Régression et corrélation

LMAFY1101

Anouar El Ghouch

LSBA, UCLouvain

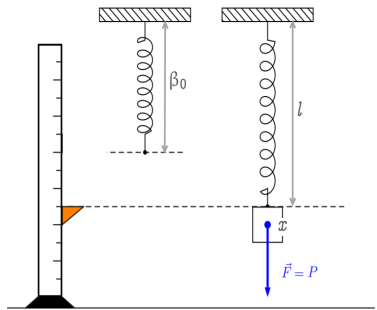
Plan

Modélisation déterministe vs modélisation statistique

Diverses lois physiques impliquent qu'une grandeur soit proportionnelle à une autre.

La loi de Hooke, par exemple, stipule que l'allongement d'un ressort varie proportionnellement à la force qu'il subit.

On se propose de vérifier cette loi à l'aide de l'expérience suivante : une masse, pour laquelle on peut facilement calculer le poids (x), est accrochée à un ressort suspendu verticalement. La masse étire le ressort par une force qui est tout simplement égale à son poids. On mesure par la suite la longueur (l) du ressort étiré.



La loi de Hooke peut-être alors exprimée sous la forme suivante

$$l - \beta_0 = \beta_1 x,$$

où β_0 est la longueur du ressort sans charge et β_1 est la constante de raideur du ressort.

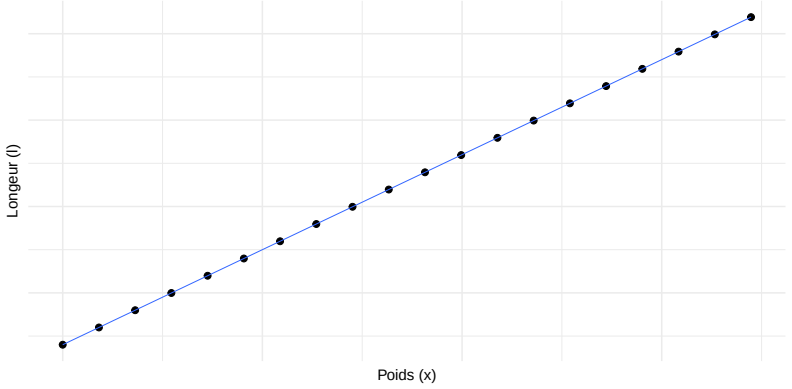
À l'aide de différentes masses, on a répété cette expérience 20 fois et on a obtenu les chiffres suivants

Poids (g)	Longueur mesurée (cm)	Poids (g)	Longueur mesurée (cm)
0.00	12.85	907.18	13.72
90.72	12.73	997.90	14.15
181.44	13.00	1088.62	13.89
272.16	13.03	1179.34	14.05
362.87	13.06	1270.06	14.25
453.59	13.11	1360.78	14.20
544.31	13.34	1451.50	14.25
635.03	13.18	1542.21	14.61
725.75	13.31	1632.93	14.43
816.47	13.87	1723.65	14.73

Si on note par $x_1, \ldots, x_n$ les poids des masses et par l_i la longueur du ressort sous la charge x_i alors, selon la loi de Hooke,

$$l_i - \beta_0 = \beta_1 x_i, \ \forall i$$

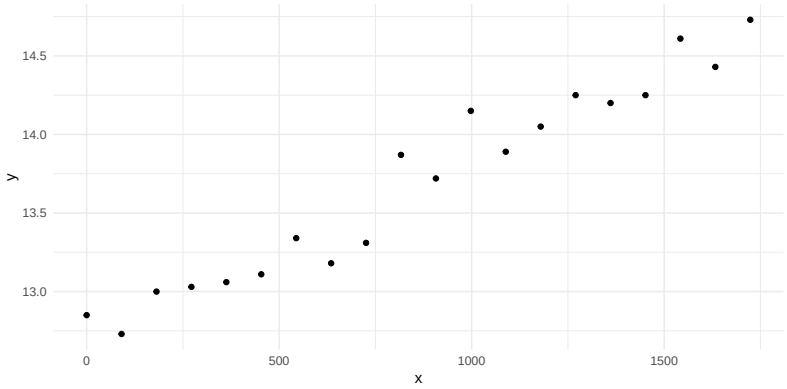
Cette équation est un exemple d'une modélisation déterministe. Graphiquement, si nous tracions ces valeurs, cela doit se traduire par des points qui s'alignent exactement selon une droite comme dans la figure suivante :



