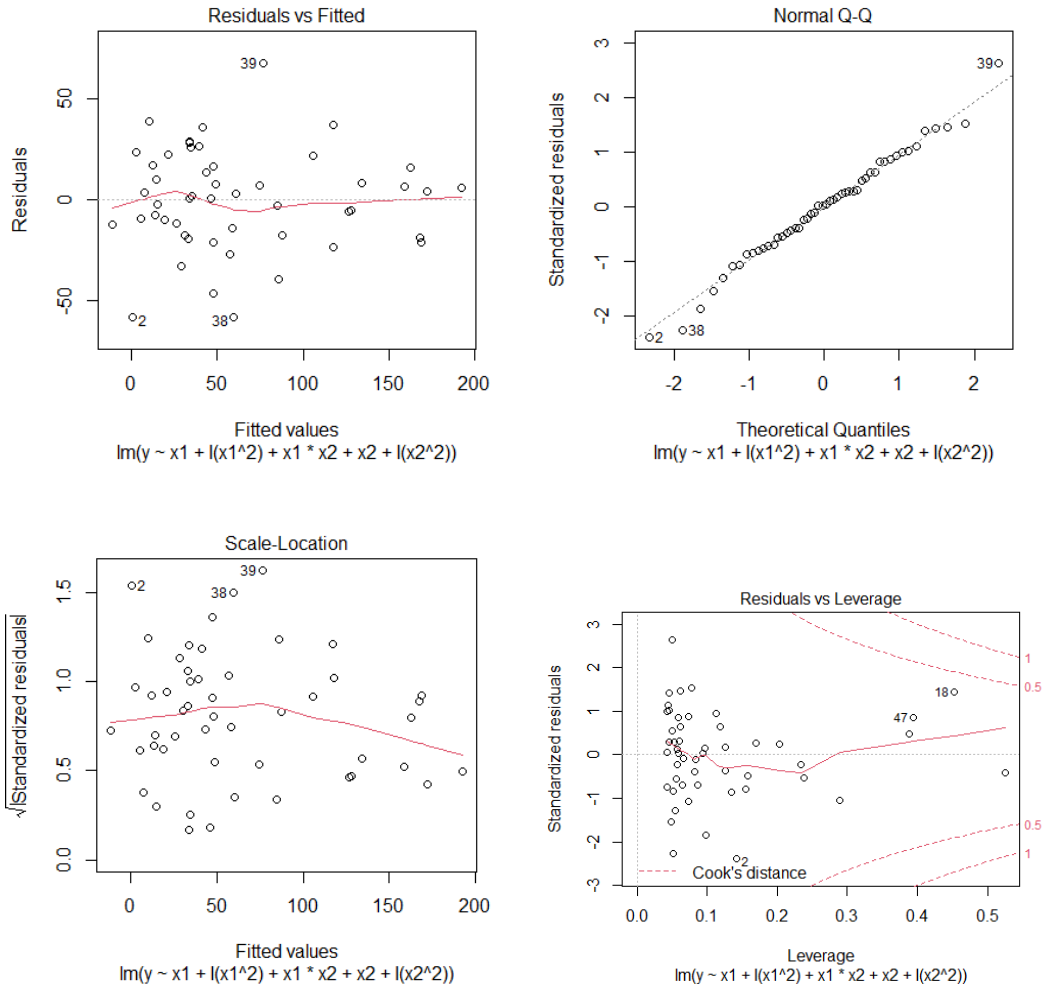
Le résumé indique que le modèle linéaire d'interaction réduite est **approprié**, ce qui est **soutenu** par les diagrammes de diagnostique :



Mais qu'en est-il du modèle complet ?

Les termes quadratiques purs sont **non significatifs**, ce qui suggère que le modèle réduit est **probablement** un meilleur choix (bien que cela ne soit **pas nécessairement** le cas).

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.25684    5.94511   9.799 1.24e-12 ***
x1          15.36026    1.38371  11.101 2.42e-14 ***
I(x1^2)      0.41459    0.46486   0.892 0.377316
x2           4.91100    1.31831   3.725 0.000553 ***
I(x2^2)      0.01042    0.26562   0.039 0.968891
x1:x2        1.56368    0.46519   3.361 0.001613 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.4 on 44 degrees of freedom
Multiple R-squared:  0.8199,Adjusted R-squared:  0.7994
F-statistic: 40.06 on 5 and 44 DF,  p-value: 2.654e-15
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Scale-Location
√|Standardized residuals|
Fitted values
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(y ~ x1 + I(x1^2) + x1 * x2 + x2 + I(x2^2))

4.4 – Modèles ANOVA pour les prédicteurs catégoriels

Nous pouvons inclure des variables catégorielles dans le cadre des moindres carrés. Supposons qu'il existe K "traitements" pour le prédicteur X.

