




Regression avancée
Chapitre 1 : introduction

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

Avant de commencer

- ▶ Les documents du cours sont disponibles ici : <http://www.math-evry.cnrs.fr/members/aguilloux/enseignements/m2upmc>
- ▶ Bibliographie (pour ce chapitre) :
 -  Pierre-André Cornillon and Eric Matzner-Løber. “La régression linéaire simple”. In: *Régression avec R* (2011), pp. 1–28.
 -  Julian J Faraway. *Practical regression and ANOVA using R*. 2002.
 -  Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.
- ▶ Contrôle des connaissances : 2 TP rendus, 1 projet (type concours de DataScience), 1 examen court.
- ▶ Pré-requis : cours de Statistique de base http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf.

Outline

Modèle linéaire

Modèle linéaire gaussien

Diagnostics sur X

Rang de la matrice X

Analyse des résidus

Dans le modèle linéaire gaussien

Dans le modèle linéaire

Influence des observations

Différentes observations atypiques

Autres mesures d'influence

Autres problèmes

Relation non-linéaire

Problèmes d'interprétation

Données “Vulnerability”, Patt et al., PNAS (2009)

Les pays les moins développés sont-ils plus vulnérables aux changements climatiques ?

Les auteurs ont voulu expliquer \ln_death_risk , log du risque mortel dû aux évènements climatiques en fonction

- ▶ du log du nombre d'évènements climatiques \ln_death_risk
- ▶ du log de la fertilité $\ln_fertility$
- ▶ de l'indice de développement humain hdi (United Nations)
- ▶ du log de la population \ln_pop

Ils concluent que le développement socio-économique a un lien sur la fragilité aux événements climatiques, et ce lien pourrait se révéler dans le deuxième quart du 21^{ème} siècle.

Modèle linéaire

Ecriture matricielle

Pour un individu i , on a

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

On peut récrire

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

ou bien, pour tous les individus

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

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

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 $\text{vect}(X)$** , c'est le point de $\text{vect}(X)$ le plus proche de Y .

$$\|Y - X\hat{\beta}\|^2 = \min_{\gamma \in \mathbb{R}^{p+1}} \|Y - X\gamma\|^2.$$

Meilleure approximation (1)

On a dit que $\hat{Y} = X\hat{\beta}$ est le point de vect(X) le plus proche de Y , on a donc

$$\|Y - X\hat{\beta}\|^2 = \min_{\gamma \in \mathbb{R}^{p+1}} \|Y - X\gamma\|^2.$$

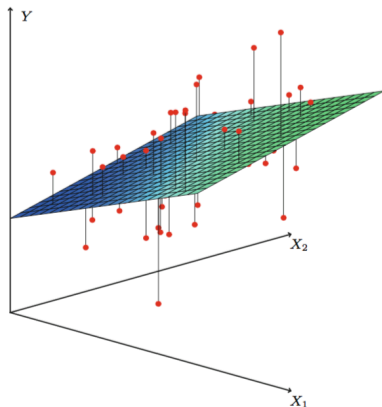


Figure 1: Dans $\mathbb{R}^{p+2} = \mathbb{R}^3$

Meilleure approximation (2)

Conséquence

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

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

- ▶ si H est la matrice de projection dans $\text{vect}(X)$

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

- ▶ on a

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

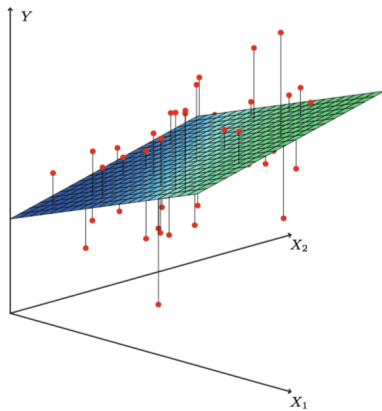
Meilleure approximation (3)

Définition : erreurs résiduelles

On note le vecteur des erreurs résiduelles $e = Y - X\hat{\beta}$ (projection de Y sur $\text{vect}(X)^\perp$) on a

$$Y - X\hat{\beta} \perp X\hat{\beta}.$$

L'estimateur des moindres carrés de σ^2 est donné par $\hat{\sigma}^2 = \frac{\|e\|^2}{n-(p+1)}$.



Suite des Conséquences des Propriétés géométriques

On obtient des mesures de l'adéquation du modèle

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

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

- ▶ 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

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

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

Quelques Propriétés de la matrice "hat"

Propriétés de la matrice $H = X(X^T X)^{-1}X^T$

1. $H^2 = H$ et $H^T = H$
2. $\text{rang}(H) = \text{tr}(H) = p + 1$
3. $H_{ii} = X_i(X^T X)^{-1}X_i^T$
4. $0 \leq H_{ii} = H_{ii}^2 + \sum_{k \neq i} H_{ik}^2 \leq 1$

Pour une preuve, voir http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf

Modèle linéaire gaussien

Définition : Modèle linéaire gaussien

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

où

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

Théorème de Cochran

Si V_1, \dots, V_k sont des s.e.v. orthogonaux dans \mathbb{R}^n de dimension n_1, \dots, n_k et si Z_1, \dots, Z_k sont les projections orthogonales d'un vecteur gaussien standard sur V_1, \dots, V_k alors

- ▶ les v.a. Z_1, \dots, Z_k sont gaussiens et deux à deux indépendants
- ▶ et, en particulier, $\|Z_j\|^2 \sim \chi^2(n_j)$ pour $j = 1, \dots, k$.

Conséquences du Théorème de Cochran

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

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

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

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

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

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

on utilisera donc le test de Student pour tester la nullité d'un coefficient.

