Introduction aux fondements de l'apprentissage statistique

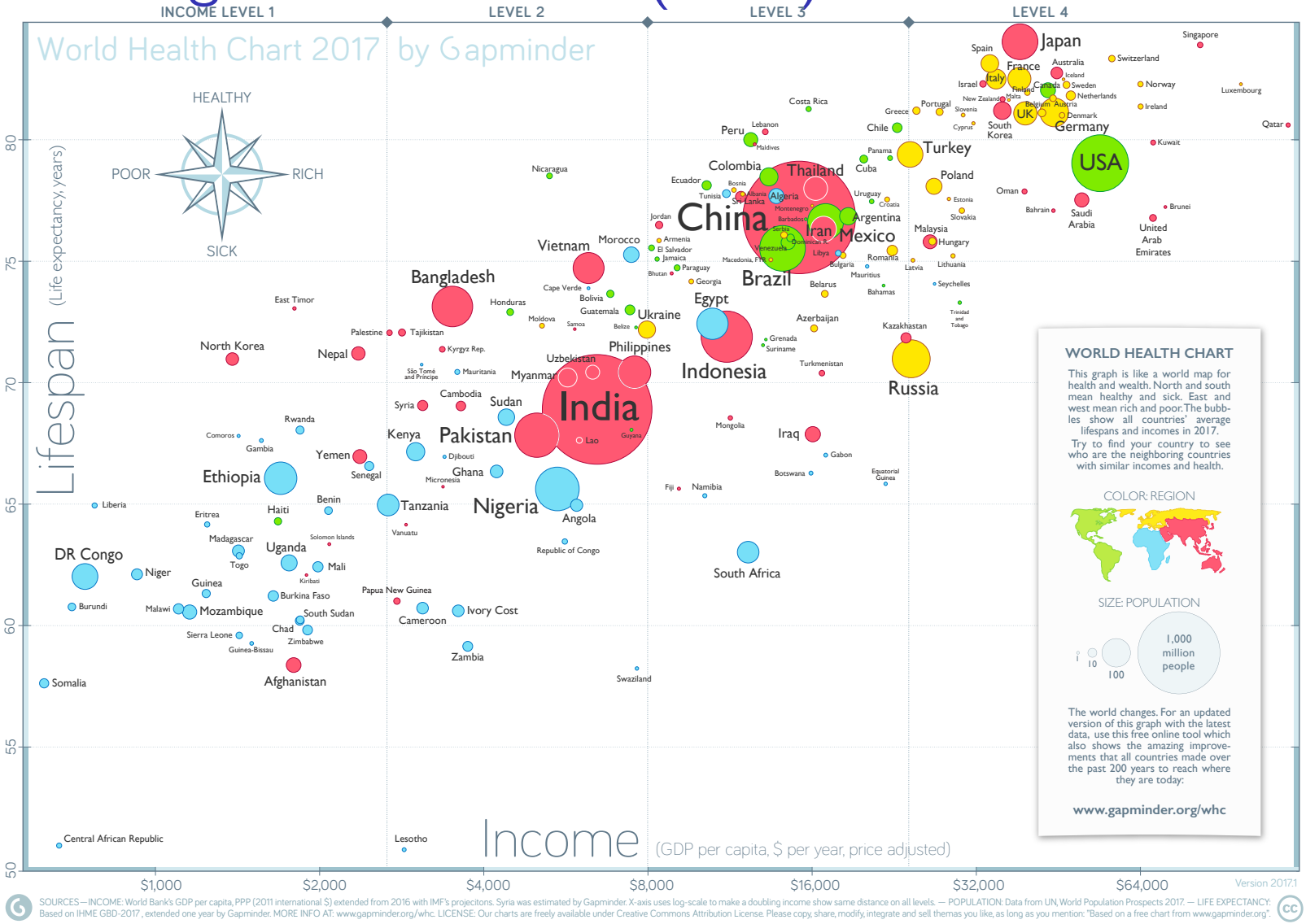
Marie-Pier Côté

École d'actuariat
Chaire de leadership en enseignement
en analyse de données massives pour l'actuariat — Intact

9 février 2019

Plan

Effet de l'argent sur la santé (2017)



Graphique en nuage de points montrant le lien entre l'espérance de vie à la naissance et le revenu domestique moyen d'un pays.

`https://www.gapminder.org/tools`

Quelques questions

- Quel serait l'impact attendu d'une augmentation du PIB sur l'espérance de vie ?
- Le continent et la taille de la population ont-ils un impact sur l'espérance de vie ?
- Si on considère un pays qui n'est pas sur le graphique et qu'on connait le PIB, peut-on prédire l'espérance de vie, ou donner une intervalle de valeurs plausibles pour celle-ci ?

But d'une analyse de régression

Étudier les relations qui existent entre des variables (ou facteurs) mesurables à partir d'observations (données) prises sur ces variables.

Modèle d'apprentissage statistique

L'apprentissage statistique est un ensemble de techniques de modélisation qui vise à expliquer une variable Y en fonction de variables $x_1, \ldots, x_p$.

Cela peut s'exprimer mathématiquement par

$$Y = f(x_1, \ldots, x_p) + \text{fluctuation aléatoire.}$$

- La variable Y est appelée variable endogène.
  - Synonymes : variable réponse, variable dépendante.
- Les variables $x_1, \ldots, x_p$ sont nommées variables exogènes.
  - Synonymes : variables explicatives, facteurs (*features*), covariables (*covariates*), et variables indépendantes.
- La fonction f est fixée mais inconnue, on vise à apprendre ou estimer f.

Modèle déterministe

Dans un monde idéal, la relation serait exacte et donc la valeur de Y serait uniquement déterminée par les valeurs de $x_1, \ldots, x_p$. Le modèle serait alors un modèle dit déterministe.

Exemple : Le modèle d'atmosphère standard de l'OACI [1] exprime la relation entre la pression et l'altitude de la manière suivante :

$$p = p_0(1 - 2.26 \times 10^{-5} h),$$

où p_0 désigne la pression au niveau de la mer en hPa ($p_0 = 1013.25\, hPa$) et h note l'altitude en m.

1. pour les altitudes jusqu'à 80 km et sans tenir compte de la vapeur d'eau

Modèle avec fluctuation aléatoire

Par contre, il est possible

- que les données soient entachées d'une erreur de mesure (expérimentale), ou
- que certains facteurs contributifs de moindre importance aient été négligés.

Généralement, les valeurs de $x_1, \ldots, x_p$ peuvent expliquer une partie de Y, alors que l'autre partie demeure inexpliquée.

Modèle avec fluctuation aléatoire — Exemple

Soient Y le poids d'un bébé fille à la naissance (en kg) et x sa taille (en cm). On peut écrire

$$Y = f(x) + \text{fluctuation aléatoire},$$

étant donné que deux bébés de même taille, par exemple 50 cm, peuvent avoir des poids très différents.

En fait, les modèles de régression sont de la forme

$$\mathrm{E}(Y) = f(x),$$

puisque l'espérance de la fluctuation aléatoire est supposée nulle.

Prévisions

Si on peut estimer f par $\hat{f}$, alors une prévision pour la variable d'intérêt Y étant donné des valeurs de $x_1, \ldots, x_p$ est

$$\hat{Y} = \hat{f}(x_1, \ldots, x_p).$$

Deux types de techniques servent à estimer f dans ce genre de problème : l'apprentissage automatique (machine learning) et la modélisation statistique (étudié aujourd'hui).

Deux noms différents pour un même concept ?

Les différences entre les deux approches résident surtout dans les objectifs de l'analyse.

Objectifs de l'apprentissage automatique

1. But principal de l'analyse est de prédire la valeur de Y.
2. L'estimation explicite/interprétable de l'effet de la valeur de x_j sur Y n'est PAS un intérêt majeur.
3. Le comportement de la fluctuation aléatoire n'est PAS un intérêt majeur.
4. On a peu/pas d'hypothèses a priori sur la forme de f, on laisse les données faire le travail.