1. Dans l'encodage de type **variable nominale**, nous définissons

$$X_j = \begin{cases} 1 & \text{traitement } j \\ 0 & \text{else} \end{cases}$$

pour $j = 1, \ldots, K-1$. Le modèle ANOVA devient ainsi

$$Y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j X_{i,j} + \varepsilon_i \quad \text{et} \quad E\{Y\} = \begin{cases} \beta_0 & \text{traitement } K \\ \beta_0 + \beta_j & \text{traitement } j \end{cases}$$

2. Dans l'encodage d'**effet du traitement**, nous définissons

$$X_j = \begin{cases} 1 & \text{treatment } j \\ -1 & \text{treatment } K \\ 0 & \text{else} \end{cases}$$

pour $j = 1, \ldots, K-1$. Le modèle ANOVA est comme dans le cas précédent, avec

$$E\{Y\} = \begin{cases} \beta_0 - (\beta_1 + \cdots + \beta_{K-1}) & \text{traitement } K \\ \beta_0 + \beta_j & \text{traitement } j \end{cases}$$

Nous étudierons des exemples concrets en R afin d'illustrer les grandes lignes.

4.5 – Moindres carrés pondérés

Le modèle de régression des moindres carrés $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ nécessite une **variance constante**. Lorsque cette hypothèse n'est pas satisfaite – de manière "monotone" : $\sigma^2\{\varepsilon_i\} = \sigma^2 x_i$, disons – on peut utiliser diverses transformations de données sur les prédicteurs X pour régler le problème.

Que peut-on faire lorsque l'hypothèse de linéarité est valide, mais que la variance σ_i ne change pas de manière **systématique** ?

On peut aborder le problème via les **moindres carrés pondérés** (WLS), qui n'exige pas que toutes les observations soient **traitées de la même manière** (c'est-à-dire qu'on ne leur donne pas toutes le **même poids**).

Soient $w_i \geq 0$ le **poids** de la $i$ème observation i et $Z_i = \sqrt{w_i}Y_i$, $i = 1, \ldots, n$.

La **matrice des poids** est $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$.

Le **problème du WLS** consiste à trouver le vecteur de coefficients $\boldsymbol{\beta}$ qui **minimise** la somme pondérée des erreurs quadratiques

$$\begin{aligned}
\mathrm{SSE}_w &= Q_w(\boldsymbol{\beta}) = \|\mathbf{Z} - \hat{\mathbf{Z}}\|_2^2 \\
&= \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\hat{\mathbf{Y}}\|_2^2 = \|\sqrt{\mathbf{W}}\mathbf{Y} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\beta}\|_2^2 \\
&= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Y}^\top \mathbf{W}\mathbf{Y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{Y} - \mathbf{Y}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

Mais $\nabla_{\boldsymbol{\beta}} Q_w(\boldsymbol{\beta}) = -2\mathbf{X}^\top \mathbf{W}\mathbf{Y} + 2\mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}$, de sorte que l'estimateur WLS $\mathbf{b}$ de $\boldsymbol{\beta}$ est

$$\nabla_{\boldsymbol{\beta}} Q_w(\boldsymbol{\beta}) = \mathbf{0} \implies \boldsymbol{b} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{Y}.$$

Toute la machinerie des moindres carrés peut être utilisés dans le contexte du WLS en remplaçant simplement $\mathbf{Y}$ par $\sqrt{\mathbf{W}}\mathbf{Y}$ et $\mathbf{X}$ par $\sqrt{\mathbf{W}}\mathbf{X}$ partout.