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

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

On a alors

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

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 ($i = 1, \dots, n$), 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 $X_i\beta$

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

Pour une preuve, voir http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf

Erreur de prévision de Y

Si on considère un nouvel individu indépendant de $1, \dots, n$ 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 :

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

Intervalle de prévision pour Y_k

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

Pour une preuve, voir http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf

Modèle avec 1 covariable

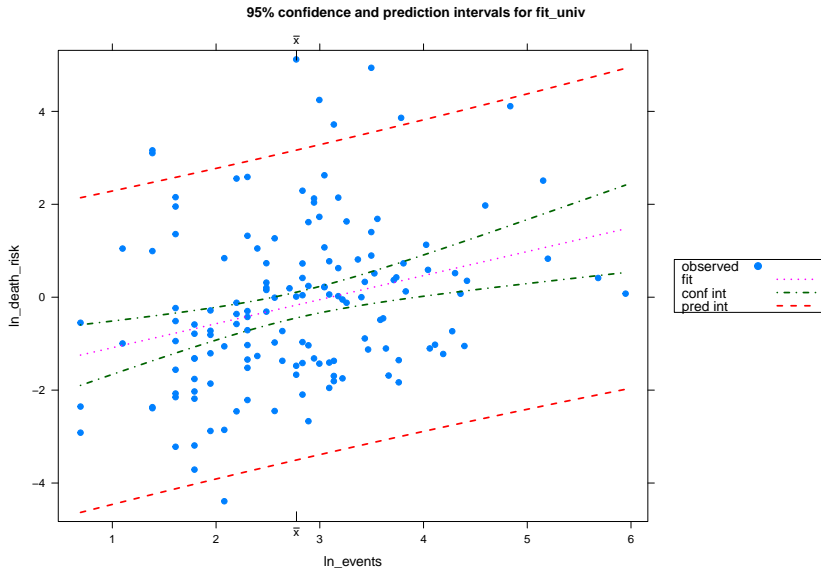
```
fit_univ = lm(ln_death_risk~ln_events)
ic=predict(fit_univ,interval="confidence")
print(ic[1:5,])
```

```
##           fit           lwr           upr
## 1 -0.41346753 -0.72116718 -0.1057679
## 2  0.20424358 -0.14006908  0.5485563
## 3 -0.02960412 -0.31691647  0.2577082
## 4  0.27723443 -0.09224715  0.6467160
## 5 -0.88753758 -1.36903423 -0.4060409
```

```
newdata=data.frame(ln_events=3.4)
pred=predict(fit_univ,newdata,interval="predict")
print(pred)
```

```
##           fit           lwr           upr
## 1 0.1543123 -3.185642  3.494266
```

```
ci.plot(fit_univ)
```



Lien entre valeur ajustée et prévision

On considère deux situations pour l'individu i

1. On a X, Y et calcule $\hat{\beta} = (X^T X)^{-1} X^T Y$ et

$$\hat{Y}_i = X_i \hat{\beta} = X_i (X^T X)^{-1} X^T Y$$

2. On efface les données de i pour obtenir $X_{(i)}$ et $Y_{(i)}$ et

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

avec lequel on prédit

$$Y_i^p = X_i \hat{\beta}_{(i)} = X_i (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}.$$

Estimation et prévision

On montre alors que

$$Y_i - \hat{Y}_i = (1 - H_{ii})(Y_i - Y_i^p)$$

Pour une preuve, voir http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf

Lemme d'inversion matricielle Sherman-Morrison

A inversible dans \mathbb{R}^n et $u, v \in \mathbb{R}^n$ alors

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}.$$

Preuve : il suffit de vérifier...

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

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

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 $X^T X$ n'est pas inversible, il n'y a donc pas de solution unique à l'équation

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

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

Valeurs propres de la matrice de corrélation

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

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

- ▶ 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 $\lambda_1 \geq \dots \geq \lambda_p$ 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

$$\kappa > 500 \text{ ou } 1000 \implies \text{colinéarité trop forte}$$

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
```

Variance inflation factor (VIF) et tolérance

Considérons la régression de la variable X^j sur les autres variables explicatives $X^1, \dots, X^{j-1}, X^{j+1}, \dots, X^p$, on note R_j^2 le R^2 associé à cette régression.

- ▶ Si $R_j^2 = 0$, X^j n'est pas fonction linéaire des autres variables
- ▶ Si $R_j^2 = 1$, X^j est fonction linéaire des autres variables \implies colinéarité

On définit les coefficients de "variance inflation factor" (VIF) pour $j = 1, \dots, p$ par :

$$VIF_j = \frac{1}{1 - R_j^2}.$$

Règle

Si $VIF > 10$ ou $100 \implies$ colinéarité