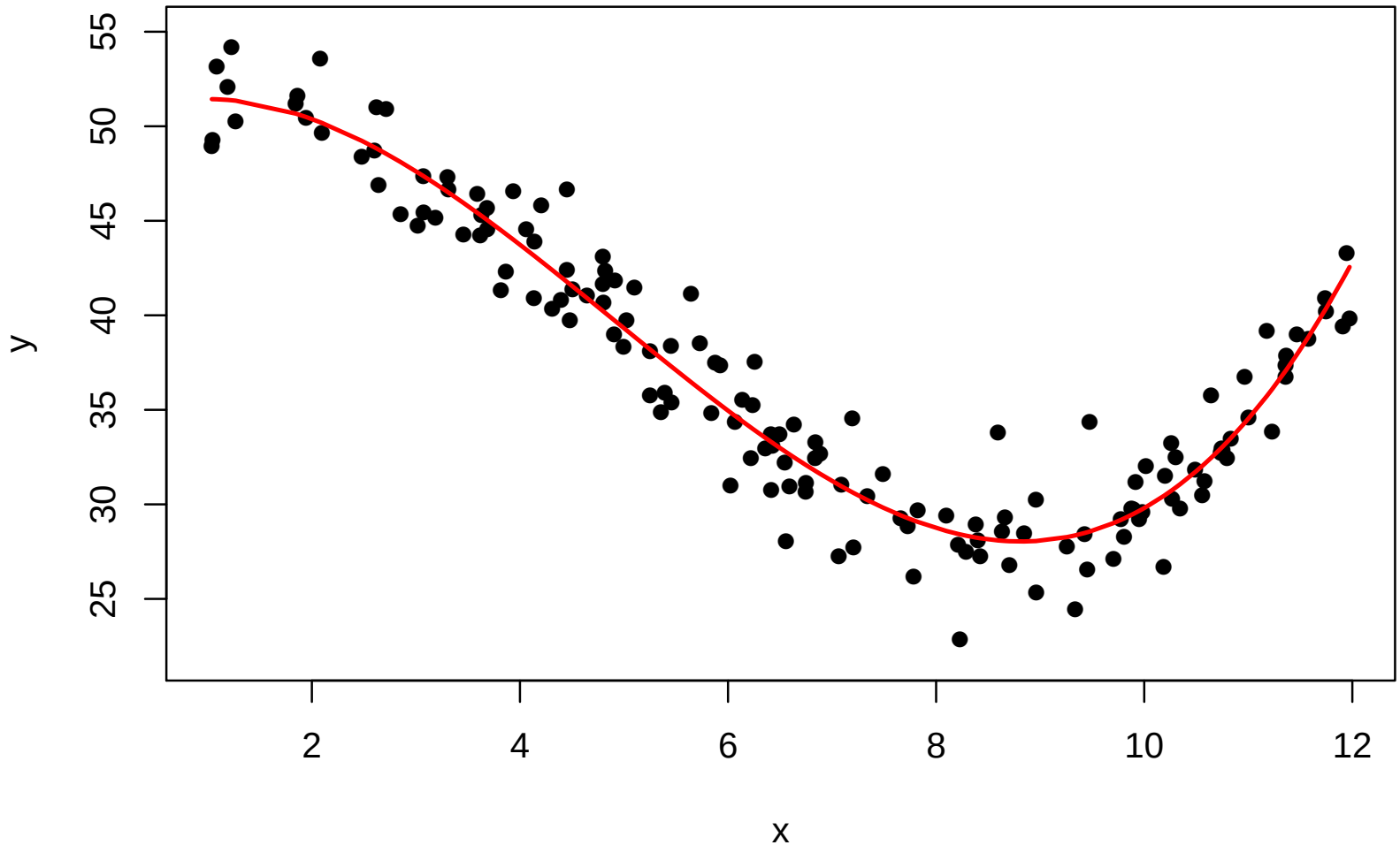
Deux noms différents pour un même concept ?

Les différences entre les deux approches résident surtout dans les objectifs de l'analyse.

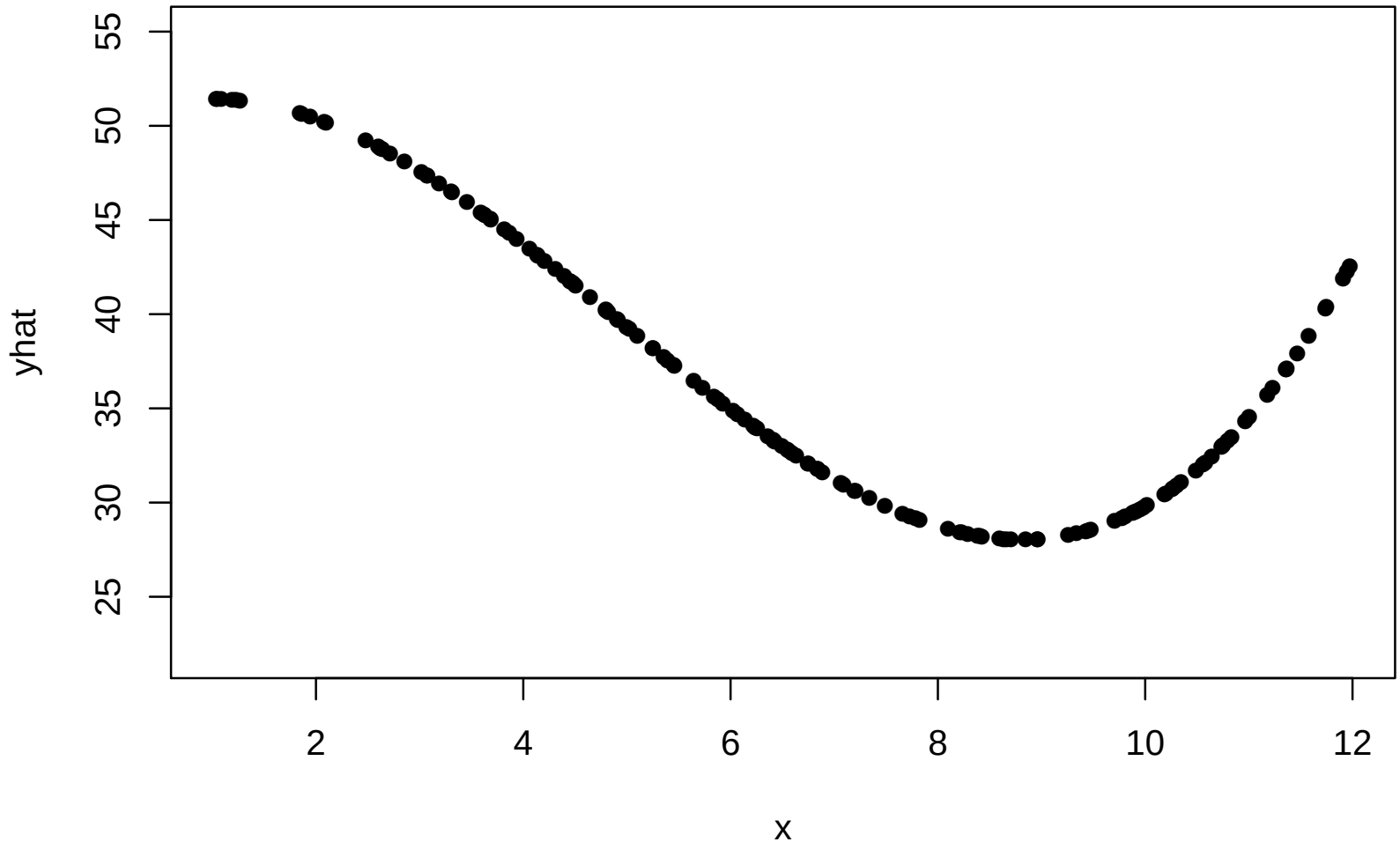
Objectifs de la modélisation statistique

1. But principal de l'analyse est d'estimer explicitement et interpréter l'effet des variables $x_1, \ldots, x_p$ sur Y.
2. On peut prédire Y à partir du modèle.
3. On veut parfois construire un modèle qui permet de simuler de nouvelles observations de Y ; on va s'intéresser aux termes de fluctuation aléatoire.
4. Des hypothèses a priori sur la forme de f sont requises, les données servent à estimer des morceaux inconnus de f et non pas f au complet.

