Diagnostics dans le modèle linéaire

Agathe Guilloux

Modèle linéaire

Ecriture matricielle

Pour un individu i, on a

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \ldots + \beta_p X_i^p + \varepsilon_i.$$

On peut récrire

$$Y_i = (1, X_i^1, \ldots, X_i^p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \varepsilon$$

ou bien, pour tous les individus

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & \ldots & X_1^p \\ 1 & X_2^1 & X_2^2 & \ldots & X_2^p \\ \ldots & & & & \\ 1 & X_n^1 & X_n^2 & \ldots & X_n^p \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix}$$

$$\begin{array}{ccccccc} Y & = & X & & \beta & + & \varepsilon. \\ n \times 1 & & n \times (p+1) & & (p+1) \times 1 & + & n \times 1 \end{array}$$

Définition complète

Modèle linéaire : définition et hypothèses

$$Y = X\beta + \epsilon$$

où

- Y est un vecteur $n \times 1$ **observé**
- X est une matrice $n \times (p+1)$ **observée** de **rang** $p+1$
- β est un vecteur $(p+1) \times 1$ de paramètres **inconnus**
- ϵ est un vecteur $n \times 1$ de v.a. **non-observées** supposées **décorrélées** avec

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}(\epsilon_i) = \sigma^2$$

où σ^2 est un paramètre **inconnu**.

L'estimateur des moindres carrés

Dans le modèle linéaire

$$Y = X\beta + \varepsilon,$$

on définit $\hat{Y} = X\hat{\beta}$ comme le **projeté orthogonal de** Y **sur** vect(X)

- Puisque $\hat{Y} = X\hat{\beta}$ est la projection de Y sur vect(X), on a

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

- si H est la matrice de projection dans vect(X)

$$X\hat{\beta} = X(X^\top X)^{-1} X^\top Y = HY$$

- on a

$$\mathbb{V}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$$

- et $\hat{\sigma}^2 = \frac{\|e\|^2}{n-(p+1)}$.

R2 et R2 ajusté

On obtient des mesures de l'adéquation du modèle

- $Y - X\hat{\beta} \perp X\hat{\beta} - \bar{Y}\mathbf{1}$ si $(1, \ldots, 1) \in \text{vect}(X)$ et donc

$$\underbrace{\|Y - \bar{Y}\mathbf{1}\|^2}_{\substack{\text{SC tot.} \\ SSTotal}} = \underbrace{\|Y - X\hat{\beta}\|^2}_{\substack{\text{SC résiduelle} \\ SSError}} + \underbrace{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2}_{\substack{\text{SC expliquée} \\ SSModel}}.$$

- On définit le R^2 par

$$0 \leq R^2 = \frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} = 1 - \frac{\|Y - X\hat{\beta}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} \leq 1$$

et le R^2 ajusté du nombre de paramètres par

$$R^2_{Adj} = 1 - \frac{(n-1)(1-R^2)}{(n-p-1)} \leq 1$$

Attention à la dimension $p+1$: c'est le nombre de variables explicatives p + 1 pour le coefficient constant (associé à $(1, \ldots, 1)$).

Modèle linéaire gaussien

Définition : Modèle linéaire gaussien

$$Y = X\beta + \epsilon$$

où

$$\epsilon \sim \mathcal{N}((0, \ldots, 0)^\top, \sigma^2 I_n).$$

Conséquences du Théorème de Cochran

1. sur $\hat{\beta}$ (projection de Y sur vect(X))

$$\hat{\beta} - \beta \sim \mathcal{N}((0, \ldots, 0), \sigma^2 (X^\top X)^{-1})$$

2. sur $\hat{\sigma}$ (projection de Y sur vect$(X)^\perp$)

$$\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2} = \frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

3. $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ donc

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim \mathcal{T}(n-p-1) \text{ où } \hat{\sigma}_j^2 = \hat{\sigma}^2 (X^\top X)^{-1}_{jj}$$

4 si $\mathbf{1} \in \text{vect}(X)$, on peut écrire

$$\mathbb{R}^n = \text{vect}(X)^{\perp} \overset{\perp}{\bigoplus} \left(\text{vect}(\mathbf{1})^{\perp_{\text{vect}(X)}} \right) \overset{\perp}{\bigoplus} \text{vect}(\mathbf{1})$$