Or voici ce qu'on obtient à partir des données récoltées

```
> x<-c(0.00,90.72,181.44,272.16,362.87,453.59,544.31,635.03,725.75,816.47,907.18,997.90,1088.62,1179.34,1270.06,1360.78,1451.50,
  1542.21,1632.93,1723.65)
> y<-c(12.85,12.73,13.00,13.03,13.06,13.11,13.34,13.18,13.31,13.87,13.72,14.15,13.89,14.05,14.25,14.20,14.25,14.61,14.43,14.73)
> mdt<-data.frame(x,y)
```

```
> plot(y ~ x, data = mdt)
```



Soit Y_i la longueur réellement mesurée du ressort sous la charge x_i. En raison d'erreurs de mesure, Y_i différera de la longueur réelle l_i. Nous écrivons

$$Y_i = l_i + \epsilon_i$$

En combinant les deux dernières équations, on obtient [1]

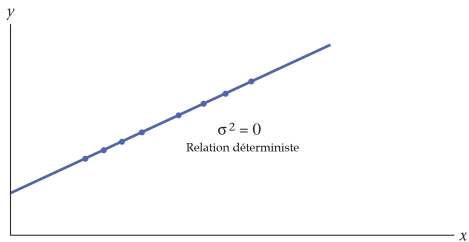
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

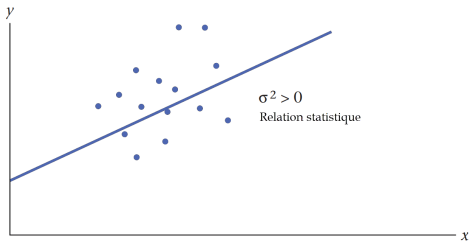
Cette équation est un exemple d'une modélisation statistique qui, à l'inverse de la modélisation déterministe, prend en compte les variations aléatoires (dues, par exemple, aux erreurs de mesure) à travers le terme d'erreur (variable aléatoire) ϵ_i.

Cette variable est souvent supposée Normale, de moyenne 0 et d'une variance σ^2 inconnue.

[1] Nous supposons ici que les incertitudes sur les poids sont négligeables. Mais même si ce n'est pas le cas, la modélisation statistique reste inchangée.

Le graphique suivant illustre la différence entre un modèle déterministe et un modèle statistique.





Dans un modèle déterministe, une entrée donnée (ici x) produira toujours la même sortie (ici y). Alors que dans modèle statistique des entrées identiques peuvent donner des sorties différentes. [2]

[2] Les deux graphiques ci-dessus sont à titre illustratif et ne représentent pas le même phénomène.

Une façon générale d'écrire un modèle statistique est la suivante

$$Y_i = f(x_i) + \epsilon_i$$

Dans cette équation, nous appelons Y la réponse ou la variable dépendante, X le prédicteur (ou la variable explicative ou encore la variable indépendante), f la fonction de régression, fonction de lien ou encore fonction de lissage, et ϵ_i l'erreur.

Ce modèle partitionne les variations observées en Y en deux composantes :

- la variabilité expliquée par le modèle : les valeurs de Y varient (en partie) car elles sont associées avec différentes valeurs de X.
- la variabilité inexpliquée par le modèle : deux valeurs de la réponse Y associées aux mêmes X peuvent différer car elles correspondent à différentes valeurs de ϵ.