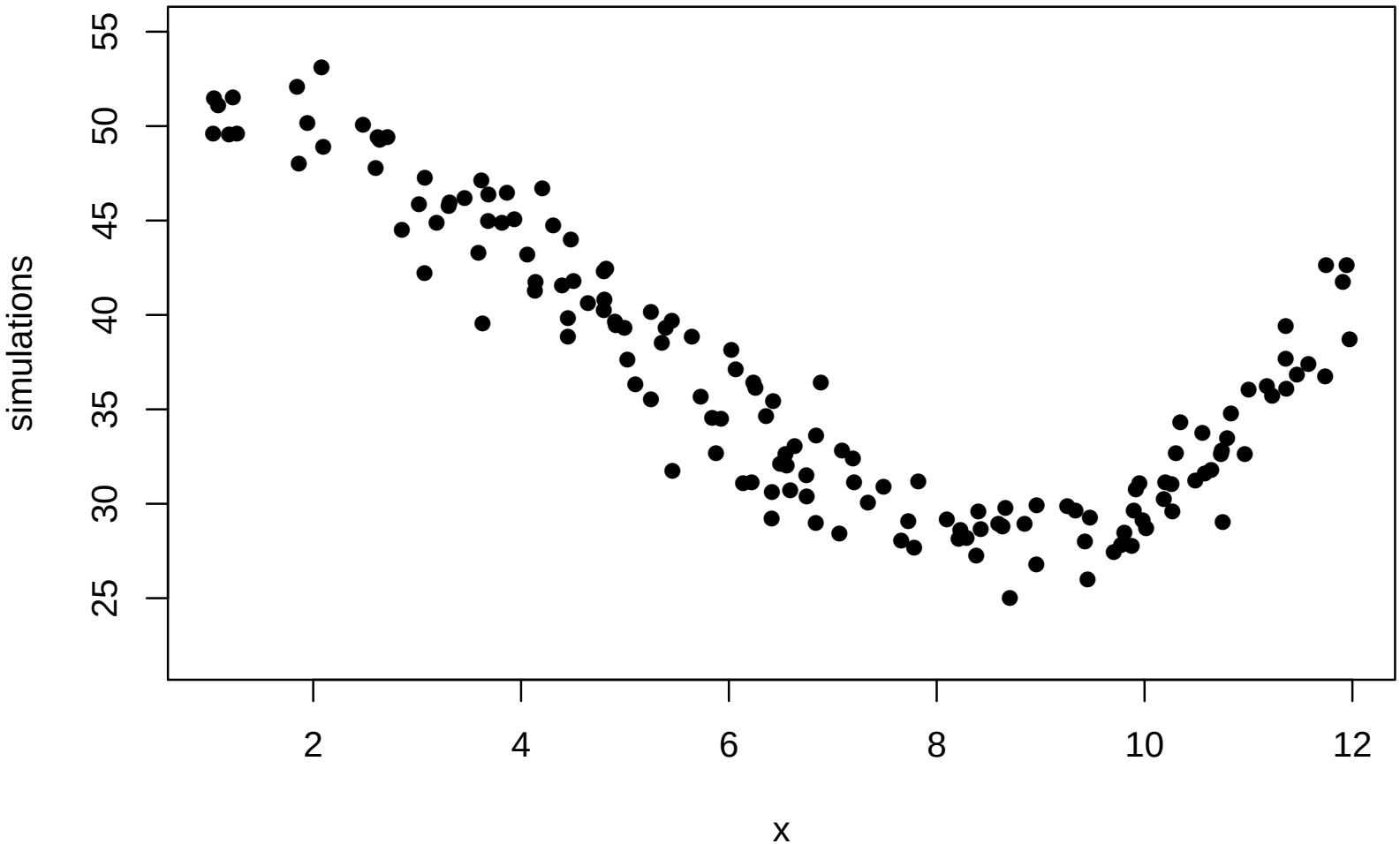
Prévision versus simulations



Prévisions (sans la fluctuation aléatoire)



Simulations (avec la fluctuation aléatoire)



Régression linéaire

Dans ce cours, on s'intéresse aux modèles de régression linéaire.

Dans un tel modèle, la valeur de la variable endogène Y est une fonction linéaire des paramètres :

$$Y = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \cdots + \beta_p f_p(x_p) + \text{fluctuation},$$

où $x_1, \ldots, x_p$ sont les variables exogènes,

$f_1, \ldots, f_p$ sont des transformations connues des covariables et

$\beta_0, \ldots, \beta_p$ sont des paramètres réels de valeur inconnue qu'on estimera à l'aide des données.

Questions pertinentes

- Est-ce raisonnable de supposer qu'il existe une relation linéaire entre Y et x_j ?
  Vérification du respect d'un certain nombre de postulats.

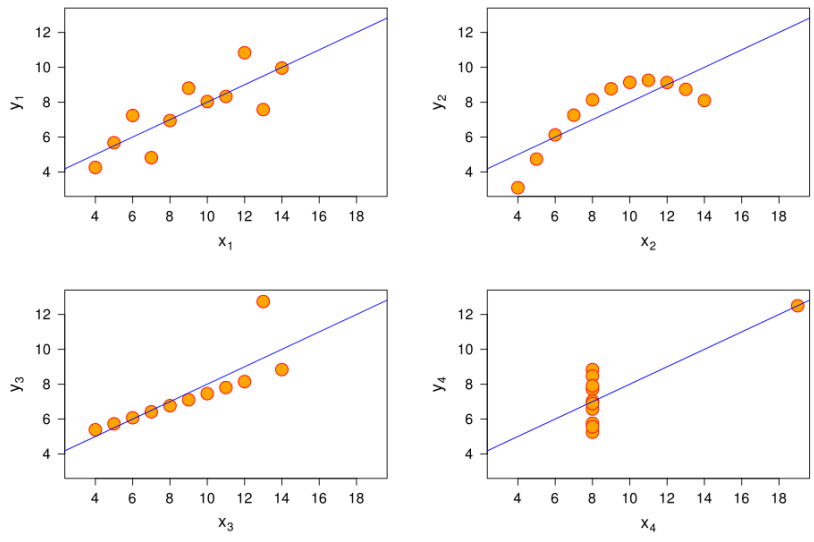
- Comment choisir les paramètres $\beta_0, \ldots, \beta_p$?
  Estimation des paramètres à l'aide de la méthode des moindres carrés.

- A-t-on vraiment besoin tous les paramètres $\beta_1, \ldots, \beta_p$ dans le modèle ? Sélection de variables.

- Comment effectuer une prévision ? Comment calculer l'erreur autour de cette prévision ?
  Intervalles de confiance pour les prévisions

- Comment mesurer l'effet d'une variable dans le modèle ?
  Interprétation des paramètres.

Première étape : tracer un nuage de points des données !



$\bar{x}$	s_x^2	$\bar{y}$	s_y^2	$\text{Corr}(x, y)$
9	10	7.5	3.75	0.816

Analyse préliminaire des données avec R

Mise en contexte, objectifs de l'analyse.

Graphiques univariés.

Le modèle et les postulats

En régression linéaire simple, on cherche à expliquer une variable endogène Y en fonction d'une seule variable exogène x.

Le modèle de régression linéaire simple est :

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

où $\beta_0, \beta_1 \in \mathbb{R}$ et

$\beta_0 =$ ordonnée à l'origine, paramètre de la régression (à estimer)

$\beta_1 =$ pente, paramètre de la régression (à estimer)

$x =$ variable exogène, explicative (pas aléatoire)

$\varepsilon =$ erreur ou fluctuation, variable aléatoire,

$Y =$ réponse, variable aléatoire.

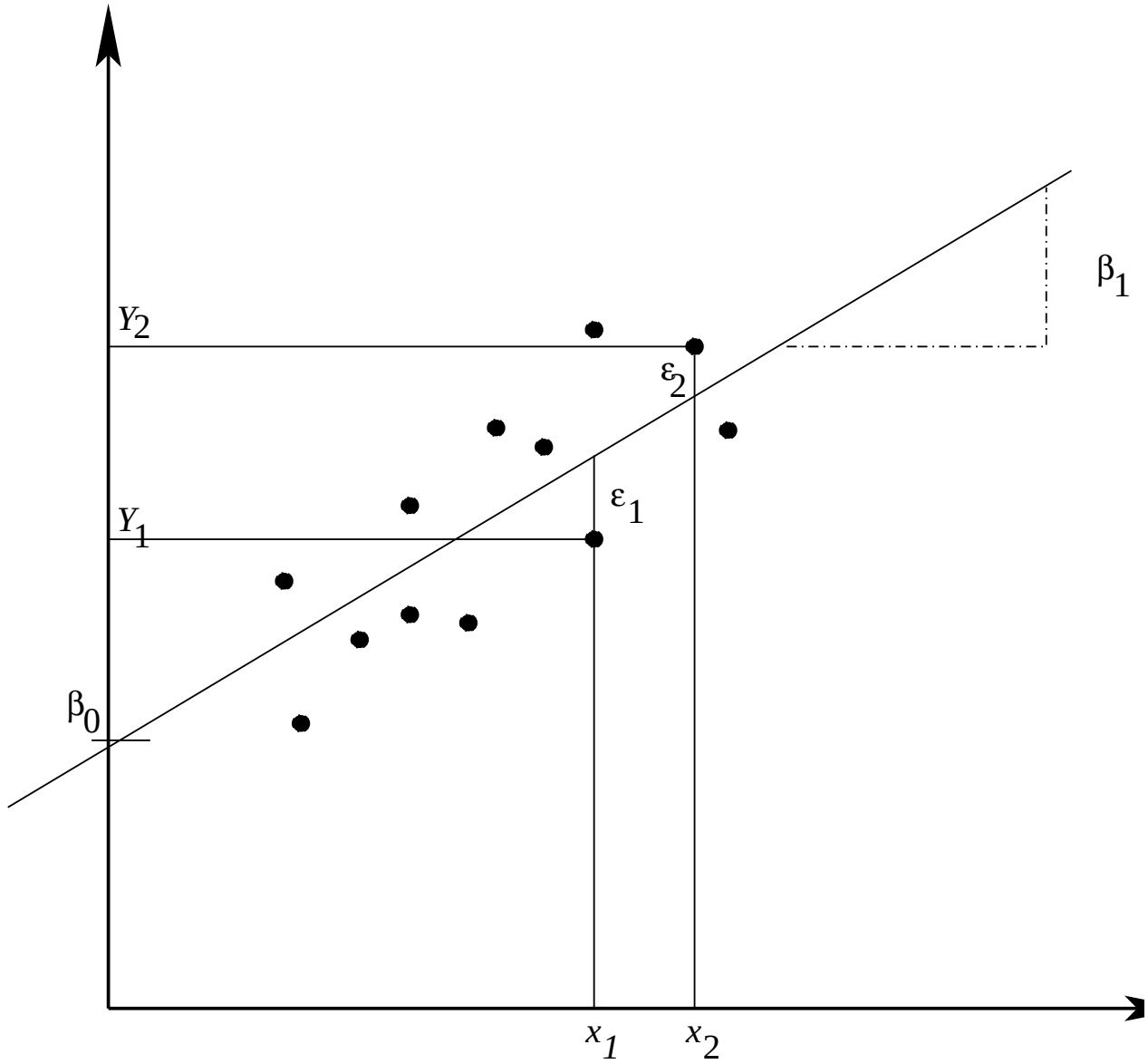
Le modèle et les postulats

Lorsqu'on a plusieurs observations, le modèle est

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1}$$

où $Y_1, \ldots, Y_n$ sont n observations de la variable endogène (variable réponse) et $x_1, \ldots, x_n$ sont n observations correspondantes de la variable exogène (variable explicative).