```
vif(fit)
```

```
## ln_events   ln_fert      hdi    ln_pop  
##  2.421759  3.642415  3.767663  2.460624
```

Propriété d'inflation de variance

On montre que

$$\mathbb{V}\hat{\beta}_j = \frac{\sigma^2}{\|\mathbf{X}^j - \bar{\mathbf{X}}\|^2} \frac{1}{1 - R_j^2}.$$

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.

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**

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

où $\hat{\sigma}_{(i)}^2$ 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

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

Pour une preuve, voir http://www.proba.jussieu.fr/pageperso/rebafka/StatBase_poly_partie2.pdf

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 $e = y - \hat{y}$

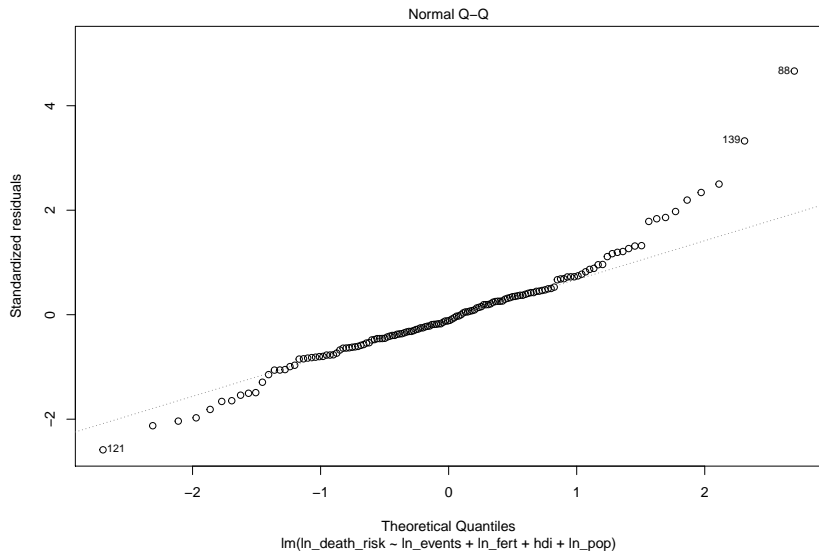
```
e = fit$residuals
```

Residus studentisés e^*

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

Normalité des erreurs

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



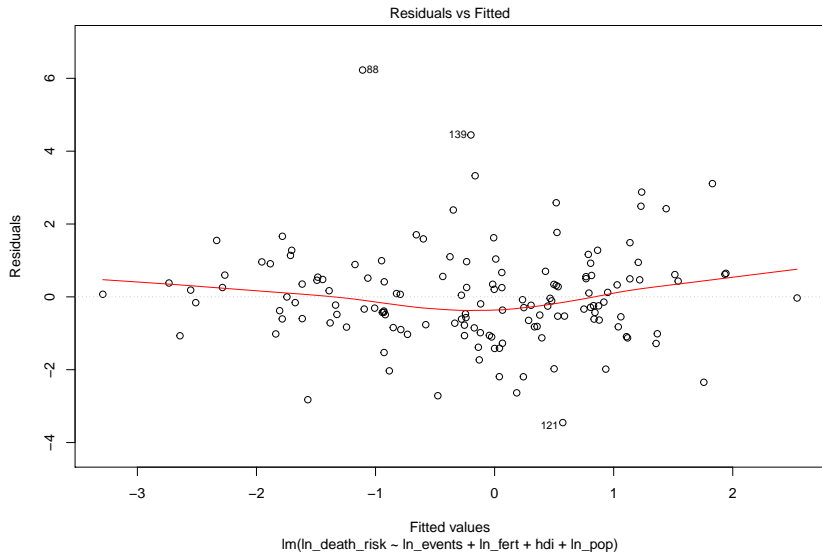
Dans le modèle linéaire

On veut vérifier les autres hypothèses sur les erreurs ϵ :

1. pour le centrage : c'est toujours vrai si on inclut l'intercept $(1, \dots, 1)^T$ dans la matrice X .
2. pour l'indépendance, il n'existe pas de test dans R. On conseille de représenter les résidus e^* contre les valeurs ajustées $X\hat{\beta}$ qui doivent être décorrélées.
3. pour l'égalité des variances, on recommande de représenter les résidus e^* contre les différents régresseurs X^j , $j = 1, \dots, p$. Cela permet d'identifier un régresseur responsable d'une hétéroscédasticité
4. si on suspecte une autocorrélation, en particulier si on étudie une série chronologique, on utilise le test de Durbin-Watson (fonction `dwtest` du package `lmtest`).

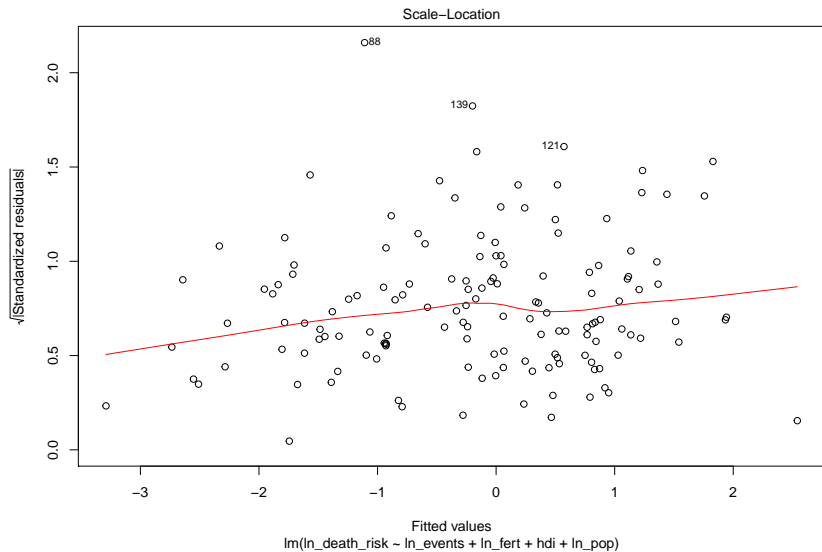
Graphique résidus/valeurs ajustées