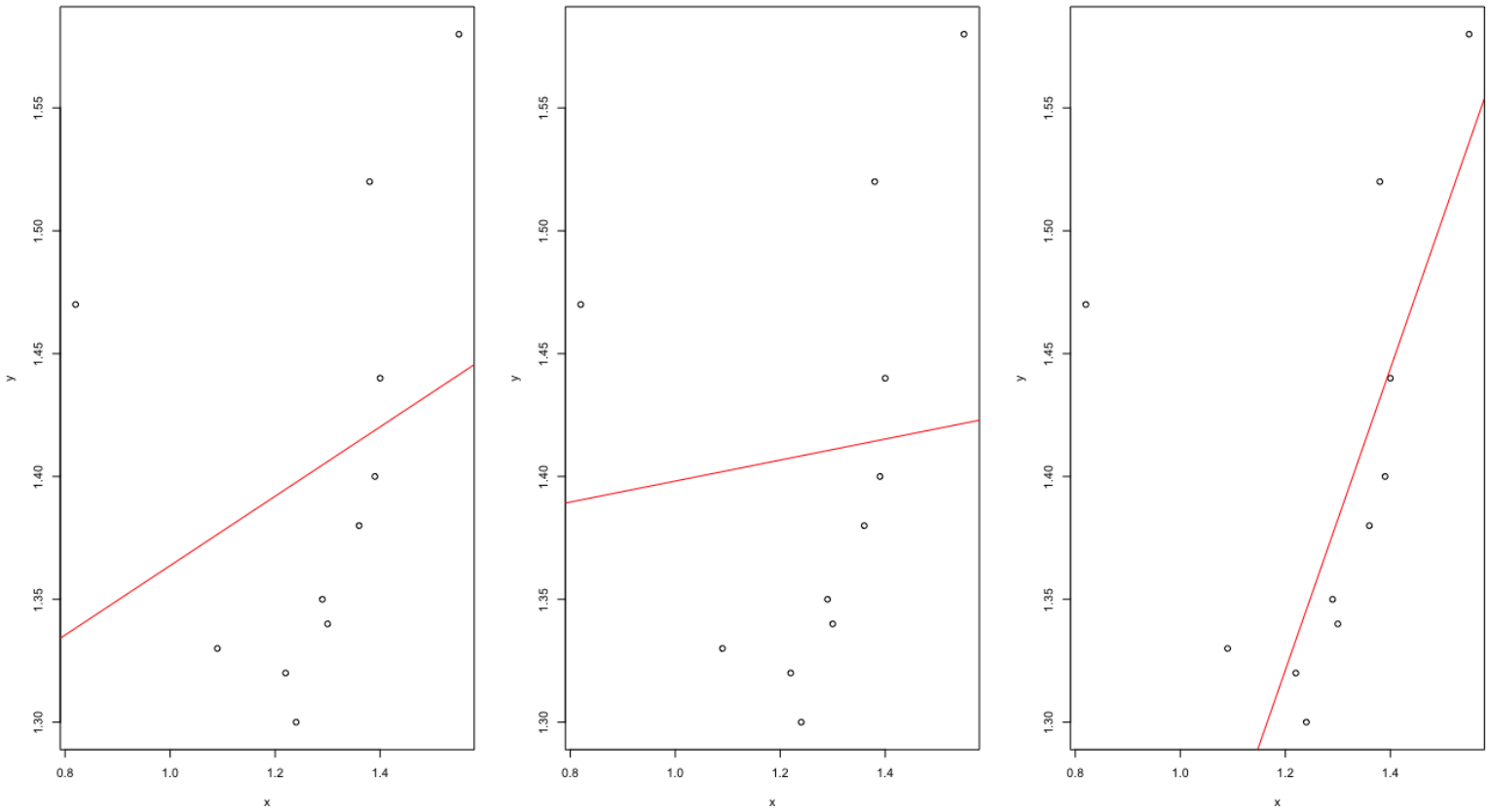
Exemple : considérons un ensemble de données avec $n = 11$ observations

i	1	2	3	4	5	6	7	8	9	10	11
x	0.82	1.09	1.22	1.24	1.29	1.30	1.36	1.38	1.39	1.40	1.55
y	1.47	1.33	1.32	1.30	1.35	1.34	1.38	1.52	1.40	1.44	1.58

Le modèle des moindres carrés est $\hat{y} = 1.223 + 0.1412x$ (à gauche); celui du WLS (avec $w_1 = 2$ et $w_i = 1$, $i = 2, \ldots, 11$) est

$$\hat{y} = 1.3553 + 0.0428x \quad \text{(au milieu);}$$

le modèle sans la première observation est $\hat{y} = 0.5848 + 0.6136x$ (à droite).


y
x
1.30
1.35
1.40
1.45
1.50
1.55
0.8
1.0
1.2
1.4

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.09759 -0.06036 -0.03454  0.06123  0.13864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2225     0.1920   6.366  0.00013 ***
x             0.1412     0.1489   0.948  0.36782
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09047 on 9 degrees of freedom
Multiple R-squared:  0.09081, Adjusted R-squared:  -0.01021
F-statistic: 0.899 on 1 and 9 DF,  p-value: 0.3678
```