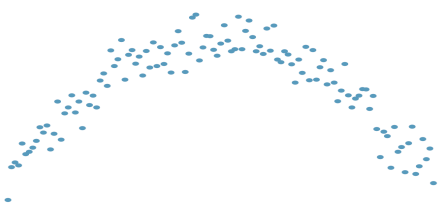
Remarquons que, puisque $E(\epsilon_i) = 0$, la fonction de régression f vérifie

$$E(Y_i) = f(x_i).$$

Souvent, en pratique, la fonction f qui relie X et Y est complètement inconnue. Dans un tel cas, on ne cherche pas à vérifier une loi théorique déjà prouvée, mais à en découvrir une nouvelle.

L'objectif principal est d'estimer f à partir d'un échantillon de n couples d'observations (x_i, y_i).

Sans terme d'erreur, càd lorsque $\sigma = 0$, il est relativement facile de trouver f à partir des (x_i, y_i). Mais en présence d'erreurs, le problème est bien plus difficile car nous ne pouvons pas distinguer la variation de Y due à X (expliquée) et celle due à l'erreur (inexpliquée).



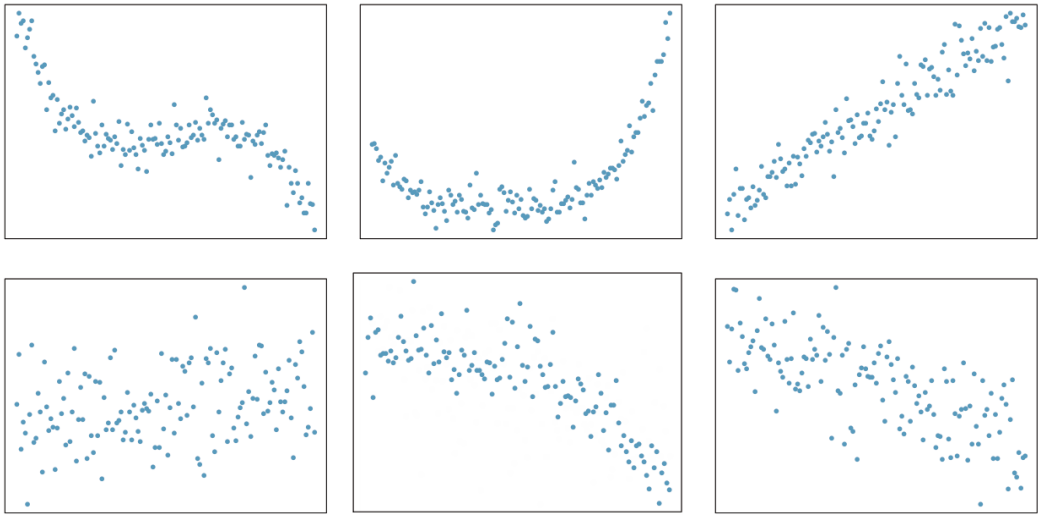
La première, et souvent la plus importante, difficulté est de spécifier la forme de f, càd choisir une **paramétrisation** adéquate.

Voici trois exemples simples :

- linéaire : $f(x) = \beta_0 + \beta_1 x$
- exponentiel : $f(x) = \beta_0 \exp(\beta_1 x)$
- sinusoidal : $f(x) = \beta_0 + \beta_1 \sin(\beta_3 + \beta_4 x)$

Une fois cette paramétrisation choisie, l'étape suivante consiste à estimer les paramètres du modèle à savoir les β's.

Le choix de la paramétrisation se fait souvent à l'aide d'un diagramme de dispersion.



Par la suite, on suppose que c'est la paramétrisation linéaire qui a été retenue.

Modèle linéaire et corrélation

Dans un modèle linéaire simple, **nous supposons** qu'il existe un β_0 et un β_1 tel que

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \forall i$$

Ce modèle est dit simple parce qu'il n'y a qu'un seul prédicteur et il est linéaire car Y dépend linéairement des β's. Un modèle linéaire avec plus d'un prédicteur est appelé régression linéaire multiple. Par la suite on ne considère que les modèles linéaires simples.