```
plot(fit,which=1)
```



Graphique scale-location

```
plot(fit,which=3)
```



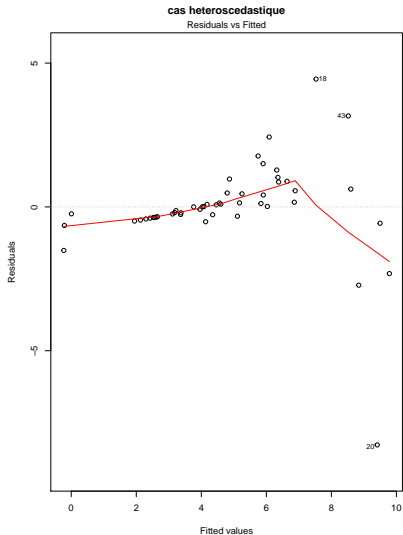
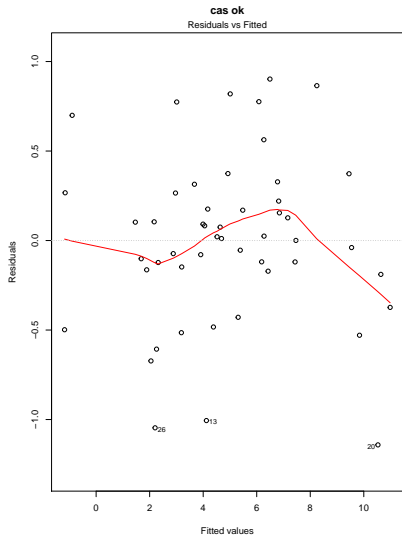
Cas avec hétéroscédasticité :

Données simulées

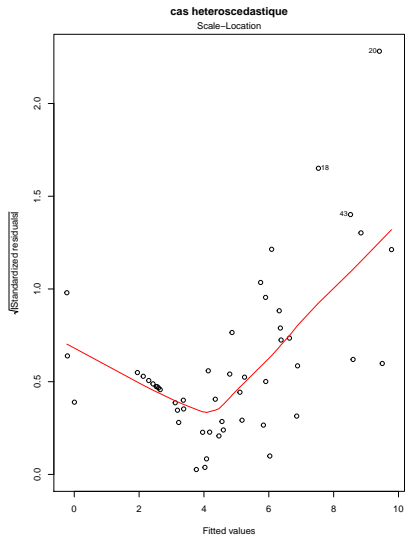
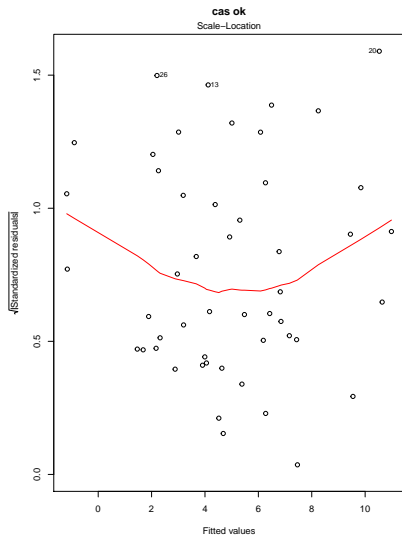
```
n =50
X=rnorm(n,1)
epsilon = rnorm(n,0,0.5)
Yok = 2 +3* X+epsilon
lmok = lm(Yok~X)