```
Weighted Residuals:
     Min       1Q   Median       3Q      Max
-0.10841 -0.07148 -0.03354  0.06517  0.15833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3553     0.1624   8.344 1.58e-05 ***
x             0.0428     0.1292   0.331    0.748
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09669 on 9 degrees of freedom
Multiple R-squared:  0.01204, Adjusted R-squared:  -0.09773
F-statistic: 0.1097 on 1 and 9 DF,  p-value: 0.748
```

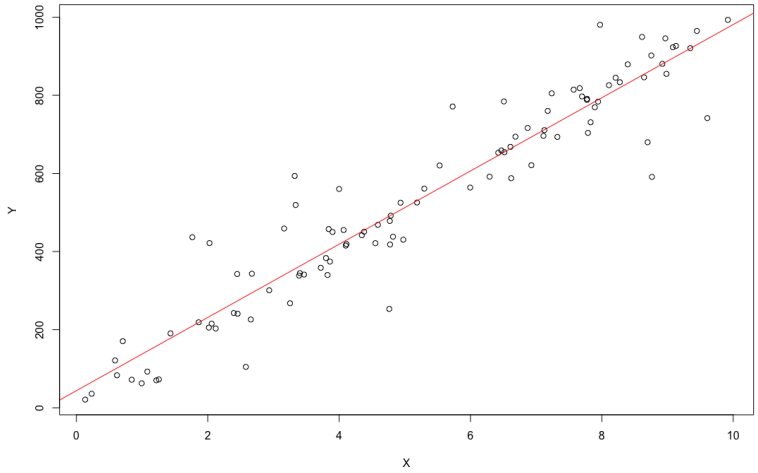
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.04568 -0.03852 -0.01341  0.02205  0.08841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5848     0.1916   3.052   0.0158 *
x             0.6136     0.1444   4.250   0.0028 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05402 on 8 degrees of freedom
Multiple R-squared:  0.693, Adjusted R-squared:  0.6546
F-statistic: 18.06 on 1 and 8 DF,  p-value: 0.002801
```

Comment le WLS peut-il nous aider avec une variance d'erreur non constante ?

Nous considérons le modèle sous-jacent

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\{\boldsymbol{\varepsilon}\}\right), \quad \text{où} \quad \sigma^2\{\varepsilon_i\} = \sigma_i^2 \not\equiv \sigma^2,$$

comme on peut le trouver dans l'image ci-dessous :



La procédure se déroule comme suit :

1. si les σ_i^2 sont connues, nous utilisons $w_i = \frac{1}{\sigma_i^2} \geq 0$ comme poids;

2. si les σ_i^2 sont inconnues,
   (a) on trouve les résidus e_i des moindres carrés ($e_i^2 \approx \sigma_i^2$ s'il n'y a pas de valeurs aberrantes en Y; autrement $|e_i| \approx \sigma_i$);
   (b) selon le choix effectué en (a), on ajuste soit e_i^2 ou $|e_i|$ par rapport à $X_1, \ldots, X_{p-1}$ pour obtenir les valeurs ajustées $\hat{v}_i$ ou $\hat{s}_i$, des estimations ponctuelles de σ_i^2 ou σ_i, respectivement ;
   (c) selon le choix effectué en (a), on utilise le WLS avec $w_i = \frac{1}{\hat{v}_i}$ ou $w_i = \frac{1}{\hat{s}_i^2}$, et on calcule SSE_w et $\mathrm{MSE}_w = \frac{\mathrm{SSE}_w}{n-p}$.
   si $\mathrm{MSE}_w \approx 1$, le choix des poids est **approprié** ; sinon, on répète les étapes (a) à (c) en utilisant les **résidus WLS** courants.

Exemple : on sait que le nombre d'articles défectueux Y produits par une machine est linéairement lié à la vitesse de réglage X de la machine :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon_i \text{ indép.}$$

Une analyste ajuste les $e_i^2 = (\hat{Y}_i - Y_i)^2$ par rapport à la vitesse X_i et obtient les $n = 12$ valeurs ajustées suivantes :

i	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{v}_i$	68.7	317.4	193	317.4	68.7	193	193	317.4	68.7	317.4	68.7	193

En utilisant le WLS avec $w_i = \frac{1}{\hat{v}_i}$, elle obtient les résidus $e_i^w = \hat{Y}_i^w - Y_i$:

i	1	2	3	4	5	6	7	8	9	10	11	12
e_i	−3.6	5.6	−13.5	−16.4	−9.6	7.5	−10.5	26.6	14.4	−17.4	−1.6	18.5

L'utilisation de cette pondérations est-elle appropriée ?

Solution: nous avons

$$\mathrm{SSE}_w = \sum_{i=1}^{12} w_i e_i^2 = \sum_{i=1}^{12} \frac{1}{\hat{v}_i} e_i^2 = 12.2953,$$

une somme de carrés avec $n - p = 12 - 2 = 10$ degrés de liberté, d'où

$$\mathrm{MSE}_w = \frac{\mathrm{SSE}_w}{n-p} = \frac{12.2953}{10} = 1.22953.$$

Comme $\mathrm{MSE}_w \approx 1$, les poids semblent **appropriés** et les $\hat{v}_i$ fournissent **des approximations raisonnables** de σ_i^2 pour $i = 1, \ldots, 12$.