On a alors

$$\underbrace{Y - X\hat{\beta}}_{\substack{\in \text{vect}(X)^{\perp} \\ \text{de dim. } n-p-1}} \quad \perp\!\!\!\perp \quad \underbrace{X\hat{\beta} - \bar{Y}\mathbf{1}}_{\substack{\in \text{vect}(\mathbf{1})^{\perp_{\text{vect}(X)}} \\ \text{de dim. } p}} .$$

donc :

$$\frac{\|X\hat{\beta} - \bar{Y}\mathbf{1}\|^2/p}{\|Y - X\hat{\beta}\|^2/(n-p-1)} \sim \mathcal{F}(p, n-p-1).$$

Cette statistique permet de tester si le modèle avec les covariables apporte significativement plus d'information sur la réponse Y que le modèle avec l'intercept $\mathbf{1}$ seulement.

Erreur d'estimation de $X_i\beta$

Pour un individu i ($i = 1, \ldots, n$), la valeur Y_i (observé) est estimée par

$$\hat{Y}_i = X_i\hat{\beta}.$$

On a

$$\mathbb{E}\,\hat{Y}_i = X_i\beta \text{ et}$$
$$\mathbb{V}(\hat{Y}_i) = X_i\mathbb{V}(\hat{\beta})X_i^\top = \sigma^2 X_i(X^\top X)^{-1}X_i^\top.$$

Intervalle de confiance pour $X_i\beta$

$$\frac{\hat{Y}_i - X_i\beta}{\sqrt{\hat{\sigma}^2 X_i(X^\top X)^{-1}X_i^\top}} \sim \mathcal{T}(n-p-1).$$

Erreur de prévision de Y

Si on considère un nouvel individu indépendant de $1, \ldots, n$ pour lequel on connaît X_+ (mais pas Y_+), on peut prédire la valeur de $Y_+ = X_+\beta + \epsilon_+$ par

$$Y_+^p = X_+\hat{\beta},$$

l'erreur commise est alors donné par :

$$Y_+^p - Y_+ = X_+\hat{\beta} - (X_+\beta + \epsilon_+) = X_+(X^\top X)^{-1}X\epsilon - \epsilon_+.$$

Intervalle de prévision pour Y_k

$$\frac{Y_+^p - Y_+}{\sqrt{\hat{\sigma}^2(X_+(X^\top X)^{-1}X_+^\top + 1)}} \sim \mathcal{T}(n-p-1).$$

On doit vérifier les hypothèses du modèle, i.e.

- les hypothèses sur les erreurs
- la présence d'individus "influents"
- les hypothèses sur X (de plein rang)
- l'hypothèse de linéarité...

Résidus

Analyse des résidus

On veut vérifier les hypothèses sur les erreurs ϵ, i.e.

- indépendantes (ou décorrélées) avec

$$\mathbb{E}(\epsilon_i) = 0 \text{ et } \mathbb{V}\text{ar}(\epsilon_i) = \sigma^2$$

- voire gaussiennes

Test de normalité

On suppose

$$\epsilon \sim \mathcal{N}((0,\dots,0)^\top, \sigma^2 I_n).$$

- Si les ϵ_i $(i = 1,\dots,n)$ étaient observables, on pourrait tracer un QQ-plot des ϵ_i/σ contre les quantiles de la $\mathcal{N}(0,1)$.
- On n'observe que les erreurs résiduelles e_i $(i = 1,\dots,n)$, qu'on prend comme estimateurs des ϵ_i.
- On a

$$e = Y - X\hat{\beta} = (I_n - H)Y = (I_n - H)X\beta + (I_n - H)\epsilon = (I_n - H)\epsilon.$$

- Les erreurs résiduelles sont gaussiennes avec

$$\mathbb{E}(e) = 0 \text{ et, } \mathbb{V}(e) = \sigma^2(1 - H) \text{ et } \mathbb{V}(e_i) = \sigma^2(1 - H_{ii}).$$

- Les erreurs résiduelles ne sont donc pas homoscédastiques. On définit les **résidus standardisés** par :

$$e_i' = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}}.$$

- On leur préfère **les résidus studentisés**

$$e_i^* = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - H_{ii})}},$$

où $\hat{\sigma}^2_{(i)}$ est l'estimateur de σ^2 calculé en enlevant l'observation i.