Yhs = 2 +3* X+ abs(X)^2*epsilon
lmhs = lm(Yhs~X)
```

Graphiques résidus/ajustés



Graphiques scale/location



Observations

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^*

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

Règle

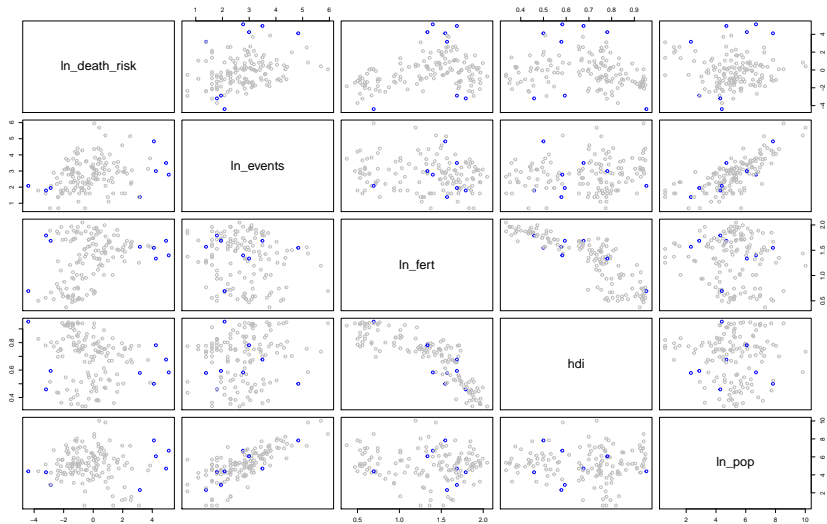
on dit qu'une observation est **aberrante** si

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

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

Où sont les points “aberrants” ?

Simple Scatterplot Matrix



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

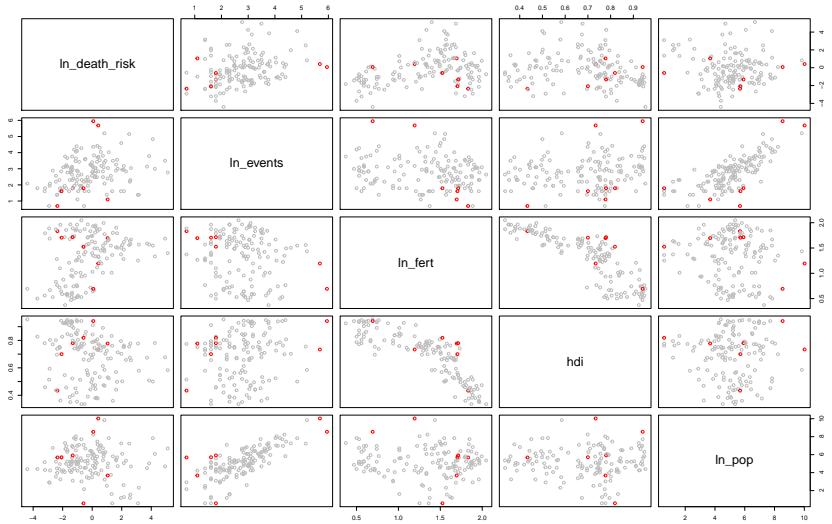
On montre que :

- ▶ $H_{ii} = 1$ ssi $\text{vect}(X_j, j \neq i)$ 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 $3p/n$).

Simple Scatterplot Matrix



Exercice

A titre d'exercice, vous pouvez montrer dans le cas d'un design orthogonal et normé

$$\langle X^j, X^k \rangle = 0 \text{ quand } j \neq k$$

$$\bar{X}^j = 0 \quad \forall j$$

$$\|X^j\| = 1/n \quad \forall j,$$

que si, pour l'individu n , $\|X_n\| \rightarrow \infty$ alors $H_{nn} \rightarrow 1$ (remarque : l'indice n de l'individu est seulement choisi pour faciliter les calculs, c'est vrai pour tout individu $i = 1, \dots, n$).

On comprend alors pourquoi le levier H_{ii} sont une mesure de l'isolement de l'individu i dans l'espace des covariables.

Autres mesures d'influence

- ▶ la **distance de Cook** est une mesure globale :

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

- ▶ les **dfbetas** permettent de mesurer l'influence d'une observation i sur une variable j

$$|DFBETA_{ij}| = \left| \frac{\hat{\beta}_j - \hat{\beta}_j(-i)}{\sqrt{\hat{\sigma}^2(-i)(X^T X)_{jj}^{-1}}} \right| > 2/\sqrt{n} \implies \text{influence sur la variable } j$$

Sur la distance de Cook (1)

Exercice A titre d'exercice, vous pouvez montrer, via la formule de Sherman-Morrison, que

$$\mathbf{X}(\hat{\beta} - \hat{\beta}_{(i)}) = (1 - H_{ii})^{-1} \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i^\top \mathbf{e}_i.$$

Montrer ensuite que

$$\frac{\|\mathbf{X}(\hat{\beta} - \hat{\beta}_{(i)})\|^2}{\sigma^2} = \frac{\mathbf{e}_i^\top \mathbf{H}_{ii}}{\sigma^2 (1 - H_{ii})^2}.$$

On comprend alors que la distance de Cook essaie de mesurer l'influence de l'observation i sur l'estimation.

Sur la distance de Cook (2)

Exercice

On souhaite déterminer l'ordre de grandeur d'un seuil critique pour la distance de Cook.

Calculer l'espérance de $e_i^2/(\sigma^2(1 - H_{ii}))$. Puis calculer pour un H_{ii} moyen, i.e. égal à $(p + 1)/n$, la valeur de $H_{ii}/(1 - H_{ii})$. Quelle alors la valeur de la pseudo-distance de Cook

$$\frac{e_i^2 H_{ii}}{(p + 1)\sigma^2(1 - H_{ii})^2}$$

on fera l'approximation $n - p \simeq n$.

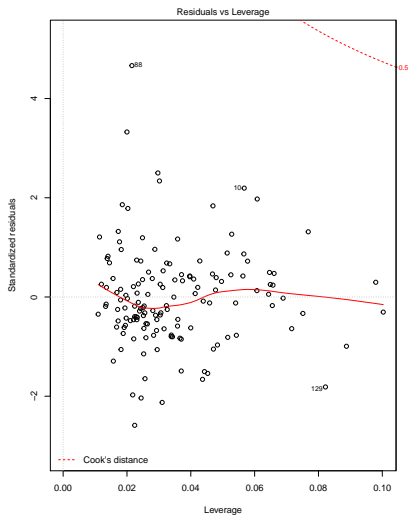
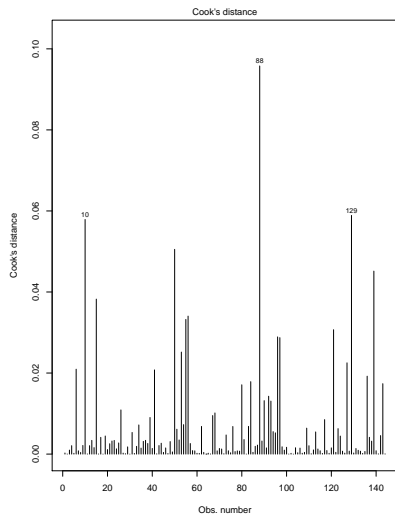
- ▶ le **covratio** permet d'évaluer si une observation i augmente ou diminue la prévision de l'estimation

$$\text{COVRATIO}_i = \frac{\hat{\sigma}^2(-i)\|(X^\top X)^{-1}(-i)\|}{\hat{\sigma}^2\|(X^\top X)^{-1}\|}$$

si $|\text{COVRATIO}_i - 1| > 3(p + 1)/n \implies$ influence

- ▶ Mesures de l'influence sur les valeurs prédites DFFITS et PRESS

Graphique DCook



Autres problèmes

Résidus partiels

Quand on suspecte un problème de linéarité entre un X^j et Y , on représente le résidu partiel

$$e_p^j = e + X^j \hat{\beta}_j$$

contre le régresseur X^j .

En effet

$$\frac{(X^j)^T e_p^j}{\|X^j\|^2} = \hat{\beta}_j$$

Résidus partiels

On peut trouver une autre "justification" en raisonnant comme suit. On veut vérifier si l'influence de X^j sur Y est bien linéaire, enlevons dans un premier temps cette hypothèse

$$Y = \sum_{k \neq j} X^k \beta_k + \psi(X^j) + \epsilon.$$

On voudrait idéalement représenter le nuage $(X_i^j, \psi(X_i^j))$ pour avoir une idée de la forme de ψ . On a accès aux X_i^j mais pas aux $\psi(X_i^j)$, estimons les. On a

$$\psi(X^j) = Y - \sum_{k \neq j} X^k \beta_k - \epsilon,$$

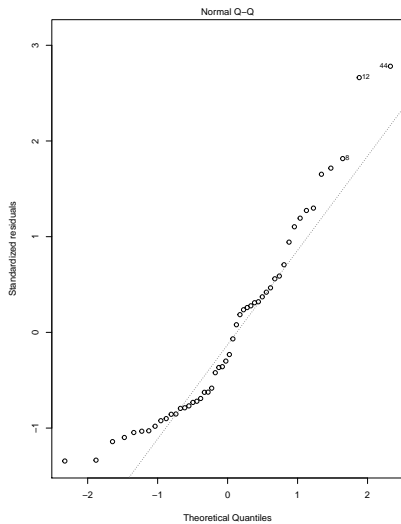
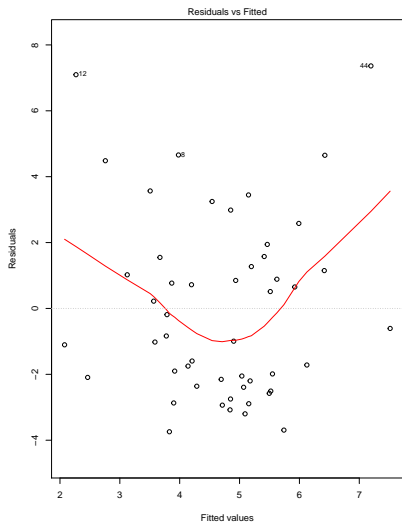
remplaçons les β_k inconnus par leurs estimateurs et ϵ par son espérance, on obtient

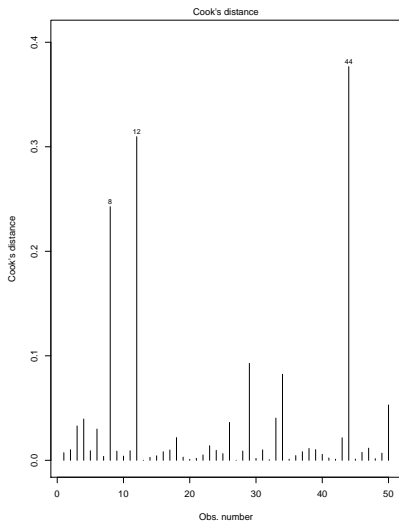
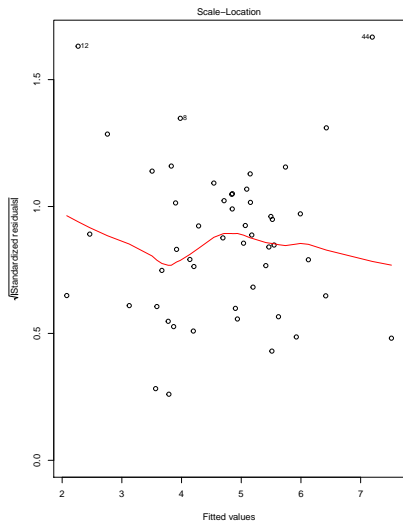
$$\widehat{\psi(X^j)} = Y - \sum_{k \neq j} X^k \hat{\beta}_k = Y - \sum_{k=1}^p X^k \hat{\beta}_k + X^j \hat{\beta}^j = e_p^j.$$

Relation non-linéaire due à une covariable

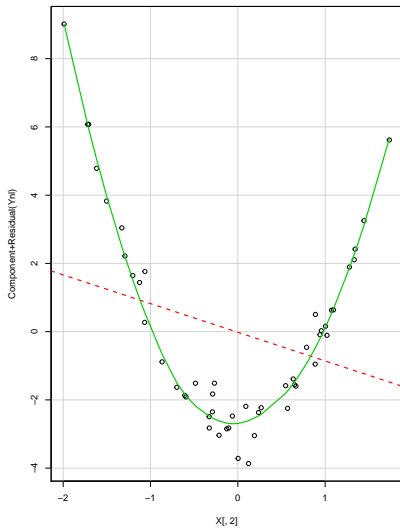
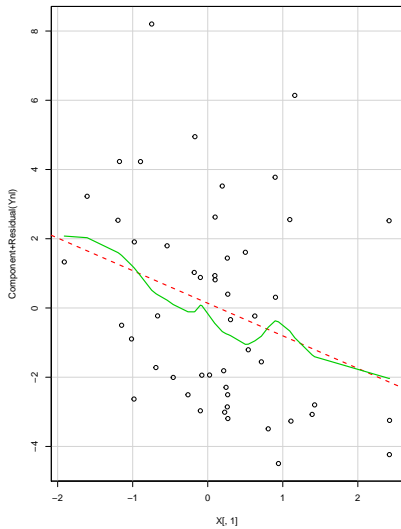
```
n = 50
X=matrix(rnorm(n*2),ncol=2)
epsilon = rnorm(n,0,0.5)