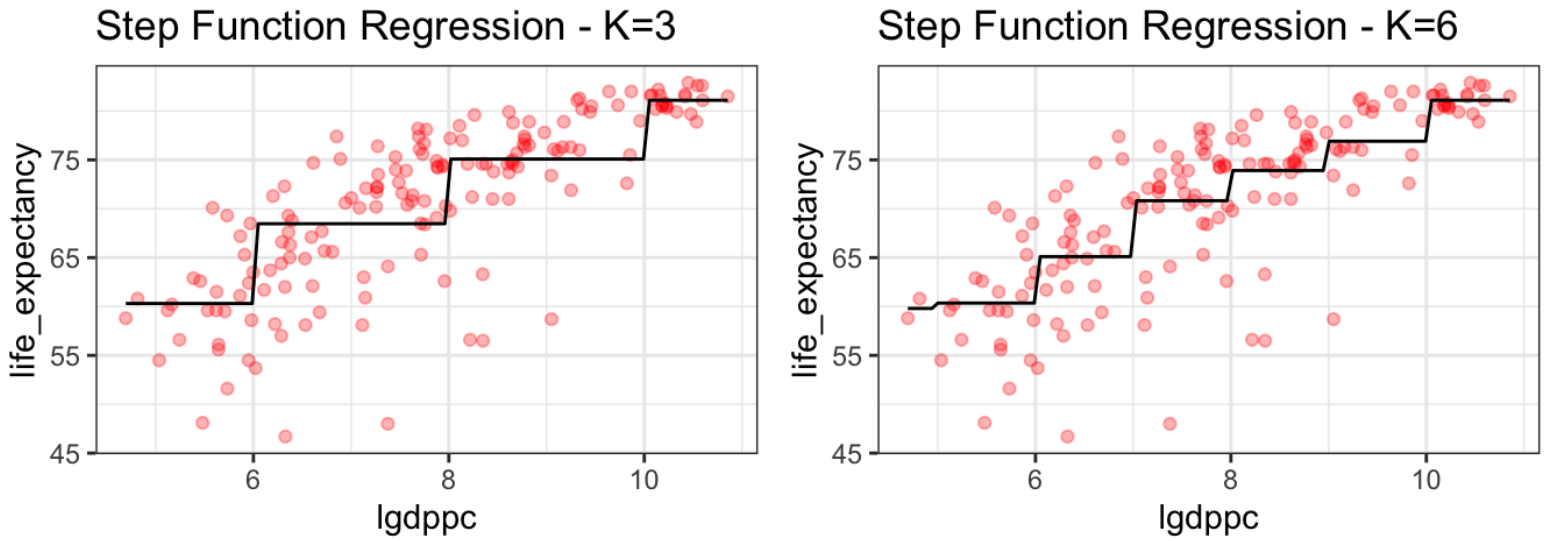
4.6 – Autres extensions

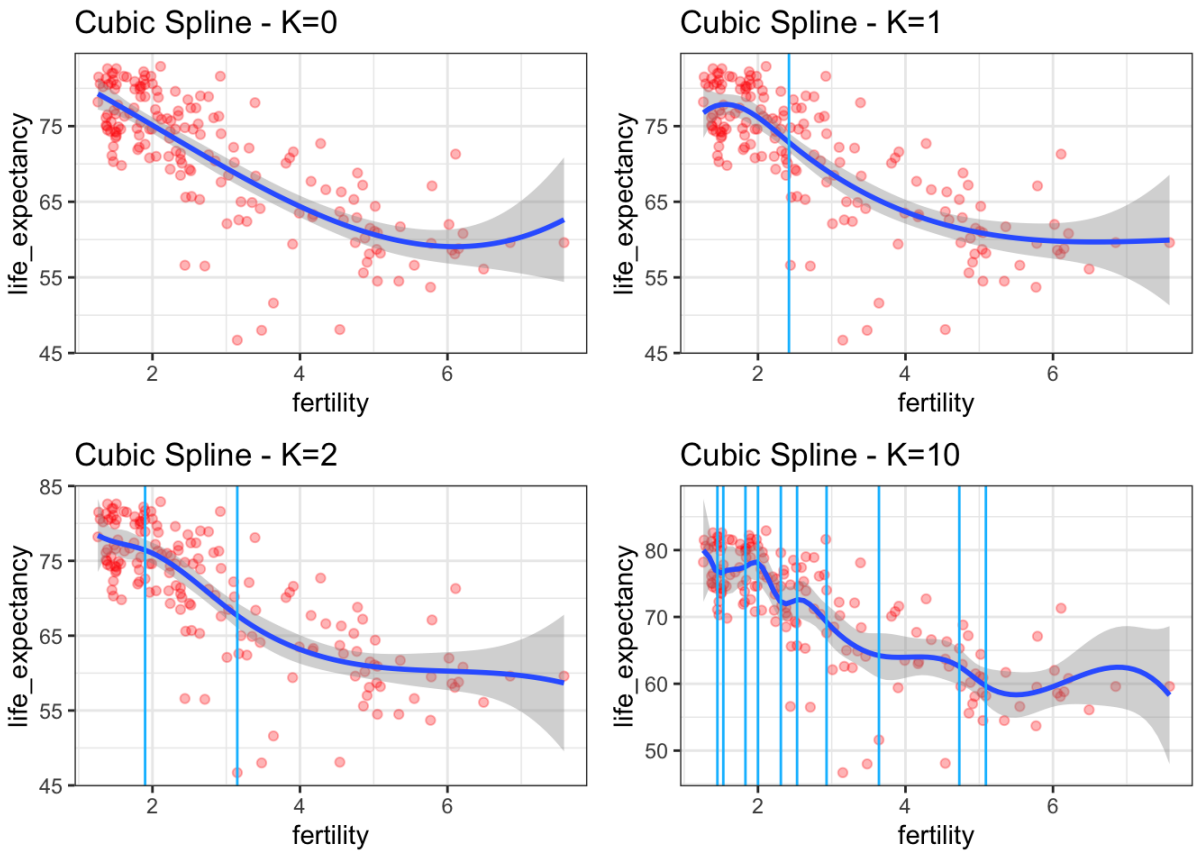
Les hypothèses des moindres carrés sont **commodes** d'un point de vue mathématique, mais elles ne sont pas toujours respectées en pratique ; nous avons vu que les transformations de données peuvent régler ce problème.