Un modèle linéaire est beaucoup plus général et beaucoup plus flexible que ce que laisse penser l'équation ci-dessus. Pour comprendre cela, notons que

les modèles suivants sont tous linéaires :

$$Y = \beta_0 + \beta_1 \ln(X) + \epsilon$$
$$Y = \beta_0 + \beta_1 X_1 X_2 + \epsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

les modèles suivants sont non-linéaires :

$$Y = \beta_0 \exp(\beta_1 X) + \epsilon$$
$$Y = \frac{\beta_0}{1 + \exp(\beta_1 X)} + \epsilon$$

Coefficient de corrélation empirique

Le degré avec lequel les points d'un diagramme de dispersion ont tendance à se regrouper autour d'une ligne reflète la force de la relation linéaire entre X et Y.

L'impression visuelle d'un nuage de points peut être trompeuse à cet égard, car un changement d'échelle des axes peut rendre le nuage plus serré ou plus détendu.

Pour cette raison, nous définissons le coefficient de corrélation, qui est une mesure numérique de la force de la relation linéaire entre deux variables. On notera ce coefficient par la lettre r.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

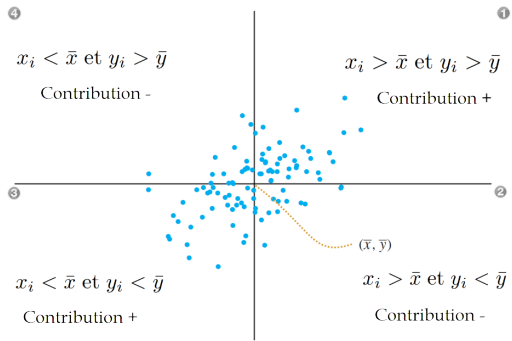
où s_x (s_y) est l'écart-type empirique (ou échantillonnale) des $x_i's$ $(y_i's)$, càd

$$s_x^2 = \frac{1}{n-1} \sum_{i}^{n} (x_i - \bar{x})^2 \quad \text{et} \quad s_y^2 = \frac{1}{n-1} \sum_{i}^{n} (y_i - \bar{y})^2$$

Comment fonctionne le coefficient de corrélation ?

Pour interpréter le signe de la corrélation, on peut considérer l'exemple suivant où les observations sont réparties en quatre quadrants comme suit.

Dans cette exemple, on peut constater qu'il y a plus d'individus dans les quadrants 1 et 3 que dans 2 et 4 $\rightarrow$ les contributions positives l'emportent sur les contributions négatives $\rightarrow$ relation globalement positive entre ces deux variables $\rightarrow r > 0$.

