

# Introduction au modèle linéaire

Retour au [plan du cours](#)

## 1 Introduction

Dans l'étude d'un phénomène comportant une ou plusieurs variables d'entrée  $X^{(1)}, X^{(2)}, \dots, X^{(k)}$  et une variable de sortie  $Y$ , la modélisation a pour but de tenter d'expliquer, par une relation mathématique, la valeur de  $Y$  en fonction des variables d'entrée :  $Y = g(X^{(1)}, X^{(2)}, \dots, X^{(k)})$ . Le modèle essaie de refléter le plus fidèlement possible la réalité. Il a pour but de mieux comprendre le phénomène étudié, et aussi de permettre de prédire  $Y$  sans devoir nécessairement réaliser des expériences.

On distingue deux types de modèles :

- Le **modèle théorique**, ou déterministe, est une équation qui émane souvent de lois physiques, chimiques, ou économiques, et représente le comportement attendu du phénomène.
- Souvent, il est difficile de développer un modèle théorique car le phénomène étudié est trop complexe. On a alors recours à un **modèle statistique** basé non pas sur une théorie mais sur des données observées.

### 1.1 Exemple 1

On étudie la pollution de l'air à New-York. On a mesuré pendant 111 jours la concentration en ozone, noté  $O_i$  (en ppm), et la température de l'air, notée  $T_i$  (en degrés Fahrenheit). Le tableau ci-dessous représente une partie des observations (celles pour lesquelles la vitesse du vent et le rayonnement solaire sont dans une certaine plage).

On constate que la concentration en ozone croît avec la température. La relation est approximativement linéaire dans la zone représentée ici. On introduit le modèle

$$O_i = a + bT_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Ce modèle est appelé **modèle de régression linéaire simple**.

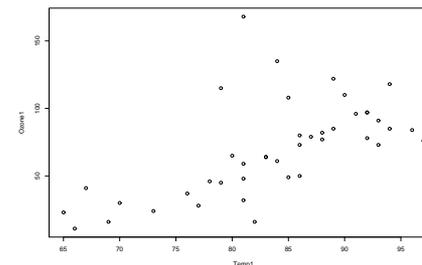


FIGURE 1 – Concentration en ozone en fonction de la température.

Une classe de modèles fréquemment utilisée en expérimentation est celle des modèles polynomiaux du type

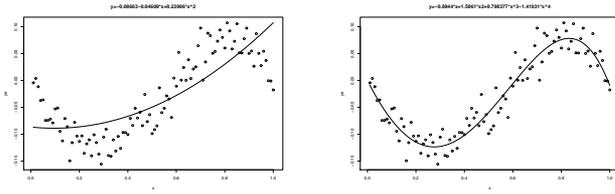
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_k X_i^k + \varepsilon_i, \quad i = 1, \dots, n$$

où l'on suppose que les variables aléatoires  $\varepsilon_i$  sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

$$\begin{aligned} E(Y_i) &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_k X_i^k \\ &= (1 \ X_i \ X_i^2 \ \dots \ X_i^k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{pmatrix} = h(\beta) \end{aligned}$$

avec  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{pmatrix}$

et  $h$  est une **application linéaire**. C'est pour cette raison que le modèle est appelé **modèle linéaire**. Voici deux exemples de modélisation polynomiale :



En statistique, on classe les modèles polynomiaux dans la classe des modèles linéaires car  $E(Y_i)$  varie linéairement par rapport aux paramètres  $\beta_i$ . Voici d'autres exemples de modèles linéaires :

$$Y_i = \beta_0 + \beta_1 \sqrt{X_i} + \beta_2 \frac{1}{X_i} + \varepsilon_i$$

ou

$$Y_i = \beta_0 + \beta_1 \sin X_i + \beta_2 e^{X_i} + \varepsilon_i.$$

Les modèles linéaires vont également être utilisés pour expliquer une variable  $Y$  à partir de plusieurs variables explicatives  $X^{(1)}, X^{(2)}, \dots, X^{(p)}$  :

$$Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i, i = 1 \dots, n,$$

on parle alors de modèle linéaire multiple.

Par exemple,  $X^{(1)}$  représente la Température,  $X^{(2)}$  la pression,  $X^{(3)}$  le carré de la température , etc.