Ynl = 2 - X[,1] + 3* X[,2]^2 + epsilon
lmnl = lm(Ynl~X[,1]+X[,2])
```



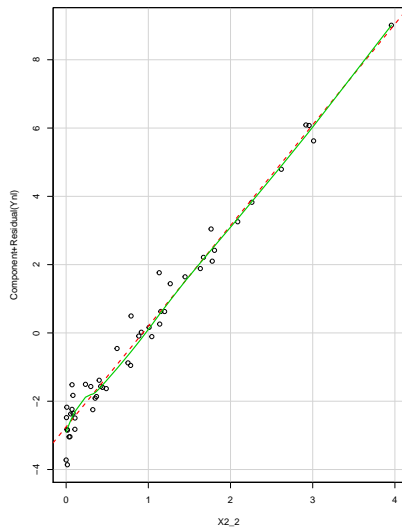
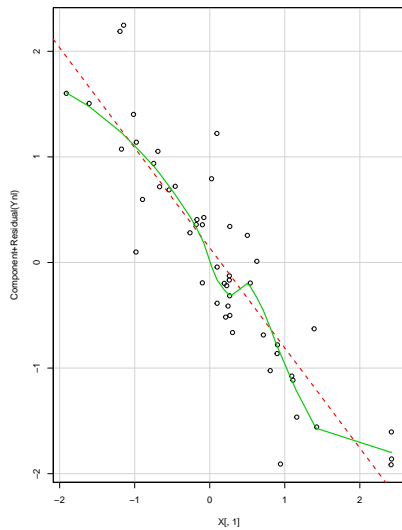


Component + Residual Plots



Transformation de la variable

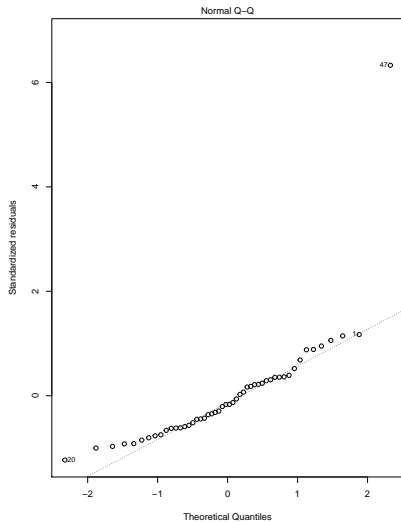
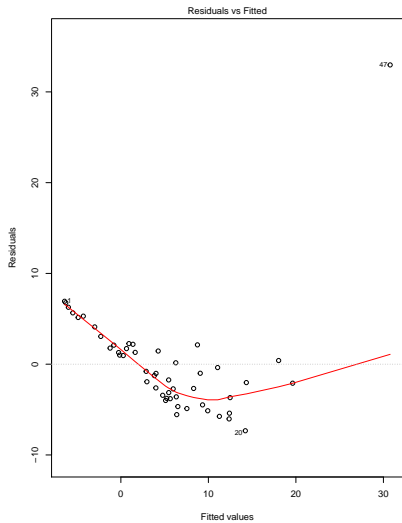
Component + Residual Plots

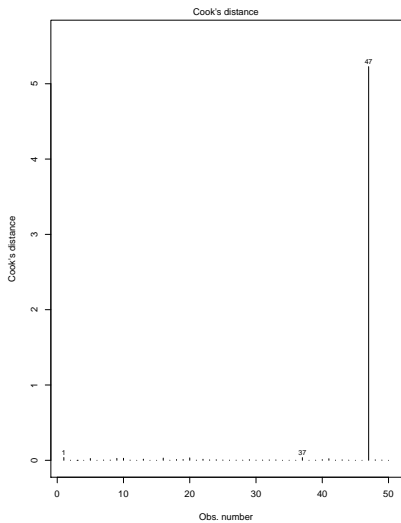
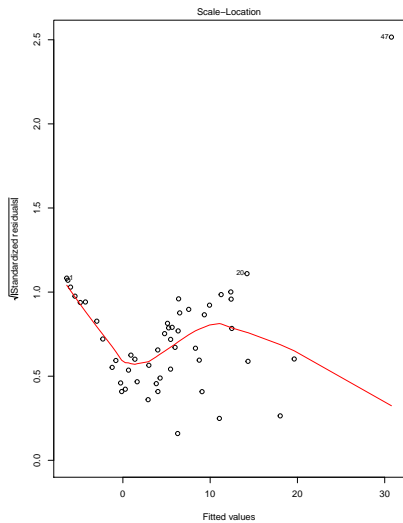


Relation non-linéaire due à Y

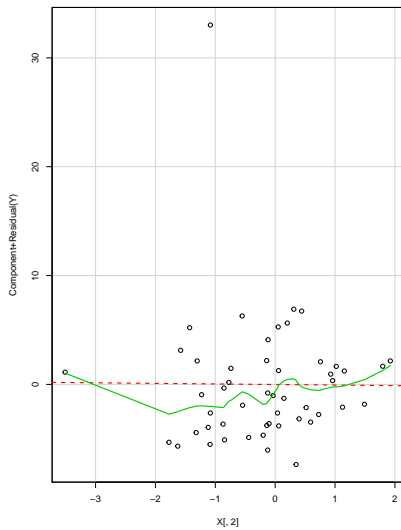
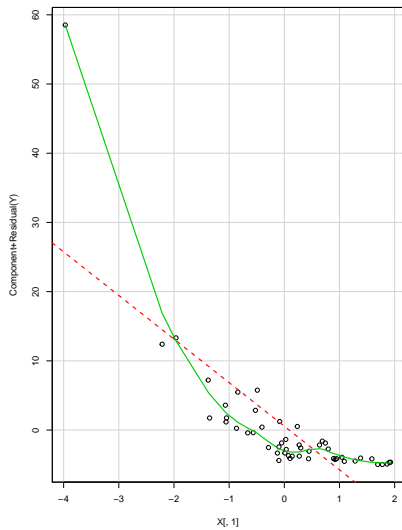
```
n =50
X=matrix(rnorm(n*2),ncol=2)
epsilon = rnorm(n,0,0.5)