Loi des résidus studentisés

On montre que

$$e_i^* \sim \mathcal{T}(n-p-1)$$

et

$$e_i^* = \frac{e_i'}{\sqrt{\frac{n-p-(e_i')^2}{n-p-1}}}$$

On conseille de faire le QQ-plot sur ces résidus (si $n-p-1$ est grand, on peut le faire avec les quantiles gaussiens)

Valeurs ajustées $\hat{y}$

```
yhat = fit$fitted.values
```

Résidus $e = y - \hat{y}$

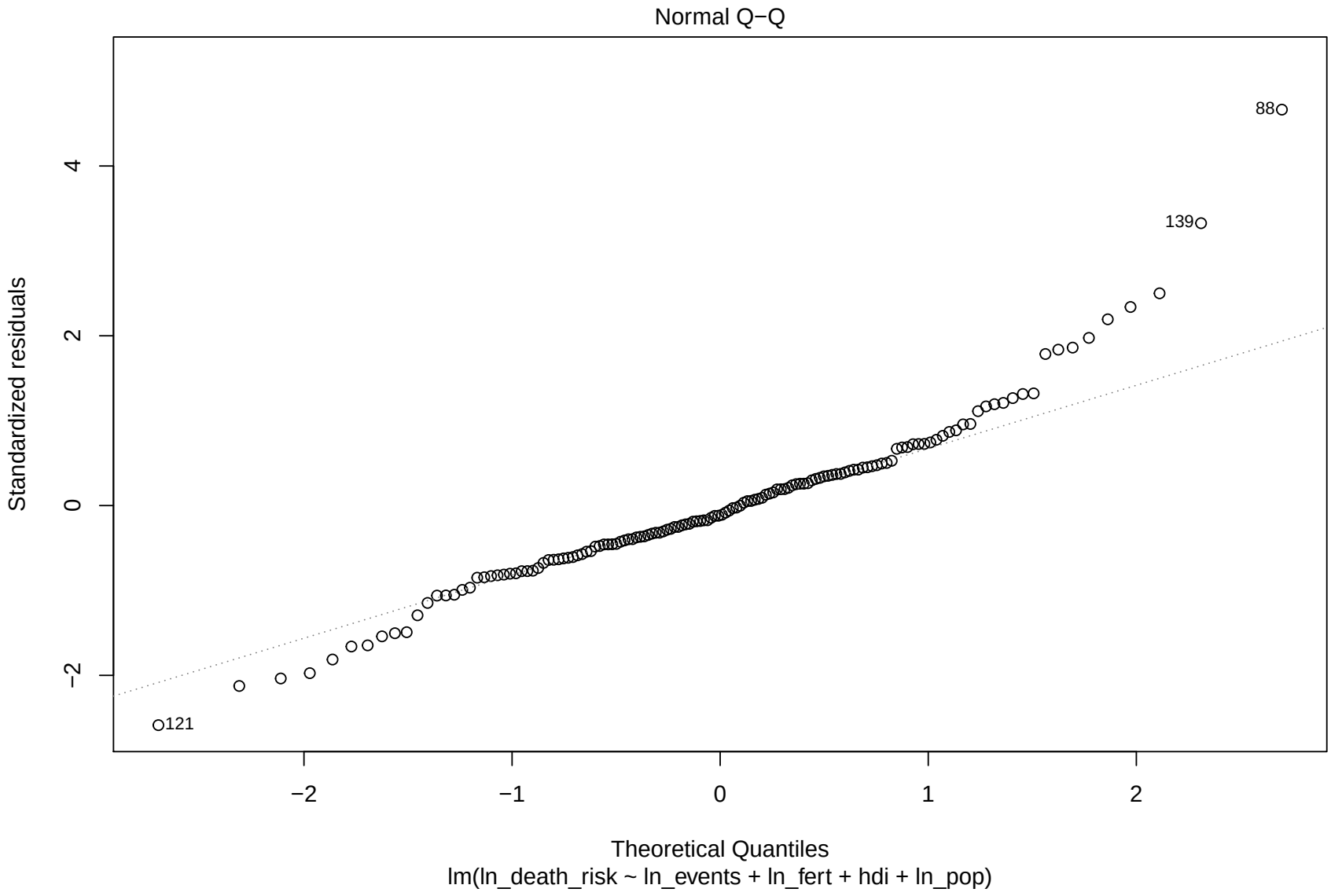
```
e = fit$residuals
```

Residus studentisés $e^{\star}$

```
e_star = rstudent(fit)
```

Normalité des erreurs

```
plot(fit,which=2)
```



Observations atypiques

Différentes observations atypiques

On cherche maintenant des mesures de l'influence des observations dans l'estimation.