## 2 Ajustement du modèle

On se place dans le modèle

$$Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i, i = 1 \dots, n.$$

Pour ajuster le modèle, on utilise la méthode des moindres carrés : on cherche les estimateurs  $(\hat{\beta}_1, \dots, \hat{\beta}_p)$  qui minimisent la somme des carrés des distances des observations  $Y_i$  à la courbe ajustée : on doit minimiser

$$f(\beta_1, \dots, \beta_p) = \sum_{i=1}^n (Y_i - \beta_1 X_i^{(1)} - \beta_2 X_i^{(2)} - \dots - \beta_p X_i^{(p)})^2.$$

La fonction  $f$  est convexe,  $(\hat{\beta}_1, \dots, \hat{\beta}_p)$  sont donc obtenus en annulant les dérivées partielles de la fonction  $f$ , ce qui donne un système de  $p$  équations à  $p$  inconnues.

### 2.1 Exemple de la régression linéaire simple

On utilise le **critère des moindres carrés**, c'est-à-dire que l'on détermine les paramètres  $a$  et  $b$  qui minimisent le critère

$$\sum_{i=1}^n (Y_i - a - bX_i)^2$$

On cherche  $(\hat{a}, \hat{b})$  qui minimise la fonction

$$f(a, b) = \sum_{i=1}^n (Y_i - a - bX_i)^2$$

On peut montrer que  $f$  est une fonction convexe, il suffit donc pour trouver  $(\hat{a}, \hat{b})$  d'annuler les dérivées partielles de  $f$ . Avec les notations suivantes :

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$\text{Cov}(Y, X) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{Y} \bar{X},$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2,$$

on obtient comme solution de ce système d'équations :

$$\hat{a} = \bar{Y} - \hat{b} \bar{X}$$

$$\hat{b} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}.$$

Si on utilise cette droite de régression pour faire de la prédiction, on prédira, pour une valeur donnée de  $X_i$ , la valeur de  $Y_i$  par la quantité

$$\hat{Y}_i = \hat{a} + \hat{b} X_i$$

On estime  $\sigma^2$  par

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i)^2.$$

## 2.2 Cas général

On considère le modèle :

$$Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i, i = 1, \dots, n.$$

Pour déterminer les estimateurs des moindres carrés, il est plus simple d'écrire le modèle sous forme matricielle :

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

où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

et  $X$  est la matrice

$$X = \begin{pmatrix} X_1^{(1)} & X_1^{(2)} & \dots & X_1^{(p)} \\ X_2^{(1)} & X_2^{(2)} & \dots & X_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_n^{(1)} & X_n^{(2)} & \dots & X_n^{(p)} \end{pmatrix}.$$

On considère la fonction  $f$  définie par

$$\begin{aligned} f(\beta) &= \sum_{i=1}^n (Y_i - (X\beta)_i)^2 \\ &= \|Y - X\beta\|^2. \end{aligned}$$

$$\begin{aligned} \forall u \in \mathbb{R}^n, \|u\|^2 &= \sum_{i=1}^n u_i^2 \\ &= u'u = \langle u, u \rangle \end{aligned}$$

où  $'$  est l'opérateur de transposition et  $u' = (u_1, \dots, u_n)$ .

On peut alors écrire  $f(\beta_1, \dots, \beta_p)$  sous la forme

$$f(\beta_1, \dots, \beta_p) = (Y - X\beta)'(Y - X\beta) = \|Y - X\beta\|^2.$$

On cherche  $\hat{\beta}$  qui minimise  $f(\beta)$ .

La solution de la minimisation est :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (X'X)^{-1}X'Y.$$

*Remarque.* — On suppose que la matrice  $X'X$  est inversible, ce qui est équivalent à supposer que les colonnes de la matrice  $X$  sont linéairement indépendantes, ou encore que l'application  $\beta \mapsto X\beta$  est injective.

*Exemple.* — Modèle de régression linéaire simple

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \dots, n.$$

*Exemple.* — Analyse de la variance à un facteur.