ln_Y = 1 - X[,1] + 0.1* X[,2] + epsilon
Y = exp(ln_Y)
lm_ln = lm(Y~X[,1]+X[,2])
```

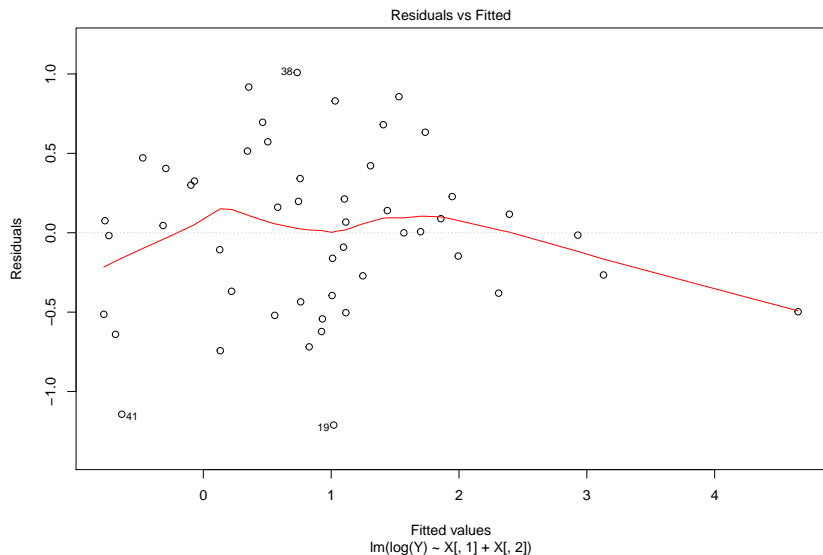


Component + Residual Plots



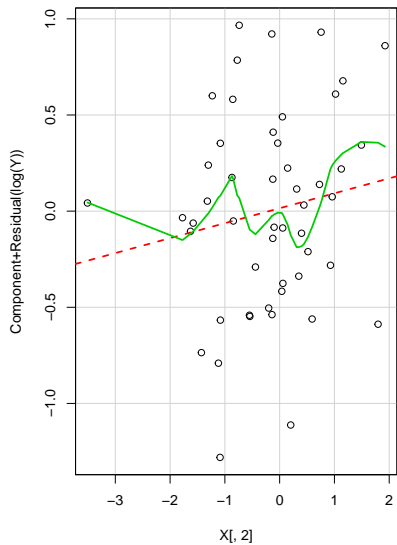
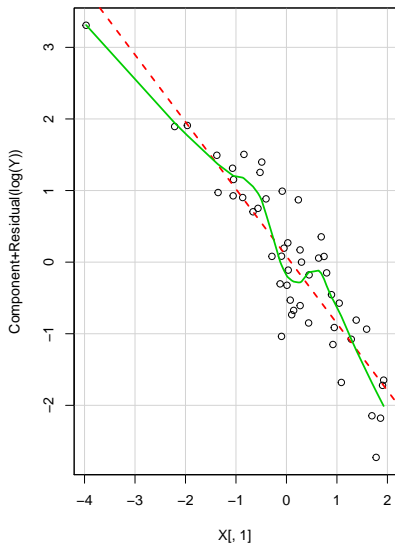
Transformation de Y