- Une "enquête" sur les observations/les individus "trop influent(e)s" devra être faite, pour déterminer notamment s'il n'y a pas eu d'erreur de mesure, de relevé, etc.
- Le rôle du statisticien est de les détecter.

On peut distinguer deux types d'observations atypiques :

- celles qui ont un "trop" grand résidu
- celles qui sont trop isolées.

Observation aberrante

On connaît la loi des résidus studentisés e_i^*

$$e_i^* \sim \mathcal{T}(n-p-1).$$

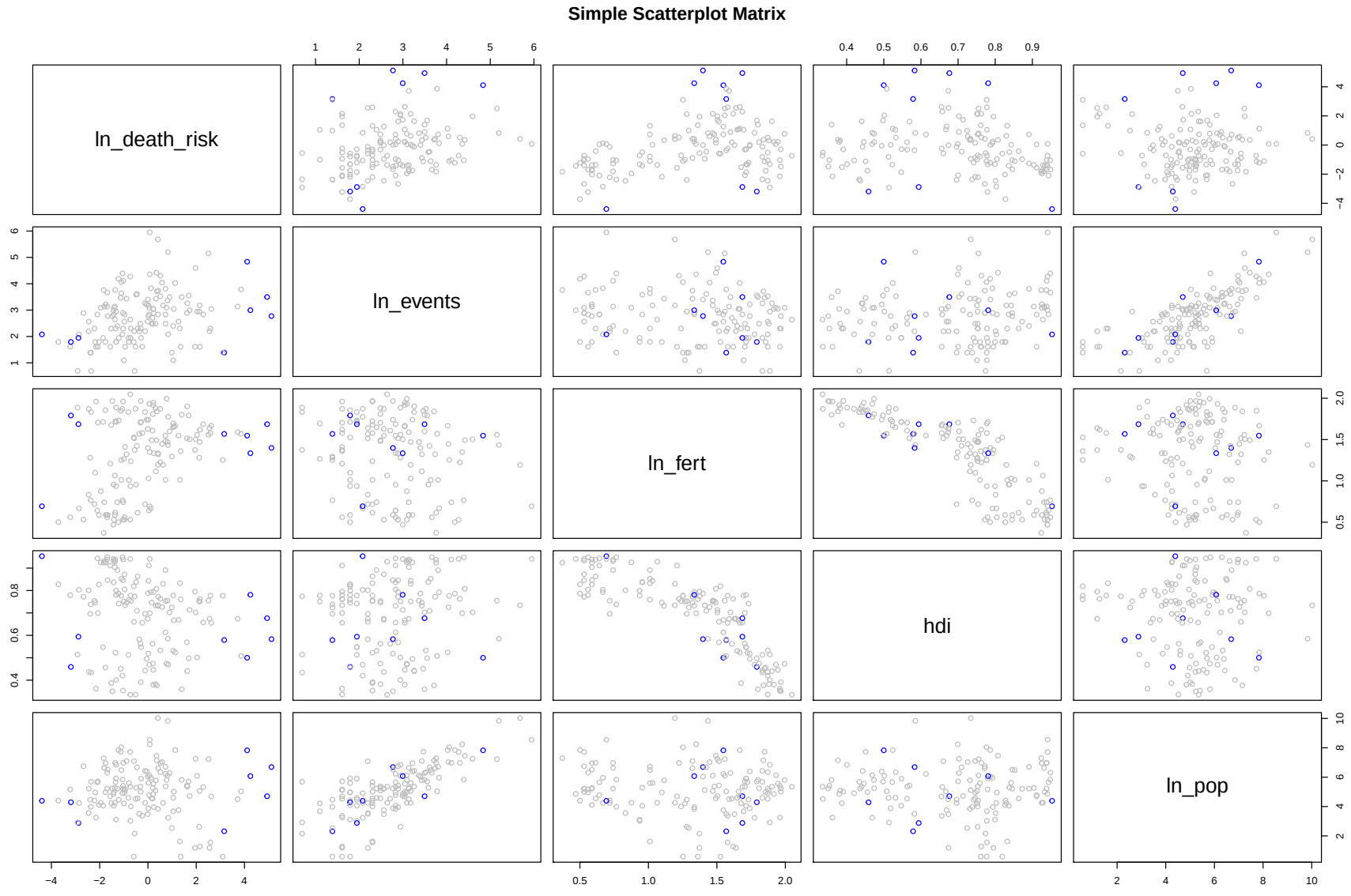
Régle

on dit qu'une observation est **aberrante** si

$$|e_i^*| > F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha).$$

On choisit souvent α de l'ordre de $1/n$ ou $F^{-1}_{\mathcal{T}(n-p-1)}(1-\alpha) = 2$.