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \dots, p, j = 1, \dots, n_i.$$

Comme dans le cas de la régression linéaire simple, on estime la variance  $\sigma^2$  du terme d'erreur à partir de la somme des carrés des résidus :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^n (Y_i - \hat{\beta}_1 X_i^{(1)} - \hat{\beta}_2 X_i^{(2)} - \dots - \hat{\beta}_p X_i^{(p)})^2 \\ &= \frac{1}{n-p} \|Y - X\hat{\beta}\|^2 \end{aligned}$$

### 3 Intervalles de confiance

On considère le modèle :

$$Y_i = \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \dots + \beta_p X_i^{(p)} + \varepsilon_i, i = 1 \dots, n,$$

où les variables  $\varepsilon_i, i = 1 \dots, n$ , sont i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ .

**THÉORÈME 1.** — *Les composantes du vecteur  $\hat{\beta}$  sont des variables aléatoires gaussiennes.*

$$E(\hat{\beta}) = \beta,$$

*c'est-à-dire que pour tout  $i = 1, \dots, p$ ,  $E(\hat{\beta}_i) = \beta_i$ , La matrice de variance-covariance du vecteur  $\hat{\beta}$ , notée  $\Gamma_{\hat{\beta}}$  est la matrice*

$$\Gamma_{\hat{\beta}} = \sigma^2 (X'X)^{-1}, \quad (\Gamma_{\hat{\beta}})_{ij} = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j).$$

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi_{(n-p)}^2$$

*et est indépendant de  $\hat{\beta}$ .*

L'élément d'indice  $(i, i)$  de la matrice  $\Gamma_{\hat{\beta}}$  représente la variance de  $\hat{\beta}_i$ , pour tout  $i$  de 1 à  $p$ . On a donc

$$\text{Var}(\hat{\beta}_i) = \sigma^2 (X'X)^{-1}_{i,i}.$$

Cette variance est estimée par

$$\hat{\sigma}^2 (X'X)^{-1}_{i,i}.$$

On montre que la variable aléatoire

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{i,i}}} \sim \mathcal{T}_{(n-p)}$$

suit une loi de Student à  $n - p$  degrés de liberté. On peut, à l'aide de cette relation, construire des intervalles de confiance pour les paramètres  $\beta_i$  et effectuer

des tests sur les valeurs de ces paramètres. L'intervalle de confiance de niveau de confiance 0.95 obtenu pour le paramètre  $\beta_i$  est l'intervalle

$$[\hat{\beta}_i - t_{n-p, 0.975} \sqrt{\hat{\sigma}^2 (X'X)^{-1}_{i,i}}, \hat{\beta}_i + t_{n-p, 0.975} \sqrt{\hat{\sigma}^2 (X'X)^{-1}_{i,i}}].$$

Si on souhaite tester l'hypothèse  $H_0 : \beta_i = b_0$  contre  $H_1 : \beta_i \neq b_0$ , on rejettera l'hypothèse  $H_0$  si  $b_0$  n'appartient pas à l'intervalle de confiance ci-dessus.

### 4 Prédiction

Un modèle linéaire peut également être utilisé pour faire de la prédiction, c'est-à-dire pour prévoir la valeur attendue pour la réponse  $Y$  lorsque les variables explicatives prennent des valeurs données. Cette question peut être vue sous deux angles :

- On peut s'intéresser au comportement moyen de  $Y$ , c'est -à-dire à  $E(Y)$  pour des valeurs données des variables explicatives.
- On peut aussi s'intéresser à la valeur réelle que prendra  $Y$  si un seul essai est réalisé en ces valeurs données des variables explicatives.

#### 4.1 Intervalle de confiance sur la réponse moyenne

A l'aide de l'estimation des paramètres effectuée sur les  $n$  essais, on souhaite prédire l'espérance de  $Y$  pour des valeurs données des variables explicatives.