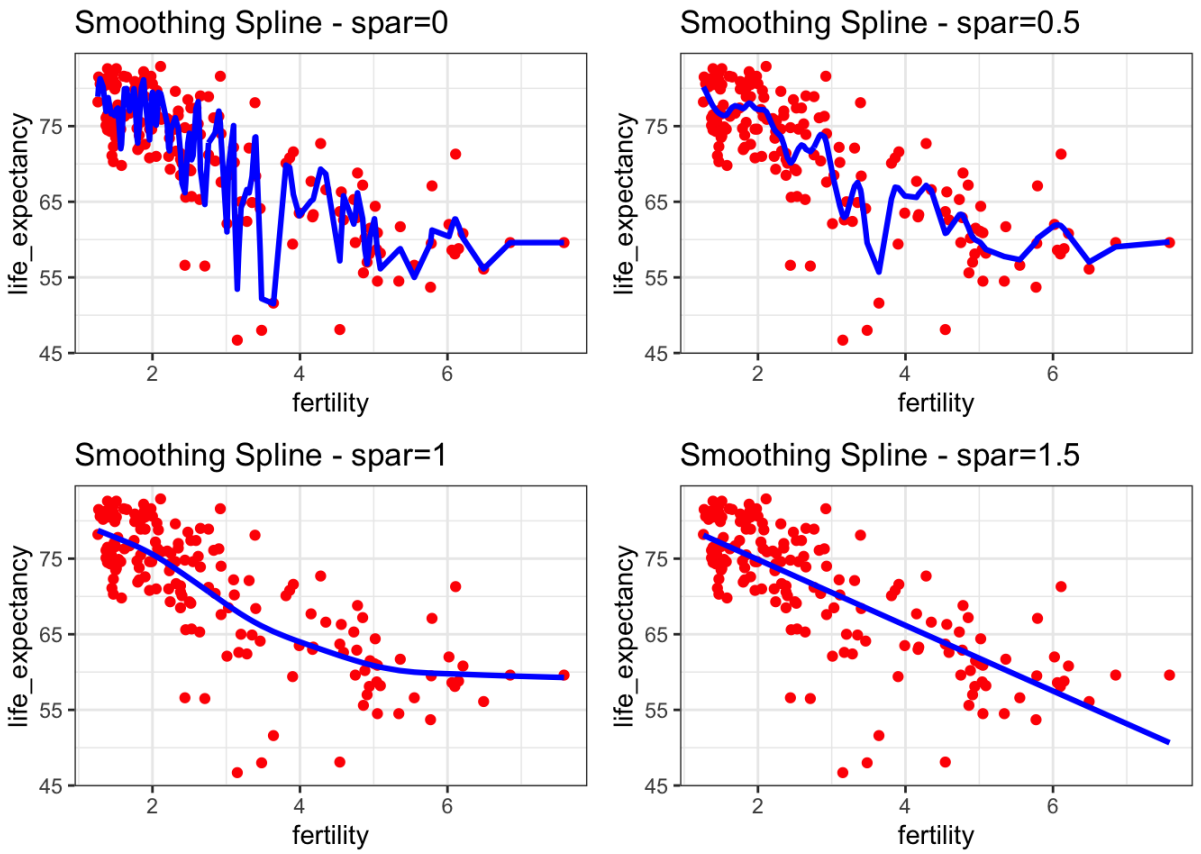
Une autre approche consiste à **étendre les hypothèses** et à élaborer le formalisme mathématique correspondant :