Où sont les points “aberrants” ?



Le levier

Une bonne mesure de l'isolement des observations est le **coefficient** H_{ii} **appelé "levier"** ("leverage") .

On sait que $0 \leq H_{ii} \leq 1$ et $0 \leq H_{ii} = H_{ii}^2 + \sum_{k \neq i} H_{ik}^2 \leq 1$

Propriétés des leviers

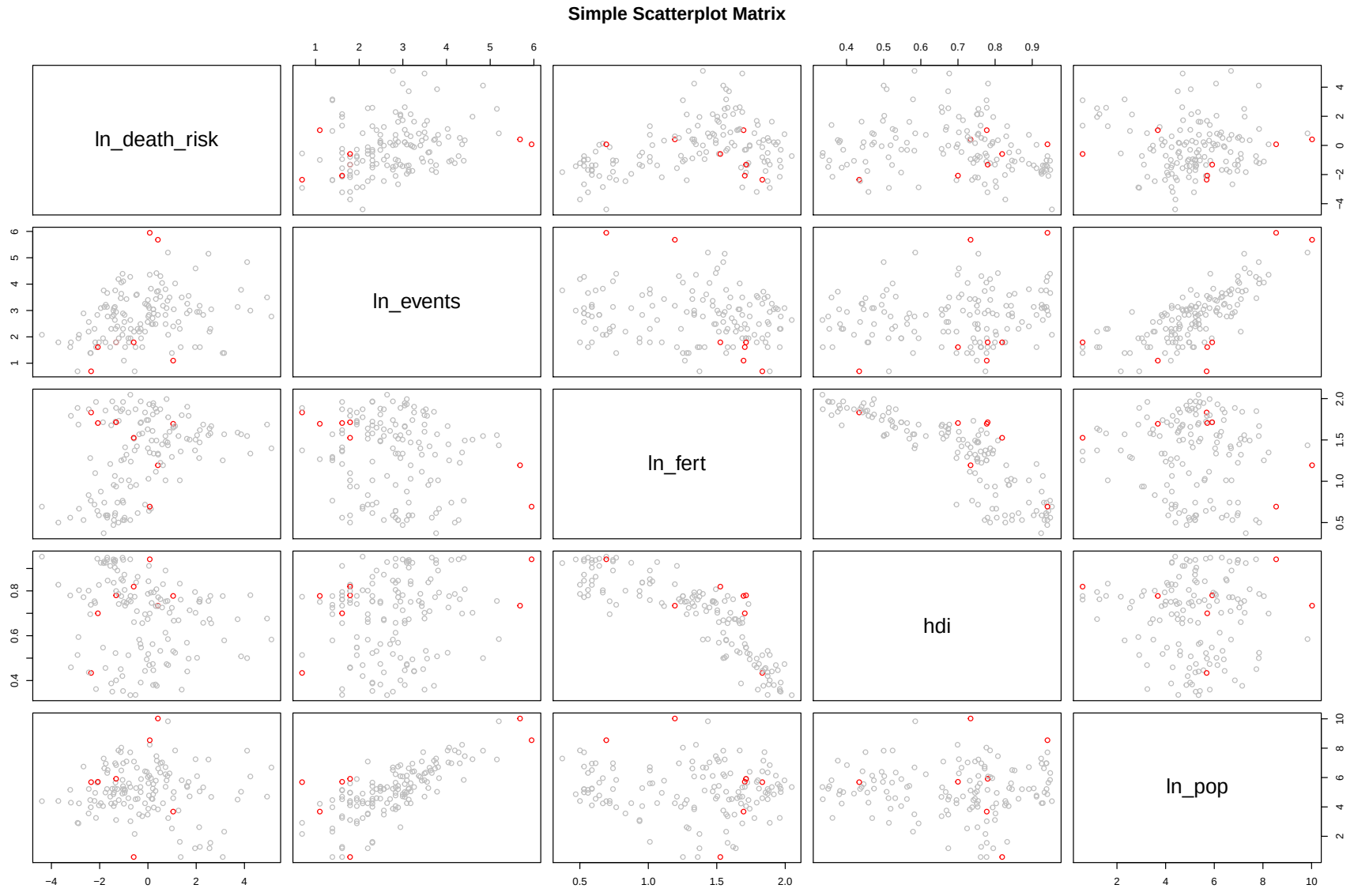
On montre que :