Pour simplifier, on se place dans le cas d'un modèle polynomial de degré  $p-1$  :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_{p-1} X_i^{p-1} + \varepsilon_i, i = 1 \dots, n.$$

La valeur de  $Y$  lorsque  $X_i = X_0$  prédite par le modèle est

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \hat{\beta}_2 X_0^2 + \dots + \hat{\beta}_{p-1} X_0^{p-1} = \hat{\beta}' \mathbf{X}_0$$

où

$$\mathbf{X}_0 = (1, X_0, X_0^2, \dots, X_0^{p-1})'.$$

La réponse moyenne pour une variable d'entrée égale à  $X_0$  est

$$\beta' \mathbf{X}_0 = \beta_0 + \beta_1 X_0 + \beta_2 X_0^2 + \dots + \beta_k X_0^{p-1}$$

On montre que  $\hat{Y}_0 = \hat{\beta}'\mathbf{X}_0$  suit une loi normale d'espérance  $\beta'\mathbf{X}_0$  et de variance  $\sigma^2\mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$ . On en déduit un intervalle de confiance pour la réponse moyenne lorsque la variable d'entrée égale à  $X_0$  :

$$\left[ \hat{\beta}'\mathbf{X}_0 - t_{n-p,0.975}\hat{\sigma}\sqrt{\mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}, \right. \\ \left. \hat{\beta}'\mathbf{X}_0 + t_{n-p,0.975}\hat{\sigma}\sqrt{\mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \right].$$

## 4.2 Intervalle de prédiction

Pour simplifier, on se place toujours dans le cas d'un modèle polynomial de degré  $p - 1$ . Si on veut prédire dans quel intervalle se trouvera le résultat d'un essai individuel pour des valeurs données des variables explicatives, on doit tenir compte de deux facteurs d'incertitude : l'incertitude sur l'estimation du résultat moyen de l'essai  $E(Y)$ , et l'incertitude sur le terme d'erreur  $\varepsilon$ .

$\hat{\beta}$  a été estimé avec les données  $(X_i, Y_i), i = 1, \dots, n$ . Une nouvelle observation  $Y_0$ , correspondant à  $X_i = X_0$  s'écrit :

$$Y_0 = \beta_0 + \beta_1 X_0 + \beta_2 X_0^2 + \dots + \beta_{p-1} X_0^{p-1} + \varepsilon_0,$$

où  $\varepsilon_0$  est indépendant de  $\hat{\beta}$ . On a donc

$$Y_0 - \hat{Y}_0 = (\beta - \hat{\beta})'\mathbf{X}_0 + \varepsilon_0$$

$Y_0$  et  $\hat{Y}_0$  sont des variables aléatoires indépendantes, de lois normales. Il en résulte que  $Y_0 - \hat{Y}_0$  suit une loi normale centrée de variance  $\sigma^2 W_0^2$  où

$$W_0^2 = 1 + \mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0.$$

Comme

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

est indépendant de  $(\hat{\beta}, \varepsilon)$ , la variable

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma}W_0}$$

suit une loi de Student à  $n - p$  degrés de liberté.

On obtient

$$P\left(\hat{Y}_0 - t_{n-p,1-\alpha/2}\hat{\sigma}W_0 \leq Y_0 \leq \hat{Y}_0 + t_{n-p,1-\alpha/2}\hat{\sigma}W_0\right) = 1 - \alpha.$$

On obtient donc un intervalle de prédiction pour la valeur d'une observation  $Y_0$  au point  $X_0$  avec coefficient de sécurité 0.95 :

$$\left[ \hat{\beta}'\mathbf{X}_0 - t_{n-p,0.975}\hat{\sigma}\sqrt{1 + \mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}, \right. \\ \left. \hat{\beta}'\mathbf{X}_0 + t_{n-p,0.975}\hat{\sigma}\sqrt{1 + \mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \right].$$

*Remarque.* — Il est recommandé de n'utiliser le modèle pour faire de la prédiction que dans le domaine couvert par les données : le phénomène peut obéir au modèle linéaire dans le domaine observé, et avoir un comportement différent dans un autre domaine.

## 5 Validation du modèle, Choix de modèles.

Pour utiliser avec confiance un modèle, il est important de vérifier sa validité.

### 5.1 Qualité de l'ajustement

On note  $\hat{Y}_i$  les valeurs prédites par le modèle :  $\hat{Y}_i = (X\hat{\beta})_i$ . On appelle résidus les variables

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

et on estime la variance  $\sigma^2$  des erreurs  $\varepsilon_i$  par

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n (\hat{\varepsilon}_i)^2.$$

A noter que  $\sum_{i=1}^n (\hat{\varepsilon}_i)^2/n$  représente la moyenne des carrés des écarts entre les  $Y_i$  et les valeurs prédites  $\hat{Y}_i$ .