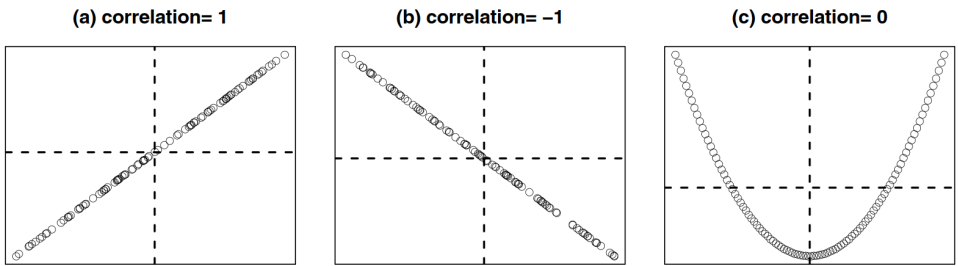


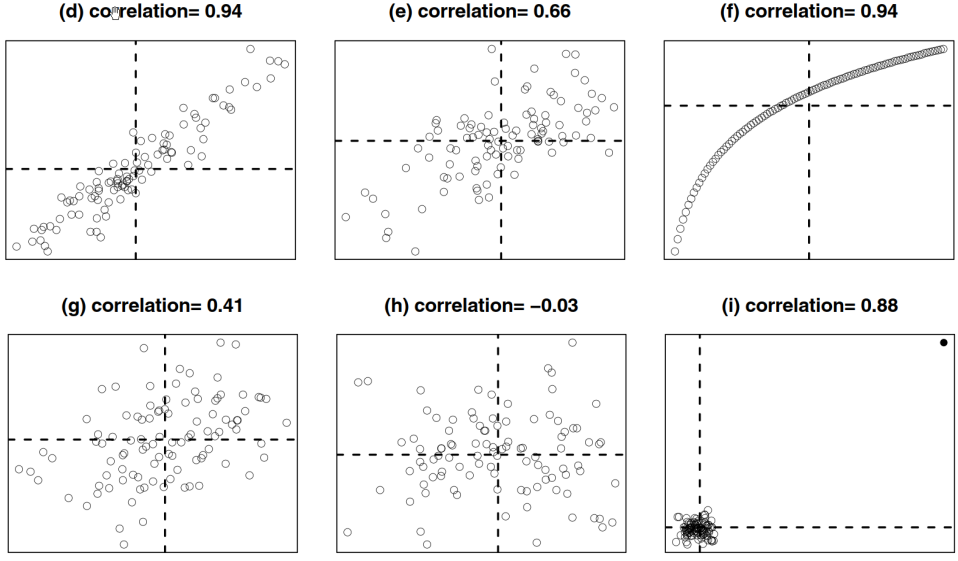
Voici quelques éléments à connaître concernant la corrélation :

- r n'a pas d'unité et il est toujours compris entre -1 et 1.
- $r > (<) 0$ indique une association positive (négative): des valeurs plus élevées d'une variable sont associées à des plus grandes (petites) valeurs de l'autre.
- $r = (-)1 \iff$ les points (x_i, y_i) sont parfaitement alignés sur une droite de pente (négative) positive.
- r est une mesure de l'intensité de la relation linéaire: un r proche de 1 ou de -1 indique une relation linéaire forte. Et un r proche de 0 indique une faible relation linéaire.
- Si $r = 0$, alors X et Y sont dites non-corrélés. Dans ce cas il n'y a pas de relation linéaire entre eux, mais il est possible qu'il existe une forte dépendance non linéaire entre les deux.

Plus tard, on verra une interprétation beaucoup plus précise d'une corrélation.

Pour l'instant, à titre d'illustration, voici quelques figures qui montrent des exemples de corrélation.



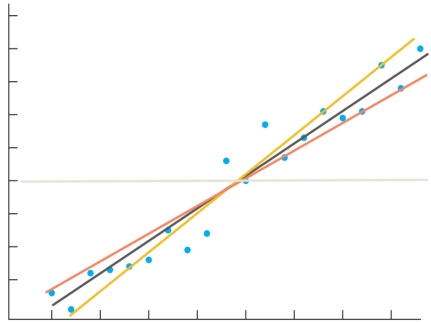

(d) correlation= 0.94
(e) correlation= 0.66
(f) correlation= 0.94
(g) correlation= 0.41
(h) correlation= −0.03
(i) correlation= 0.88

Pour calculer la corrélation dans R, il suffit d'utiliser la fonction *cor*.

```
> cor(x, y)

[1] 0.974
```

On peut donc dire qu'il y a une forte liaison linéaire positive entre le poids de la masse et la longueur de l'allongement du ressort.

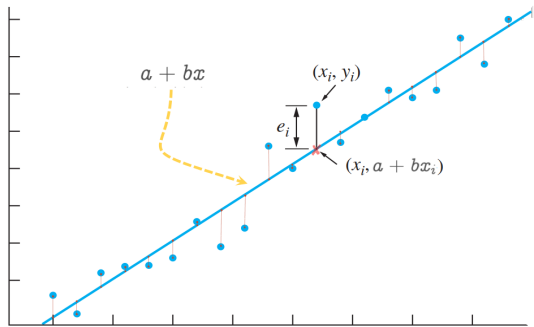
Maintenant, il ne reste qu'à trouver la bonne droite !



La méthode des moindres carrés

Quelle droite choisir ?