- $H_{ii} = 1$ ssi vect$(X_j, j \neq i)$ est de dimension $p + 1 - 1$.
- $e_i = Y_i - \sum_k H_{ik} Y_k = (1 - H_{ii}) Y_i - \sum_{k \neq i} H_{ik} Y_k$

Règle pour les leviers

On sait aussi que $\sum_i H_{ii} = p + 1$, on considère donc qu'une observation est **isolée** quand a un levier sup. à $2p/n$ (ou $(2p + 2)/n$ ou $3p/n$).

```
influences =  lm.influence(fit)
leviers =   influences$hat
```


Simple Scatterplot Matrix
ln_death_risk
ln_events
ln_fert
hdi
ln_pop

Distance de Cook

La **distance de Cook** est une mesure résumée :

$$\text{DCOOK}_i = \frac{(e_i)^2 H_{ii}}{(p+1)(1-H_{ii})^2} > 4/n \text{ ou } 1 \Longrightarrow \text{influence}$$

Diagnostics sur X

Rang de la matrice X

- On veut vérifier l'hypothèse que X est de plein rang, i.e. que les $p+1$ colonnes de X engendrent un s.e.v. de $\mathbb{R}^n$ de dimension $p+1$.
- Si ce n'est pas le cas, la matrice $X^\top X$ n'est pas inversible, il n'y a donc pas de solution unique à l'équation

$$X^\top Y = X^\top X\hat{\beta}.$$

- On veut donc vérifier qu'il n'y pas de colinéarité entre les colonnes $\mathbf{1}, X^1, \ldots, X^p$ de X.

Valeurs propres de la matrice de corrélation

On définit la matrice R des corrélations empiriques entre les variables X^j, $j = 1, \ldots, p$:

$$R_{jj'} = \frac{\sum_{i=1}^{n}(X_i^j - \bar{X}^j)(X_i^{j'} - \bar{X}^{j'})}{\sqrt{\sum_{i=1}^{n}(X_i^j - \bar{X}^j)^2 \sum_{i=1}^{n}(X_i^{j'} - \bar{X}^{j'})^2}} = \mathbb{C}\text{orr}(X^j, X^{j'}).$$

- C'est une matrice symétrique positive de rang $= \dim(\text{vect}(X)) \leq p$ ($< p$ si il y a colinéarité).
- On calcule les p valeurs propres $\lambda_1 \geq \ldots \geq \lambda_p$ de cette matrice.
  - S'il y a une relation linéaire parfaite entre des X^j, une des valeurs propres vaut 0.

Règle

On définit l'indice de conditionnement $\kappa = \lambda_1/\lambda_p$ et la règle

$$\kappa > 100 \Longrightarrow \text{ colinéarité trop forte}$$

Si on veut une étude plus fine, il faut étudier les vecteurs propres associées aux trop petites valeurs propres.

Matrice de correlations

Definition de la matrice

```
X = vul[,c(3:6)]
cor_mat = cor(X)
```

Calcul des valeurs propres et vecteurs propres

```
propres = eigen(cor_mat)
propres$values[1] / propres$values
```

```
## [1]  1.000000  1.248879  7.351330 14.939503
```

Règle à suivre pour les problèmes de colinéarité

- Si on détecte un problème de colinéarité, il faut enlever les variables posant problème **une à une**.
- Le choix des variables devrait se faire avec ceux qui ont fourni le jeu de données.

A cette étape, on doit avoir un jeu de données propre pour le modèle linéaire :

- relations linéaires entre variables explicatives et variable à expliquer
- matrice X de plein rang
- résidus normaux
- pas d'observation aberrante ou trop influente

Il reste à sélectionner un modèle et à l'interpréter !