## 5.2 $R^2$ et $R^2$ ajusté

Un critère utilisé dans les logiciels pour mesurer la qualité de l'ajustement consiste à comparer la variance de  $Y$  avec la variance des résidus : on introduit

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

$R^2$  est appelé coefficient de détermination ou encore pourcentage de la variance de  $Y$  expliquée par le modèle.

On note

$$SCT = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad SCR = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, \quad SCE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

Si le vecteur  $\mathbf{1} = (1, 1, \dots, 1)' \in Im(X)$ , a la relation suivante :

$$SCT = SCR + SCE.$$

On en déduit que

$$R^2 = \frac{SCE}{SCR} = 1 - \frac{SCR}{SCT}$$

$$0 \leq R^2 \leq 1.$$

La qualité de l'ajustement est d'autant plus grande que cette valeur est proche de 1. Ce coefficient est appelé *coefficient de détermination*.

On peut montrer que, si on ajoute des variables explicatives, par exemple, si on utilise un modèle polynomial de degré plus grand, la valeur de  $R^2$  augmente. Néanmoins, on n'a pas toujours intérêt à prendre trop de variables explicatives car cela conduit à devoir estimer plus de paramètres, avec le même nombre d'observations, et la variance des paramètres estimés augmente. C'est la raison pour laquelle on utilise plutôt un critère, appelé  $R^2$  ajusté, qui tient compte du nombre de paramètres dans le modèle :

$$R'^2 = 1 - \frac{SCR/(n - k - 1)}{SCT/(n - 1)}.$$

On pourra choisir, entre deux modèles, celui qui maximise le  $R^2$  ajusté.

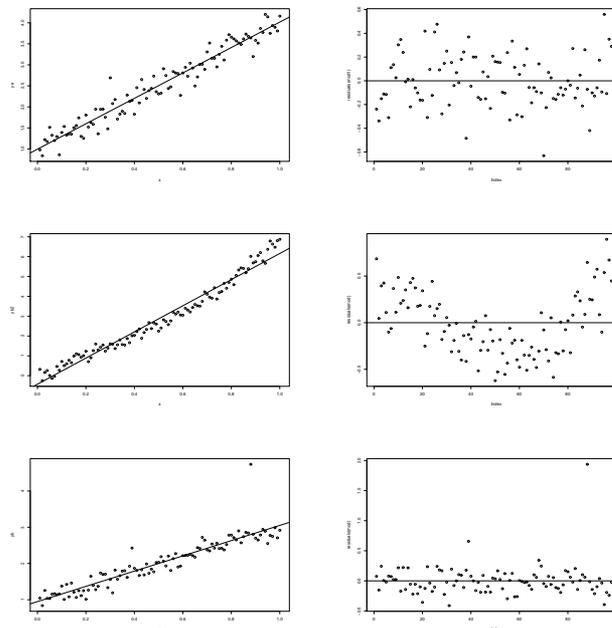
## 5.3 Analyse des résidus

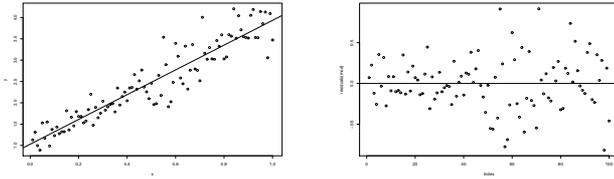
On se place ici dans un modèle de régression linéaire simple, mais les commentaires ci-dessous s'appliquent à des modèles généraux.

Un graphe des résidus du modèle a pour but de regarder si les résidus estimés

$$\hat{\varepsilon}_i = Y_i - \hat{a} - \hat{b}X_i$$

se comportent selon les hypothèses du modèle. Le modèle suppose en effet que les variables aléatoires  $\varepsilon_i$  sont indépendantes et de même loi  $\mathcal{N}(0, \sigma^2)$ . Les résidus estimés doivent donc montrer un comportement aléatoire de ce type. On représente en abscisse la valeur des  $X_i$  et en ordonnée la valeur  $\hat{\varepsilon}_i$  (colonne de droite). Voici 4 exemples :





- Le premier exemple ne pose pas de problème, les points sont bien distribués autour de la droite et le graphe des résidus montre bien un comportement aléatoire.
- Pour le deuxième exemple, il apparaît que l'équation du modèle n'est pas valide : la relation de dépendance entre  $X$  et  $Y$  n'est pas linéaire.
- La troisième régression est assez bonne mais fait apparaître un point aberrant très éloigné des autres, ce point tire la droite de régression vers le haut.
- Sur le quatrième exemple, on voit que la variabilité des points autour de la droite est d'autant plus grande que  $X$  augmente. Cette observation remet en cause l'hypothèse selon laquelle la variance des résidus  $\sigma^2$  est identique dans tout le domaine. On observe alors une allure "en entonnoir" du graphe des résidus.