- les **modèles linéaires généralisés (GLM)** permettent des réponses suivant des lois conditionnelles **non-normales** (voir chapitre 7) ;
- la **classification**, tels que la régression logistique, les arbres de décision, les machines à vecteurs de support, la méthode naïve de Bayes, les réseaux neuronaux, etc., étend la régression aux **réponses catégorielles** ;

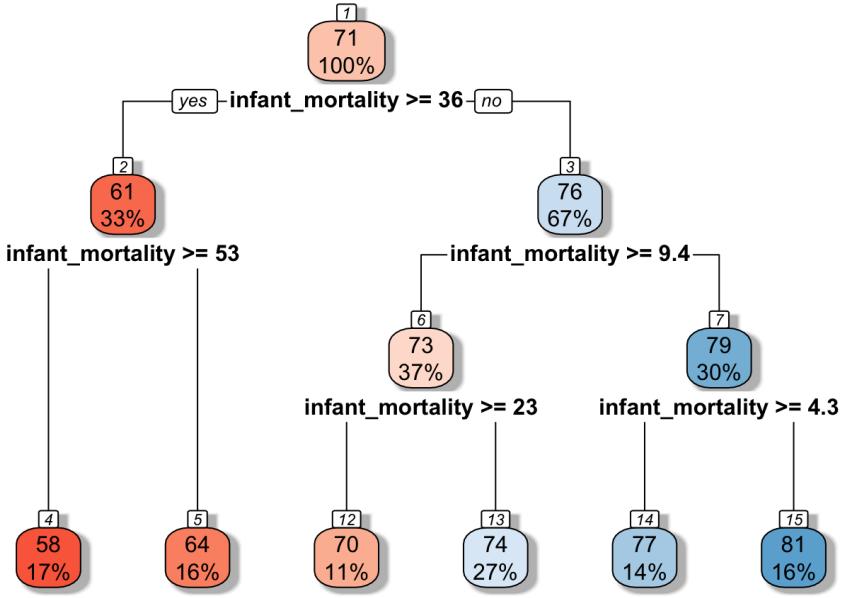
- les **méthodes non-linéaires**, telles que les splines, les modèles additifs généralisés (MAG), les méthodes du plus proche voisin, les méthodes de lissage à noyau, etc., sont utilisées pour les réponses qui ne sont **pas des combinaisons linéaires des prédicteurs** ;