Schéma illustratif du modèle de régression linéaire simple.



Interprétations des paramètres

- L'ordonnée à l'origine β_0 est la valeur moyenne de Y lorsque $x = 0$.
- La pente β_1 est l'accroissement moyen de Y lorsque x augmente d'une unité.
- Si $\beta_1 = 0$, la distribution de Y ne dépend pas de la valeur prise par x.

Le modèle et les postulats

Pour pouvoir utiliser la régression linéaire, il est nécessaire de formuler certains **postulats** :

$\mathcal{H}_1$ Linéarité. $\mathrm{E}[\varepsilon_i] = 0$ pour $i = 1, \ldots, n$

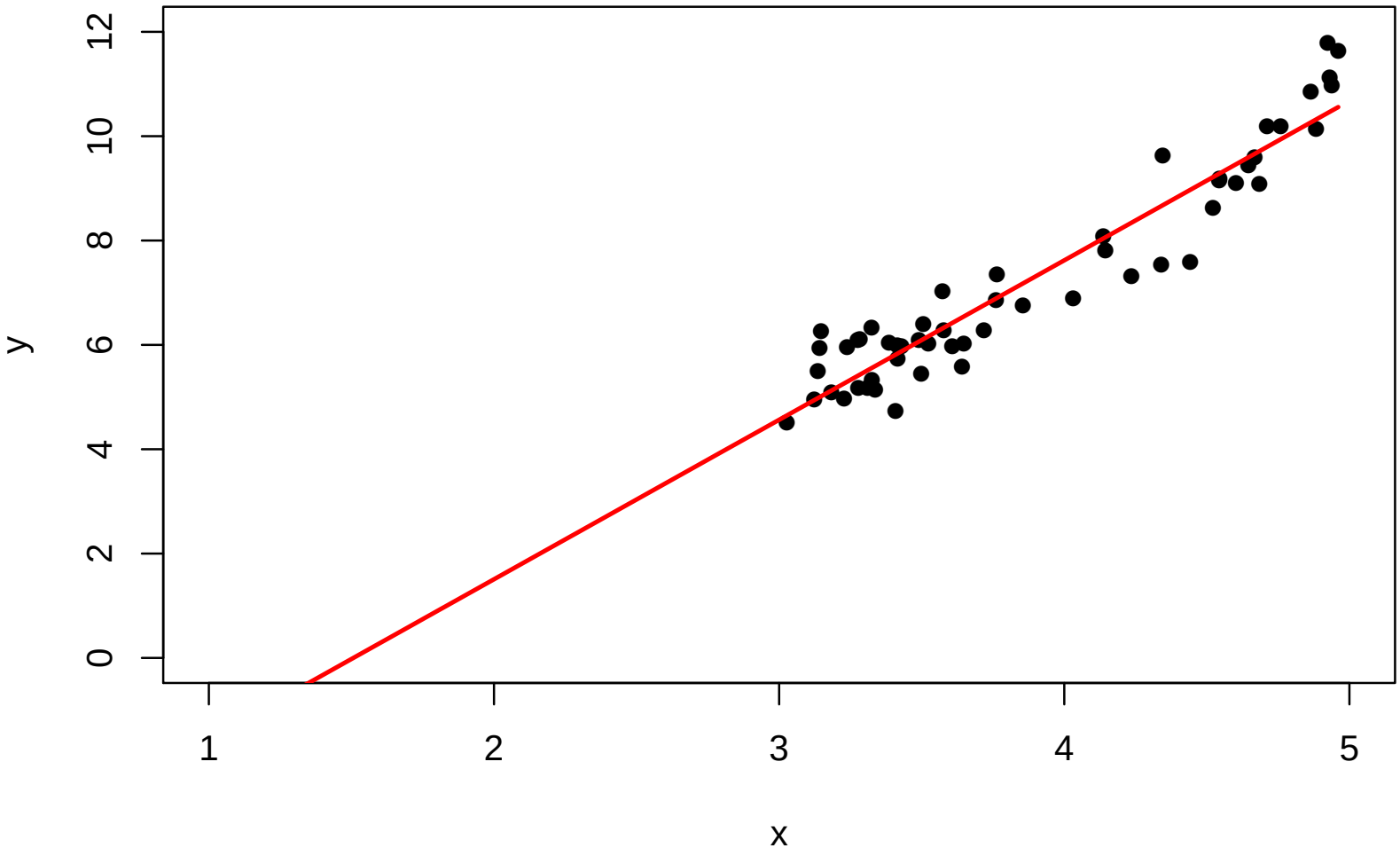
$\mathcal{H}_2$ Homoscédasticité. $\mathrm{var}[\varepsilon_i] = \sigma^2$ pour $i = 1, \ldots, n$

$\mathcal{H}_3$ Non-corrélation. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$, $i, j \in \{1, \ldots, n\}$

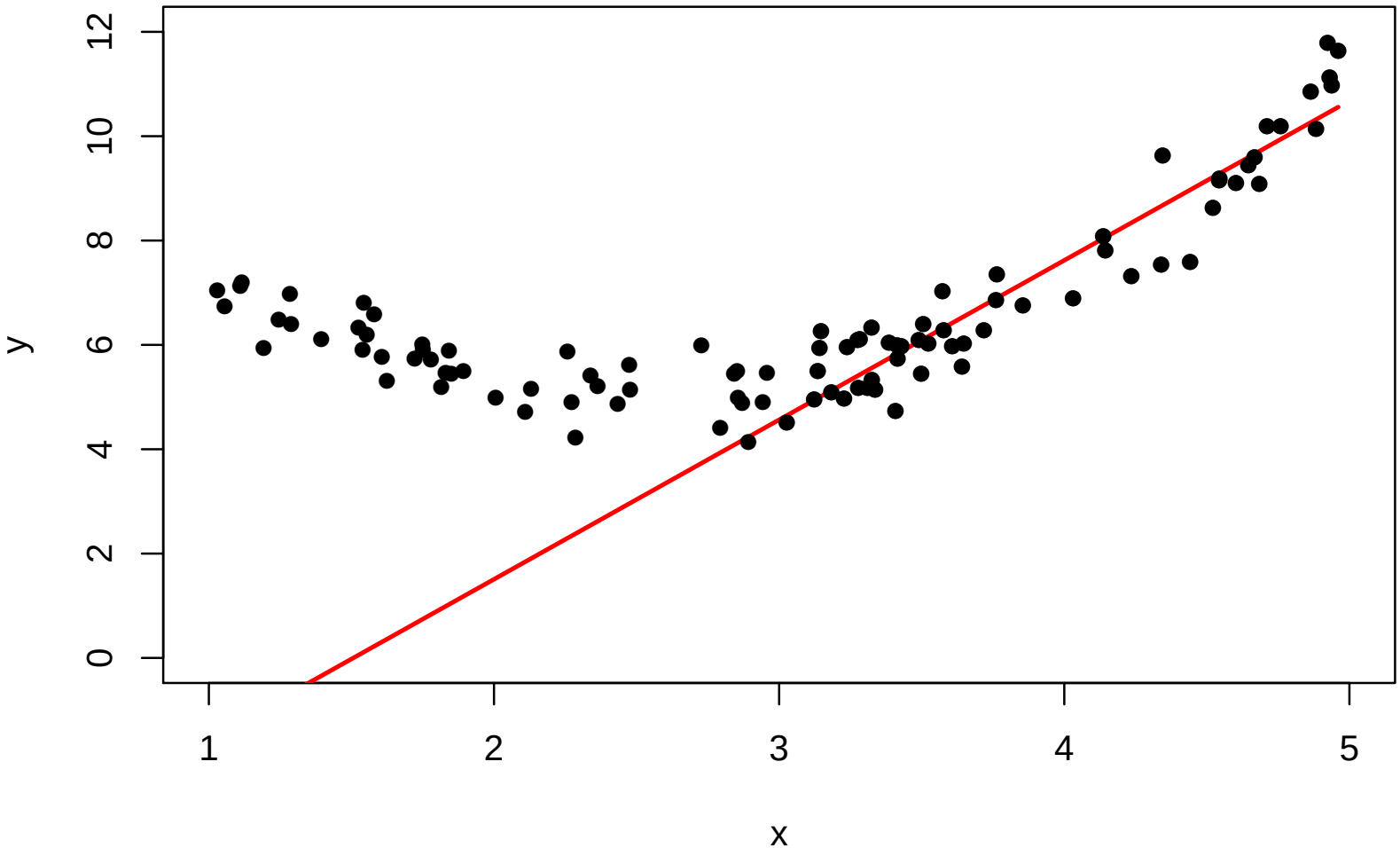
On doit parfois ajouter l'hypothèse de normalité :

$$\mathcal{H}_4. \ \varepsilon_1, \cdots, \varepsilon_n \sim N(0, \sigma^2).$$

Attention à l'extrapolation !