Pour tester la normalité des résidus, on peut utiliser la droite de Henry, ou les tests vus au chapitre précédent (Shapiro-Wilk, Kolmogorov)...

### Que faire si les hypothèses ne sont pas vérifiées ?

- Choisir un modèle plus complexe (polynomial de degré au moins 2) si on s'aperçoit que le modèle de régression linéaire simple ne convient pas.
  - Eliminer les éventuels points aberrants qui nuisent à la qualité de l'estimation.
  - Si le modèle est à variance hétérogène, on peut parfois utiliser des transformations qui stabilisent la variance.
- Si on connaît  $\sigma_i^2 = Var(Y_i)$  pour tout  $i$ , on minimise

$$\sum_{i=1}^n \frac{(Y_i - a - bX_i)^2}{\sigma_i^2}.$$

Ceci permet d'accorder plus de poids aux variables  $Y_i$  de faible variance dans l'estimation de  $(a, b)$ .

## 6 Test de Fisher d'un sous-modèle

Supposons que l'on ait utilisé un modèle de régression polynomiale de degré 3, et que l'on souhaite effectuer un test statistique pour vérifier si un modèle de régression polynomiale de degré 2 est acceptable, ou bien on souhaite tester l'hypothèse  $H_0$  selon laquelle parmi les coefficients  $\beta_1, \dots, \beta_p$ , les  $k$  derniers coefficients sont nuls. Plus généralement, on dispose d'un modèle, appelé modèle (1) :

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

avec  $\beta \in \mathbb{R}^p$  et d'un modèle, appelé modèle (2) :

$$Y = \tilde{X}\theta + \varepsilon,$$

avec  $\theta \in \mathbb{R}^l$ .

DÉFINITION 2. — Soit

$$V = \{X\beta, \beta \in \mathbb{R}^p\}$$

et

$$W = \{\tilde{X}\theta, \theta \in \mathbb{R}^l\}$$

On dit que le modèle (2) est un sous-modèle du modèle (1) si  $W$  est un sous-espace vectoriel de  $V$ .

Exemples

$$Y = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (H_0)$$

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \quad (H_1)$$

On souhaite tester l'hypothèse

$H_0$  : "le vecteur  $Y$  des observations obéit au modèle (2)"  
contre l'hypothèse

$H_1$  : “le vecteur  $Y$  des observations obéit au modèle (1)”.

Dans le modèle (2), on estime  $\theta$  par

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_l \end{pmatrix} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y.$$

On montre à l’aide du théorème de Cochran que, sous l’hypothèse  $H_0$ , la statistique

$$F = \frac{\|X\hat{\beta} - \tilde{X}\hat{\theta}\|^2/(p-l)}{\|Y - X\hat{\beta}\|^2/(n-p)}$$

suit une loi de Fisher de paramètres  $(p-l, n-p)$ .

Sous l’hypothèse  $H_1$ , la statistique  $F$  aura tendance à prendre des valeurs plus grandes que sous  $H_0$ . On rejette l’hypothèse  $H_0$  si

$$F > f_{p-l, n-p, 1-\alpha}$$

où  $f_{p, q, 1-\alpha}$  est le  $(1-\alpha)$  quantile de la loi de Fisher de paramètres  $(p, q)$ . Dans les logiciels, on trouve pour ce type de test la  $p$ -value, il s’agit de la quantité

$$P_{H_0}(F > F_{obs})$$

où  $F_{obs}$  est la valeur observée pour la statistique de test  $F$ . Le test décrit ci-dessus équivaut à rejeter  $H_0$  si la  $p$ -value est inférieure au niveau du test  $\alpha$ .