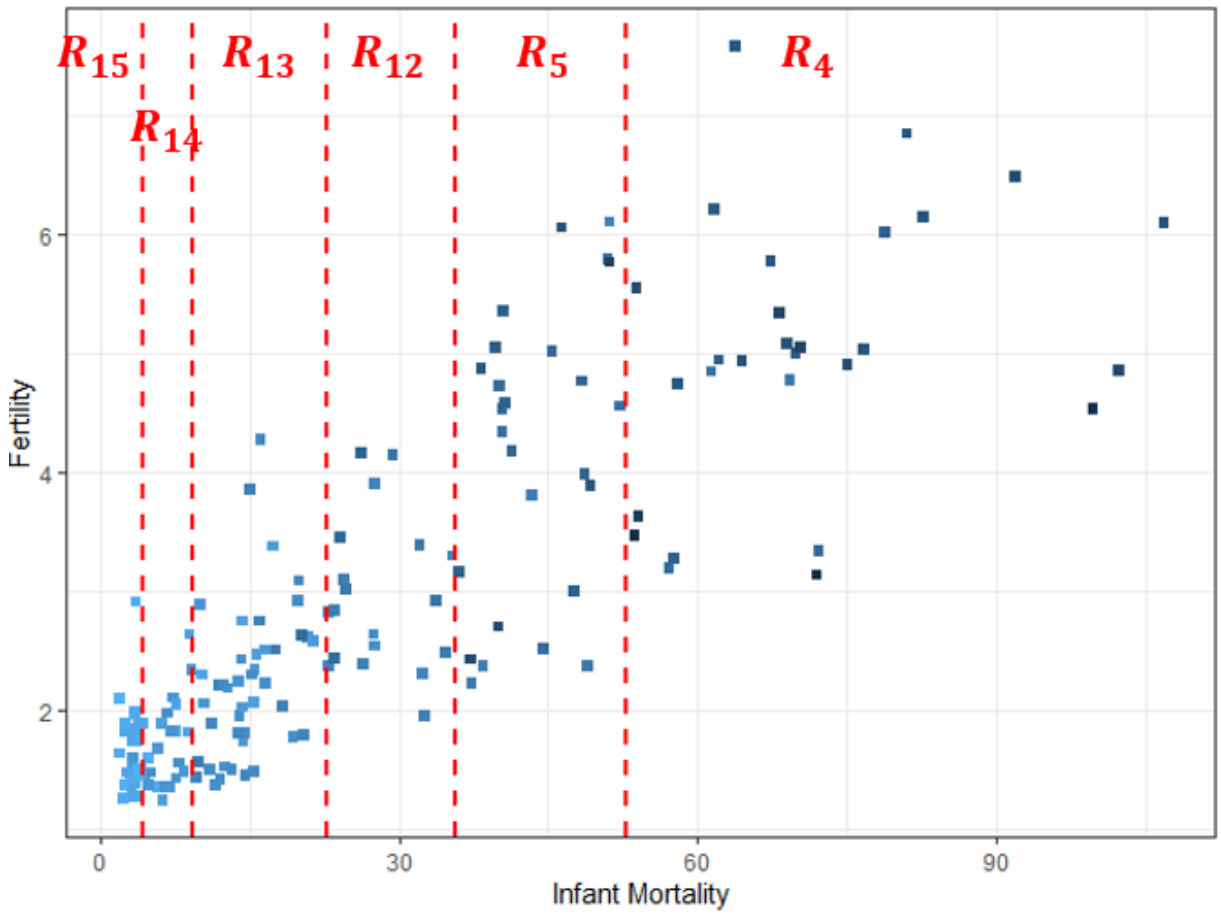



Cubic Spline - K=0
Cubic Spline - K=1
Cubic Spline - K=2
Cubic Spline - K=10
life_expectancy
fertility


Smoothing Spline - spar=0
Smoothing Spline - spar=0.5
Smoothing Spline - spar=1
Smoothing Spline - spar=1.5
life_expectancy
fertility
45
55
65
75
2
4
6

- les **méthodes basées sur les arbres** et **d'apprentissage par ensemble**, telles que le bagging, les forêts aléatoires et le boosting, sont utilisées pour simplifier la modélisation des **interactions entre prédicteurs** ;




R_{15}
R_{14}
R_{13}
R_{12}
R_5
R_4
Fertility
2
4
6
0
30
60
90
Infant Mortality

- les **méthodes de régularisation**, telles que la régression en crête, le LASSO, et les réseaux élastiques, facilitent le processus de **sélection de modèles** et de **sélection de caractéristiques**.

En ce qui concerne le dernier sujet, supposons que l'ensemble d'apprentissage se compose de n prédicteurs $\mathbf{x}_i$ **centrées** et **mises à l'échelle**, ainsi que de n réponses y_i.

Soit $b_{\mathrm{LS},j}$ le $j$ième coefficient des moindres carrés, et fixons un **seuil** $\lambda > 0$, dont la valeur dépend des données.

Nous avons vu que $\mathbf{b}_{\mathrm{LS}}$ est la solution exacte du problème

$$\mathbf{b}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\} = \arg\min_{\boldsymbol{\beta}} \{\mathrm{SSE}\}.$$

En général, nous n'assumons **aucune restriction** sur les valeurs $b_{\text{LS},j}$; de grandes magnitudes impliquent que les caractéristiques correspondantes **jouent un rôle important** dans la prédiction de la réponse.

La **régression en crête** (RR) est une méthode qui **régularise** les coefficients $b_{\text{LS},j}$.

Elle réduit ces derniers en **pénalisant** les solutions de magnitude élevée – si la magnitude d'un coefficient spécifique demeure **élevée**, alors il doit réllement avoir une **forte** pertinence dans la prédiction de la réponse.

Cela conduit à un problème des moindres carrés modifié :

$$\mathbf{b}_{\text{RR}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\text{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_2^2}_{\text{pénalité}} \}.$$

La quantité à minimiser est petite lorsque SSE est **faible** (c'est-à-dire que le modèle s'ajuste bien aux données) et lorsque la **pénalité** est petite (c'est-à-dire lorsque chaque β_j est petit) ; les solutions RR sont généralement obtenues via des méthodes numériques.

L'hyperparamètre λ contrôle l'**impact relatif** des deux composantes. Il y a des variantes, telles que la **meilleure régression par sous-ensemble** (BS) et le **LASSO**, qui tendent toutes deux à donner $\beta_j = 0$ pour certains j :

$$\mathbf{b}_{\mathrm{BS}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\mathrm{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_0}_{\text{pénalité}} \}, \quad \|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} \mathrm{sgn}(|\beta_j|)$$

$$\mathbf{b}_{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}} \{ \underbrace{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\mathrm{SSE}} + \underbrace{\lambda n \|\boldsymbol{\beta}\|_1}_{\text{pénalité}} \}, \quad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|.$$