Attention à l'extrapolation !



Estimation : méthode des moindres carrés

On estime les paramètres β_0, β_1 par les valeurs qui minimisent la distance entre les observations (les points) et le modèle (la droite) :

$$\arg\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2.$$

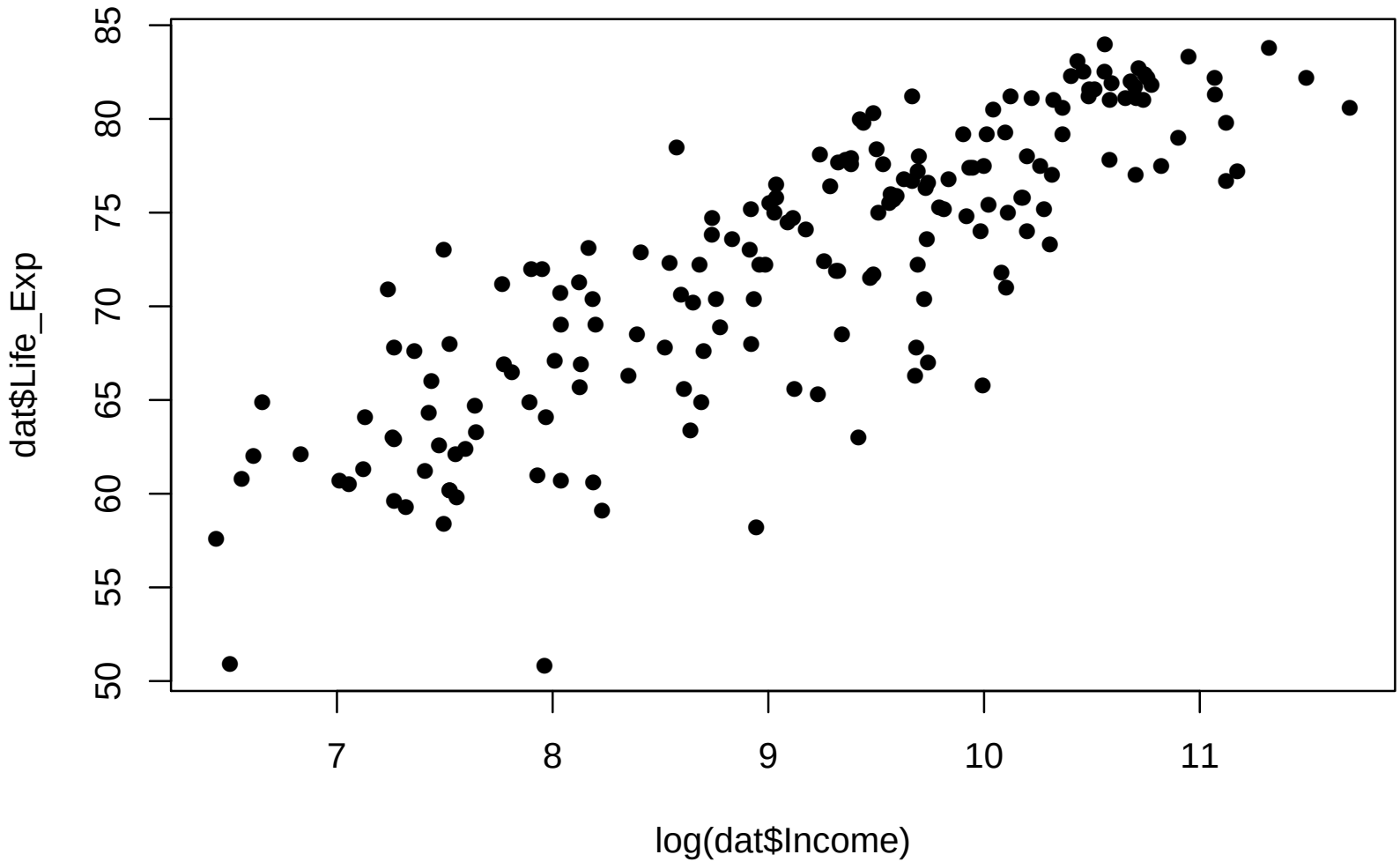
On trouve :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Les formules sont explicites !

Sous les postulats $\mathcal{H}_1 - \mathcal{H}_3$, les estimateurs obtenus ont de belles propriétés (sans biais, convergents).

Exemple : Espérance de vie en fonction du PIB



Exemple : Estimation avec R

```
mod1 <- lm(Life_Exp~I(log(Income)),data=dat)
mod1

##
## Call:
## lm(formula = Life_Exp ~ I(log(Income)), data = dat)
##
## Coefficients:
##    (Intercept)  I(log(Income))
##         26.736           4.993
```

Dans ce cas, l'expression de la droite de régression (et de la prévision) est

$$\hat{Y} = 26.736 + 4.993x.$$

Estimation de σ^2

On rappelle le postulat d'homoscédasticité :

$$\mathcal{H}_2 \ : \quad \text{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n.$$

Le paramètre σ^2 représente la variabilité de l'erreur autour de la droite, et son estimateur

$$s^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \frac{\text{SSE}}{n-p-1}$$

est sans biais, sous $\mathcal{H}_1$–$\mathcal{H}_4$.

Note : les formules sont présentées dans le cas général, p est le nombre de variables explicatives.

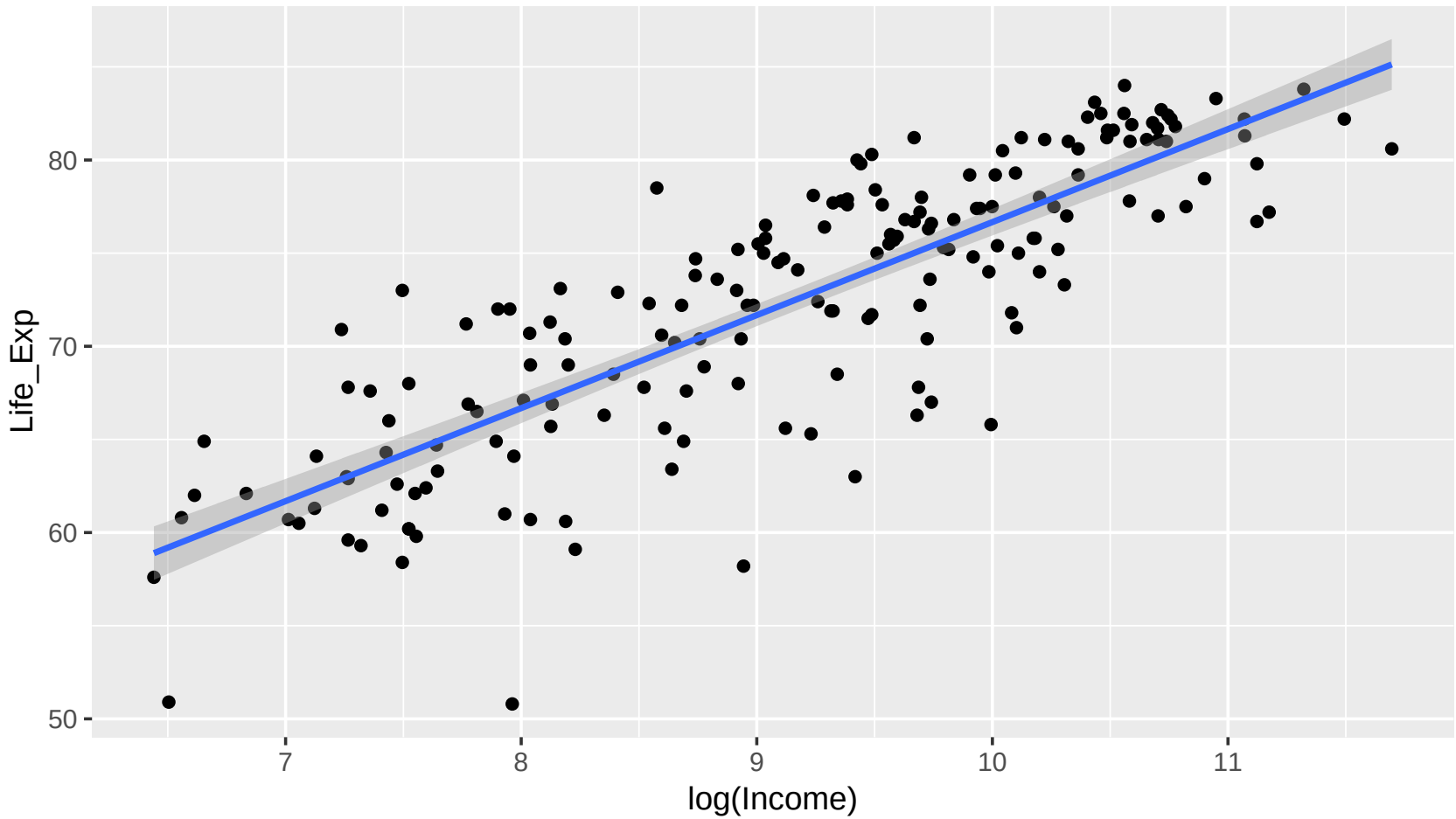
Exemple : Estimation de σ^2

```
summary(mod1)

##
## Call:
## lm(formula = Life_Exp ~ I(log(Income)), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6903  -2.2286   0.7689   2.7357   8.9470
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     26.7358     2.2575   11.84   <2e-16 ***
## I(log(Income))   4.9930     0.2445   20.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 4.053 on 182 degrees of freedom
## Multiple R-squared:  0.6961,Adjusted R-squared:  0.6944
## F-statistic: 416.9 on 1 and 182 DF,  p-value: < 2.2e-16
```

Exemple : Illustration

```
library(ggplot2)
ggplot(dat,aes(x=log(Income),y=Life_Exp))+
  geom_point()+geom_smooth(method='lm',formula=y~x)
```



Ce modèle est-il bon ?