```
lm_ln_trans = lm(log(Y)~X[,1]+X[,2])  
plot(lm_ln_trans)
```



```
crPlots(lm_ln_trans)
```

Component + Residual Plots



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 !

Interprétation

```
?mtcars
```

```
mtcars_simple = mtcars[c(1,2,6)]
```

```
summary(mtcars_simple)
```

##	mpg	cyl	wt
##	Min. :10.40	Min. :4.000	Min. :1.513
##	1st Qu.:15.43	1st Qu.:4.000	1st Qu.:2.581
##	Median :19.20	Median :6.000	Median :3.325
##	Mean :20.09	Mean :6.188	Mean :3.217
##	3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:3.610
##	Max. :33.90	Max. :8.000	Max. :5.424


```
glimpse(mtcars_simple)
```

```
## Observations: 32
```

```
## Variables: 3
```

```
## $ mpg <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2...
```

```
## $ cyl <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 4, 4, 4...
```

```
## $ wt <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3...
```

```
mtcars_simple<- dplyr::mutate(mtcars_simple, cyl = factor(cyl))
glimpse(mtcars_simple)
```

```
## Observations: 32
```

```
## Variables: 3
```

```
## $ mpg <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2...
```

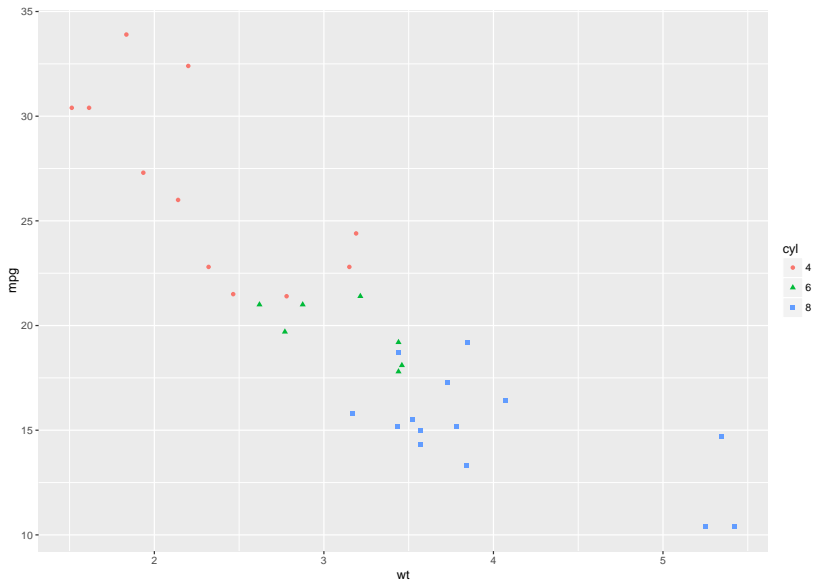
```
## $ cyl <fctr> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, ...
```

```
## $ wt <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3...
```

```
fit_simple = lm(mpg~wt+factor(cyl),data=mtcars_simple)
summary(fit_simple)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(cyl), data = mtcars_simple)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.9908     1.8878  18.006 < 2e-16 ***
## wt            -3.2056     0.7539  -4.252 0.000213 ***
## factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:  0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```
ggplot(mtcars_simple, aes(x=wt, y=mpg, color=cyl, shape=cyl)) +  
  geom_point()
```



```
fit_croise = lm(mpg ~ wt * cyl, data = mtcars_simple)
summary(fit_croise)
```

```
##
## Call:
## lm(formula = mpg ~ wt * cyl, data = mtcars_simple)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1513 -1.3798 -0.6389  1.4938  5.2523
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.571      3.194  12.389 2.06e-12 ***
## wt            -5.647      1.359  -4.154 0.000313 ***
## cyl16        -11.162      9.355  -1.193 0.243584
## cyl18        -15.703      4.839  -3.245 0.003223 **
## wt:cyl16       2.867      3.117   0.920 0.366199
## wt:cyl18       3.455      1.627   2.123 0.043440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.449 on 26 degrees of freedom
## Multiple R-squared:  0.8616, Adjusted R-squared:  0.8349
## F-statistic: 22.26 on 3 Df, 1.00 on 22 Df, p-value: 0.0000000e+00
```