Un bon modèle de régression devrait expliquer une bonne partie de la variabilité dans les variables réponses.

Le coefficient de détermination

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

est la proportion de variabilité dans les Y expliquée par le modèle.

On a que $0 \leq R^2 \leq 1$ et on cherche à maximiser ce critère.

Exemple : Coefficient de détermination

```
summary(mod1)

##
## Call:
## lm(formula = Life_Exp ~ I(log(Income)), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6903  -2.2286   0.7689   2.7357   8.9470
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     26.7358     2.2575   11.84   <2e-16 ***
## I(log(Income))   4.9930     0.2445   20.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 4.053 on 182 degrees of freedom
## Multiple R-squared:  0.6961,Adjusted R-squared:  0.6944
## F-statistic: 416.9 on 1 and 182 DF,  p-value: < 2.2e-16
```

Régression linéaire simple avec R

Ajustement du modèle.

Interprétations des paramètres et calcul du R^2.

Et si on regardait le délai en heures plutôt qu'en minutes ?

Les paramètres β_0 et β_1 s'ajustent automatiquement : les unités n'importent pas du tout.

Morale : il ne faut jamais juger si un paramètre est "important" ou "grand" seulement par sa valeur absolue.

Test sur les paramètres

On teste si l'effet d'une variable x_j sur Y est significatif à l'aide d'un test d'hypothèse

$$H_0 : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0.$$

La statistique du test est

$$t = \frac{\hat{\beta}_j}{\text{erreur standard de } \hat{\beta}_j}.$$

On conclut que $\beta_j \neq 0$ lorsque t prend une valeur qui est trop loin de 0.

Sous les postulats $\mathcal{H}_1$–$\mathcal{H}_4$, le seuil observé (p-value) du test peut être calculé avec la loi Student avec $n - p - 1$ degrés de libertés. On rejette $H_0 : \beta_j = 0$ au niveau de confiance $1 - \alpha$ lorsque le seuil observé est inférieur à α.

Exemple : Test sur les paramètres

```
summary(mod1)

##
## Call:
## lm(formula = Life_Exp ~ I(log(Income)), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6903  -2.2286   0.7689   2.7357   8.9470
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     26.7358     2.2575   11.84   <2e-16 ***
## I(log(Income))   4.9930     0.2445   20.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 4.053 on 182 degrees of freedom
## Multiple R-squared:  0.6961,Adjusted R-squared:  0.6944
## F-statistic: 416.9 on 1 and 182 DF,  p-value: < 2.2e-16
```

Plusieurs types de variables

En pratique, plusieurs variables explicatives peuvent influencer Y. On peut aussi s'intéresser à l'effet de plusieurs types de variables sur Y, à savoir des variables :

- dichotomiques (ex. présence/ absence de précipitations)
- discrètes (ex. une classe de revenu)
- continues (ex. la taille de la population, le nombre de sièges dans l'avion)
- qualitatives (ex. le continent, l'aéroport de départ).

La régression linéaire multiple

L'équation du modèle de régression linéaire multiple est

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

où, pour $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, p\}$,

- Y_i est l'observation i de la variable endogène et est aléatoire.
- x_{ij} dénote la covariable j pour l'observation i, qui est connue, non aléatoire.
- β_0 et β_j sont les paramètres du modèle. Ils sont inconnus et par conséquent doivent être estimés.
- ε_i est le terme d'erreur pour l'observation i et est une variable aléatoire inconnue.

Modèle et notation

Dans sa version matricielle, le modèle s'écrit comme suit :

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}_{n\times 1}} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{\boldsymbol{X}_{n\times(p+1)}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}_{(p+1)\times 1}} + \underbrace{\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}_{n\times 1}}$$

soit

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon.$$

On a que $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$ est la ie ligne de $\boldsymbol{X}$ et, pour $i = 1, \ldots, n$,

$$Y_i = \boldsymbol{x}_i\beta + \varepsilon_i.$$

Estimation

L'estimation des paramètres par la méthode des moindres carrés donne encore une formule explicite :

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

Le nombre d'observations doit être supérieur au nombre de paramètres afin d'être en mesure d'estimer ces derniers adéquatement.

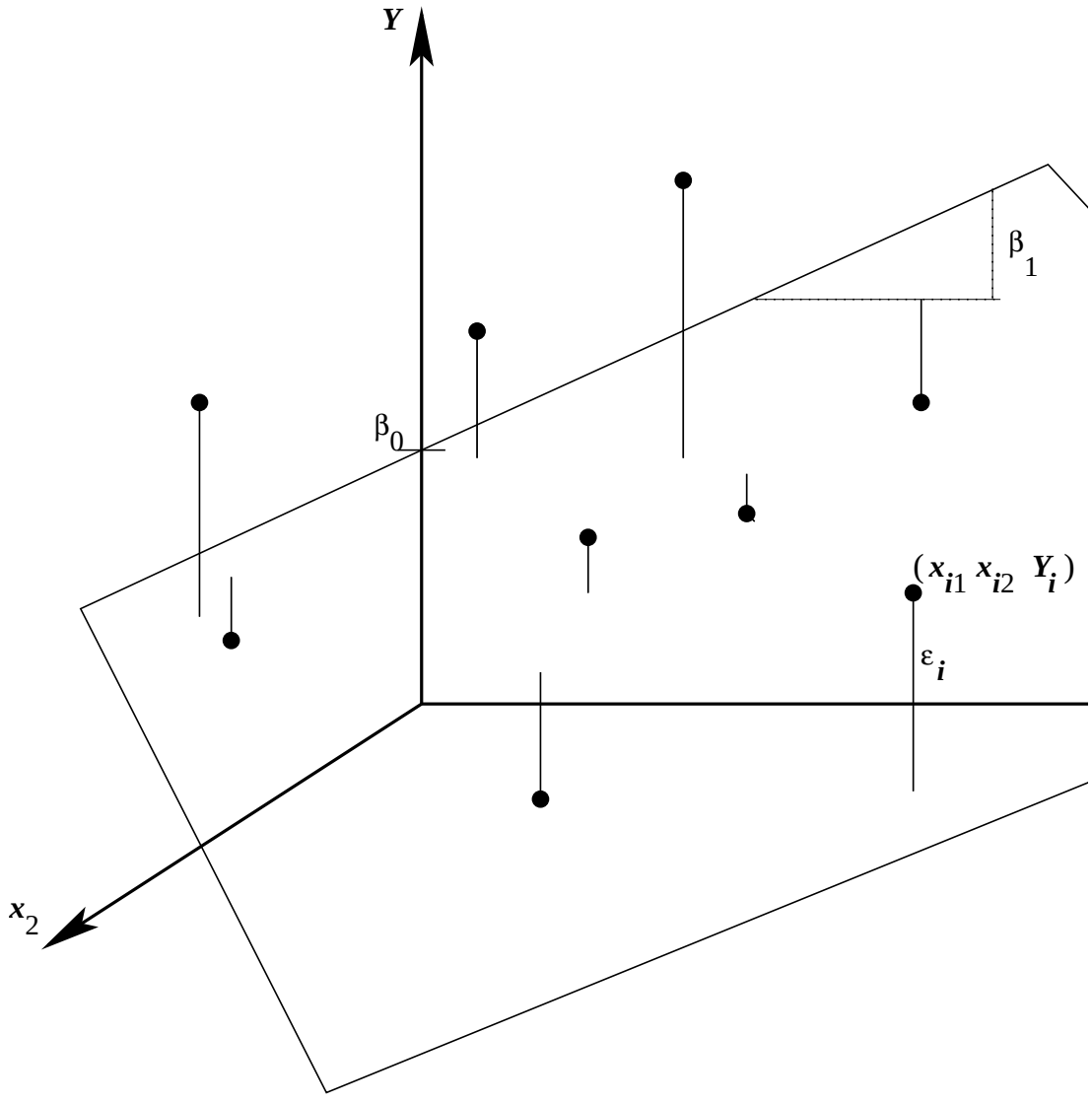
Il ne doit pas y avoir de variables redondantes (ex : taux de chômage et taux d'emploi). La plupart du temps, dans les données réelles, cela cause problème, surtout dans le cas de "redondances approximatives".

Interprétation des paramètres

Les observations Y_i sont placées autour d'un hyperplan, et le coefficient de régression β_j représente la pente dans la direction de la variable explicative x_j.

- β_0 est la moyenne de Y lorsque toutes les variables explicatives sont égales à 0 simultanément.
- β_j est l'augmentation de la moyenne de Y lorsque la $j$ième variable explicative x_j est augmentée de 1 unité et que toutes les autres variables explicatives sont inchangées

Représentation géométrique du modèle



Traitement des variables qualitatives ou catégorielles

Certaines variables exogènes catégorielles ne sont pas ordinales.
Quelques exemples :

$$x_{i1} = \begin{cases} H, & \text{si } i \text{ est un homme} \\ F, & \text{si } i \text{ est une femme} \end{cases}$$

$$x_{i1} = \begin{cases} 1, & \text{si } i \text{ est une ville} \\ -1, & \text{si } i \text{ est un village} \end{cases}$$

$$x_{i1} = \begin{cases} 1, & \text{si } i \text{ est à ON} \\ 0, & \text{si } i \text{ est à OFF} \end{cases}$$

Régression avec variables qualitatives

Dans ces cas, il est toujours plus prudent de coder x_i sous la forme d'une variable indicatrice, c'est-à-dire sous la forme

$$x_i = \begin{cases} 1, & \text{si } \ldots \\ 0, & \text{sinon} \end{cases}$$

Si ce type de codage n'est pas utilisé, alors les β peuvent être difficiles à interpréter. Par exemple si

$$x_i = \begin{cases} -1, & \text{si } i \text{ est un homme} \\ 1, & \text{si } i \text{ est une femme,} \end{cases}$$

alors on a que le modèle $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ nous donne

$$\begin{aligned} \mathrm{E}[Y_i; i \text{ est un homme}] &= \beta_0 - \beta_1 \\ \mathrm{E}[Y_i; i \text{ est une femme}] &= \beta_0 + \beta_1 \end{aligned}$$

$\Rightarrow \beta_1$ représente 0.5 fois la différence entre la valeur moyenne de Y pour les hommes et celle pour les femmes.

Régression avec variables qualitatives

Si on prend plutôt

$$x_i = \begin{cases} 1, & \text{si } i \text{ est un homme} \\ 0, & \text{si } i \text{ est une femme} \end{cases}$$

alors on a

$$\begin{aligned} \mathrm{E}[Y_i; i \text{ est un homme}] &= \beta_0 + \beta_1 \\ \mathrm{E}[Y_i; i \text{ est une femme}] &= \beta_0 \end{aligned}$$

et maintenant β_1 représente exactement la différence entre la valeur moyenne de Y pour les hommes et celle pour les femmes.

Exemple avec variable polytomique

Soit

- x_{i1}, le numéro de lot du produit i,
- x_{i2}, la concentration de sel dans le produit i et
- Y_i, l'indice de qualité du produit i.

Les variables Y_i et x_{i2} sont des variables continues, alors que x_{i1} est un variable polytomique prenant une des valeurs $\{1, 2, 3, 4\}$. On considère le modèle

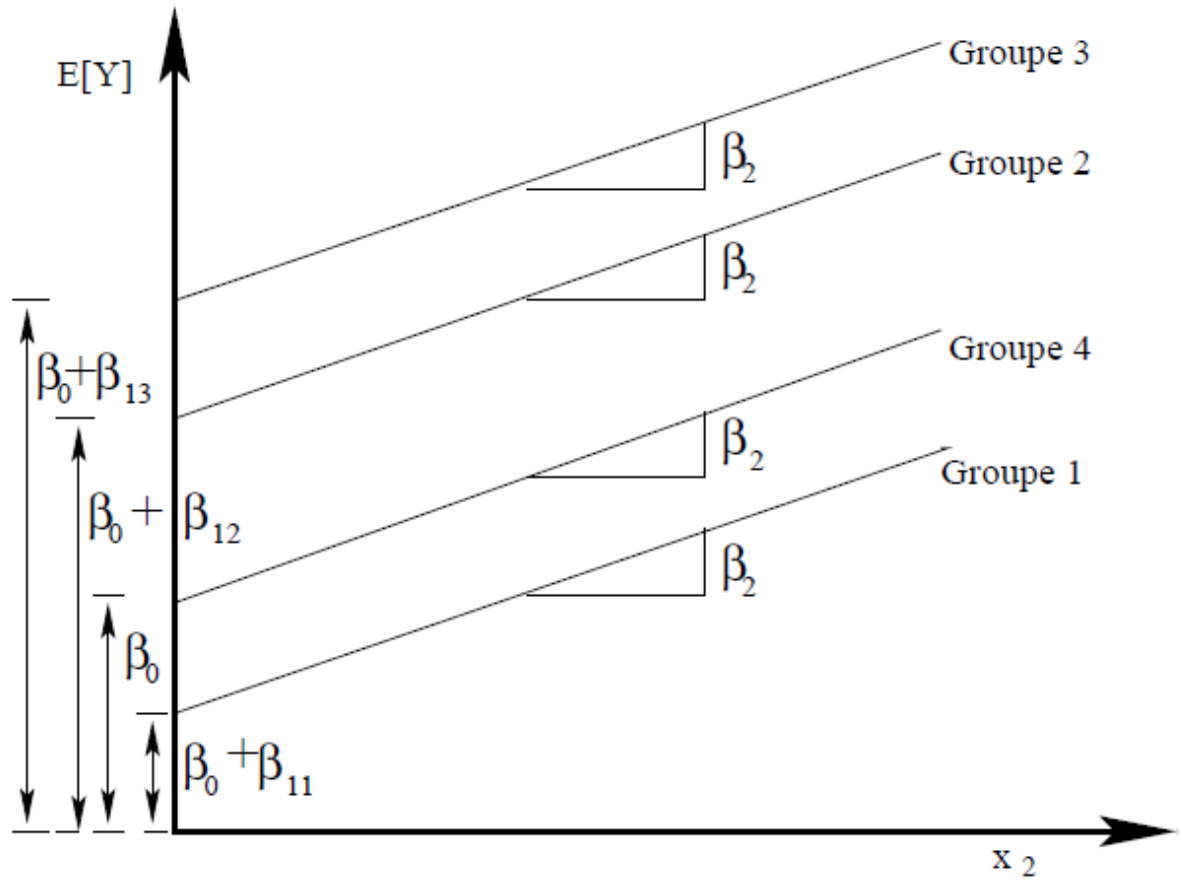
$$Y_i = \beta_0 + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_2 x_{i2} + \varepsilon_i,$$

où

$$x_{i11} = \begin{cases} 1, & x_{i1} = 1 \\ 0, & x_{i1} \neq 1 \end{cases} \qquad x_{i12} = \begin{cases} 1, & x_{i1} = 2 \\ 0, & x_{i1} \neq 2 \end{cases}$$

$$x_{i13} = \begin{cases} 1, & x_{i1} = 3 \\ 0, & x_{i1} \neq 3. \end{cases}$$

Exemple avec variable polytomique



Relations plus complexes

Le modèle est linéaire en termes des paramètres $\beta_1, \ldots, \beta_p$, pas nécessairement en terme des variables exogènes.

Exemples :

1. $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
2. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

Par convention, si on inclut x^2 dans un modèle, on inclut aussi x. Si une variable est contenue dans une interaction $x_1 x_2$, on inclut aussi les effets principaux x_1 et x_2.

Prévisions

La prévision de Y pour une valeur donnée $\boldsymbol{x}^*$ est

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^* = \boldsymbol{x}^* \hat{\boldsymbol{\beta}}.$$

La vraie valeur de $Y^* = \boldsymbol{x}^* \hat{\boldsymbol{\beta}} + \varepsilon^*$ sera très probablement différente de notre prévision $\hat{Y}^*$:

- Les paramètres sont estimés, donc il y a une incertitude sur les valeurs de β_0 et β_1.
- Le terme d'erreur ε^* est centré à 0 mais a une certaine variance σ^2.

Inférence sur la valeur moyenne $\mathrm{E}[Y|\mathbf{x}^*]$

On veut estimer

$$\mathrm{E}[Y|\mathbf{x}^*] = \mathbf{x}^*\beta.$$

Un intervalle de confiance de niveau $100(1-\kappa)\%$ pour $\mathrm{E}[Y|\mathbf{x}^*]$ est

$$\left[\mathbf{x}^*\hat{\beta} \pm t_{n-p-1}(1-\kappa/2)\sqrt{s^2\mathbf{x}^*(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{*\top}}\right].$$

On l'obtient avec la fonction `predict` en R, avec l'argument `interval="confidence"`.

Inférence sur une prévision de Y étant donné $\mathbf{x}^*$

On peut aussi s'intéresser à la valeur même d'une réalisation de la variable endogène pour une combinaison de valeurs $\mathbf{x}^*$ fixées des variables exogènes.

Un intervalle de confiance à $100(1-\kappa)\%$ pour Y^* étant donnée la combinaison des variables exogènes $\mathbf{x}^*$ est donné par

$$\left[\mathbf{x}^*\hat{\boldsymbol{\beta}} \pm t_{n-p-1}(1-\kappa/2)\sqrt{s^2\left(1+\mathbf{x}^*(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{*\top}\right)}\right].$$

Cet intervalle est donné avec l'argument `interval="prediction"` de la fonction `predict`.

Exemple de prévisions

```
predict(mod1, newdata=data.frame(Income=10000),
        interval="confidence", level=0.95)

##        fit      lwr      upr
## 1 72.72293 72.13274 73.31311

predict(mod1, newdata=data.frame(Income=10000),
        interval="prediction", level=0.95)

##        fit      lwr      upr
## 1 72.72293 64.70523 80.74062
```

Régression linéaire multiple avec R

Ajustement du modèle.

Interprétations des paramètres, tests d'hypothèses et prévisions.

Importance de la sélection des variables

- Toutes les variables confondantes (reliées à la fois Y et à au moins une variable explicative incluse dans le modèle) doivent être incluses dans le modèle afin que les prévisions soient sans biais.

- Inclure trop de variables explicatives qui sont inutile mène à de grandes erreurs standards des estimations et du sur-ajustement.

Il faut donc choisir un sous-ensemble approprié des variables explicatives à inclure dans le modèle.

Sélection de sous-modèles

Lorsqu'il n'y a pas trop de variables explicatives, l'approche à privilégier est d'ajuster tous les sous-modèles et de choisir le meilleur selon les critères de comparaison de modèle.

Critère d'information d'Akaike

Ce critère est souvent utilisé dans la pratique et permet de comparer la qualité de l'ajustement de modèles :

$$\text{AIC} = n \ln(SSE/n) + 2(p+1).$$

- Un modèle qui s'ajuste bien sur les observations est associé à une faible valeur de SSE.
- La complexité du modèle est prise en compte en ajoutant le double du nombre de paramètres.

Un bon modèle est donc associé à une faible valeur d'AIC.

Critère d'information bayésien de Schwarz

Gideon E. Schwarz a développé son critère BIC à partir d'une argumentation bayésienne :

$$\text{BIC} = n \ln(SSE/n) + (p+1) \ln(n).$$

Le critère BIC est semblable au AIC, mais la pénalité pour le nombre de paramètres dépend de la taille de l'échantillon.

Un bon modèle est associé à une faible valeur de BIC.

Autres critères de comparaison et sélection de modèle

- Le coefficient de détermination ajusté :

$$R^2_{\text{a}} = 1 - \frac{SSE/(n-p-1)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)}$$

  Lorsque R^2_a est plus proche de 1, le modèle a un meilleur ajustement.

- Les critères basés sur la puissance de prévision.

- Le Cp de Mallows.

Régression régularisée (LASSO)

Lasso signifie *Least Absolute Shrinkage and Selection Operator*. Plutôt que de minimiser *SSE*, la fonction à minimiser inclut un terme de pénalité sur la norme $L1$ du vecteur de coefficients :

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|.$$

L'hyperparamètre $\lambda > 0$ est calibré avec une validation croisée.

La forme de cette pénalité est pratique pour la sélection de variables, puisque certains coefficients estimés seront exactement 0.

Régression régularisée (LASSO)

Cela est équivalent à la minimisation

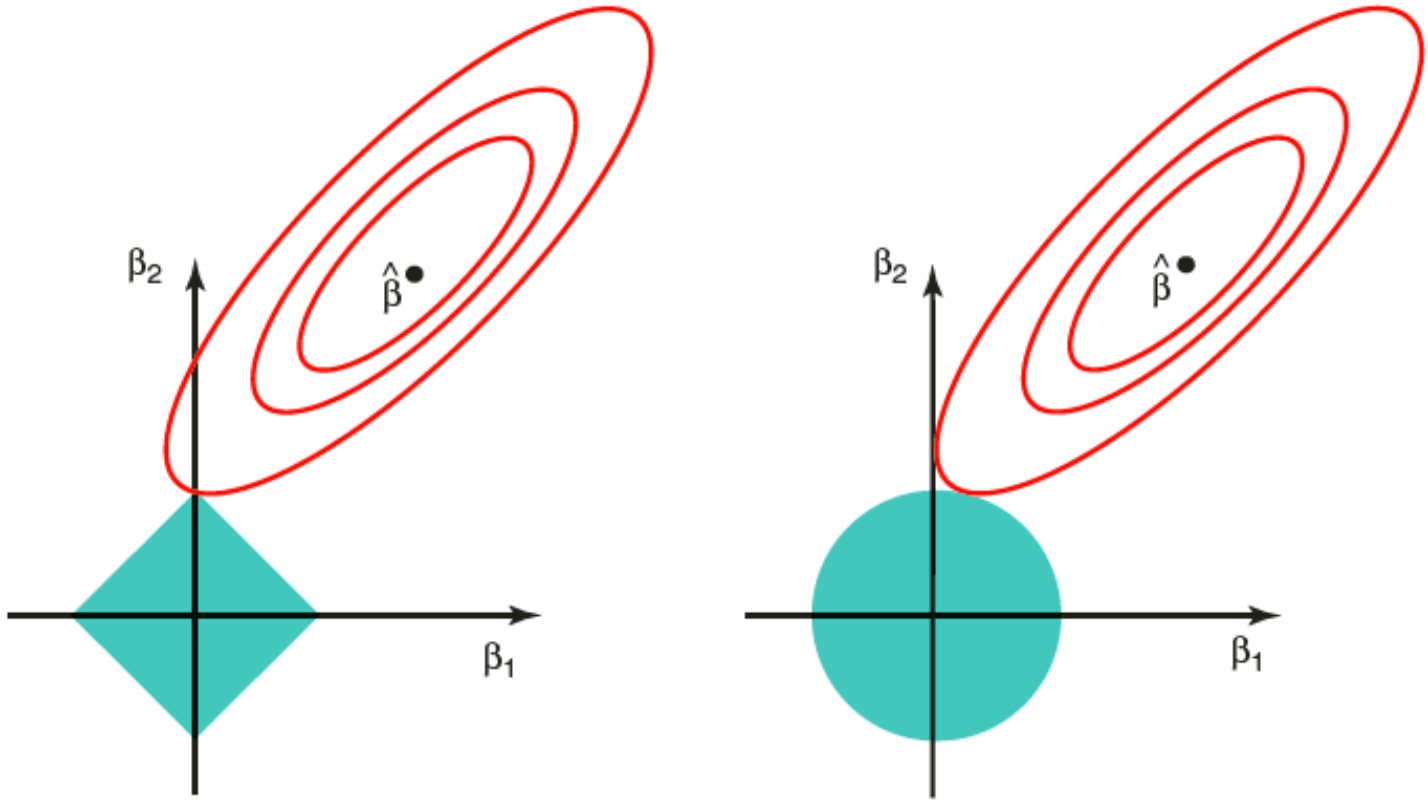
$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

sous contrainte que

$$\sum_{j=1}^{p} |\beta_j| \leq t,$$

où t est le "budget total" pour les coefficients et dépend du paramètre d'ajustement $\lambda > 0$.

Illustration de la régression régularisée



Source de l'image : James et coll. (2013)

Validation du modèle

- La validité des estimations, inférences et prévisions repose sur la vérification des postulats $\mathcal{H}_1 - \mathcal{H}_4$.
- Il existe de nombreux diagnostiques basés sur les résidus $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Multicollinéarité

- S'il y a de l'information redondante dans les variables explicatives $\boldsymbol{X}^\top \boldsymbol{X}$ n'est pas inversible et cela cause problème pour l'estimation. Par contre, on peut simplement enlever la ou les colonnes superflues.

- Si l'information est " presque" redondante, alors c'est plus difficile à détecter, mais $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ est instable et les variances des estimations sont artificiellement gonflées.

- La régression régularisée de type Ridge permet de limiter les problèmes liés à la multicollinéarité.

Exemple en R

Sélection de variables et Lasso.

Généralisations

- Lorsque Y est discrète ou non-normale : modèle linéaire généralisé
$$g\{\mathrm{E}(Y|\boldsymbol{x})\} = \boldsymbol{x}\boldsymbol{\beta},$$
pour une fonction de lien g appropriée.

- Lorsque l'effet de x sur Y n'est pas linéaire mais est lisse : modèle additif généralisé.

- Lorsque le postulat d'homoscédasticité n'est pas vérifié : régression pondérée, GLM ou GAMLSS.

Références

- James G., Witten, D., Hastie, T. & Tibshirani R. (2013). *An Introduction to Statistical Learning (with applications in R).* Springer, New York. Diponible gratuitement en PDF : `http://www-bcf.usc.edu/~gareth/ISL/`.

- Montgomery, D.C., Peck, E.A., Vining, G.,G. (2012). *Introduction to linear regression analysis.* 5e edition.