- La meilleure droite est celle qui se rapproche le plus possible du nuage des points formé par les données (x_i, y_i), $i = 1, \ldots, n$.
- La "proximité" d'une droite $a+bx$ à un point (x_i, y_i) est mesurée par la "distance" verticale : $e_i = y_i - (a + bx_i)$
- Donc, la meilleure droite est celle qui minimise la somme de toutes ces distances au carré: $\sum_i e_i^2$.
- Par conséquent, cette droite est appelée la droite des moindres carrés.



Plus précisément, soit $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$.
On montre facilement que

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{s_y}{s_x} r$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

où $\bar{x} = n^{-1} \sum_i x_i$, $\bar{y} = n^{-1} \sum_i y_i$, $\bar{x}^2 = \bar{x} \times \bar{x}$, $\overline{x^2} = n^{-1} \sum_i x_i^2$, et $\overline{xy} = n^{-1} \sum_i (x_i y_i)$.

La droite des moindres carrés résultante est donnée par

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + r \frac{s_y}{s_x} (x - \bar{x})$$

C'est la ligne qui traverse le centre de gravité $(\bar{x}, \bar{y})$ des points (x_i, y_i) avec une pente de $r \frac{s_y}{s_x}$.

La droite des moindres carrés avec R

Voici les paramètres estimés calculés à l'aide de des formules ci-dessus.

```
> b1 <- (mean(x * y) - mean(x) * mean(y))/(mean(x^2) - mean(x)^2)
> b0 <- mean(y) - b1 * mean(x)
> c(b0, b1)

[1] 12.70000  0.00115
```

Mais c'est bien plus pratique de faire appel à la fonction 'lm' (abréviation de linear model) :

```
> mod <- lm(y ~ x, data = mdt)
> mod

Coefficients:
(Intercept)            x
   12.70000      0.00115
```

Dans la sortie ci-dessus, « (Intercept) » signifie l'ordonnée à l'origine, càd $\widehat{\beta}_0$, et « x » indique la pente de la droite, càd $\widehat{\beta}_1$.

L'équation de la droite recherchée est donc

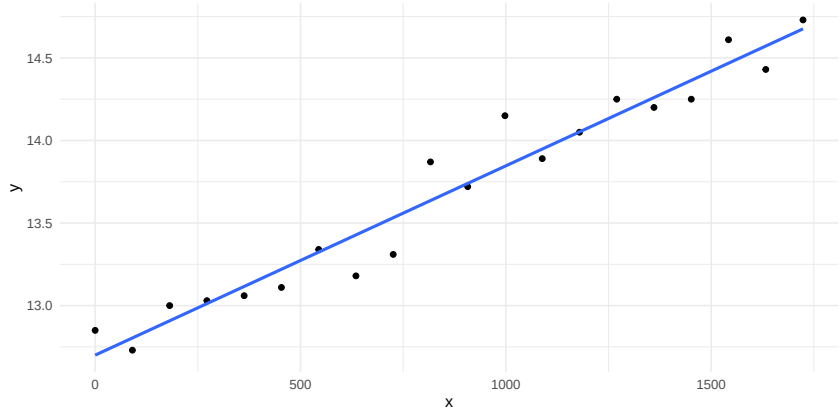
$$\widehat{\text{Longueur}} = 12.7 + 0.0011 \times \text{Poids}.$$

Interprétation des paramètres

- La longueur initiale (sans masse) est estimé à $12.7\,\text{cm}$.
- Lorsque le poids de la masse (x) augmente d'un gramme, on estime que la longueur du ressort (y) augmente en moyenne par $0.0011\,\text{cm}$.

Voici le diagramme de dispersion avec la droite de régression.

```
> plot(y ~ x, data = mdt, pch = 16)
> abline(mod, col = "blue")
```



Valeurs ajustées et résidus

À partir de la droite $\hat{y} = 12.7 + 0.0011x$, on peut calculer

1. Les valeurs ajustées (fitted values, en anglais):

$$\hat{y}_i = 12.7 + 0.0011x_i,$$

ce sont les valeurs prédites de Y prises aux X observées/mesurées.

2. Les résidus (residuals, en anglais)

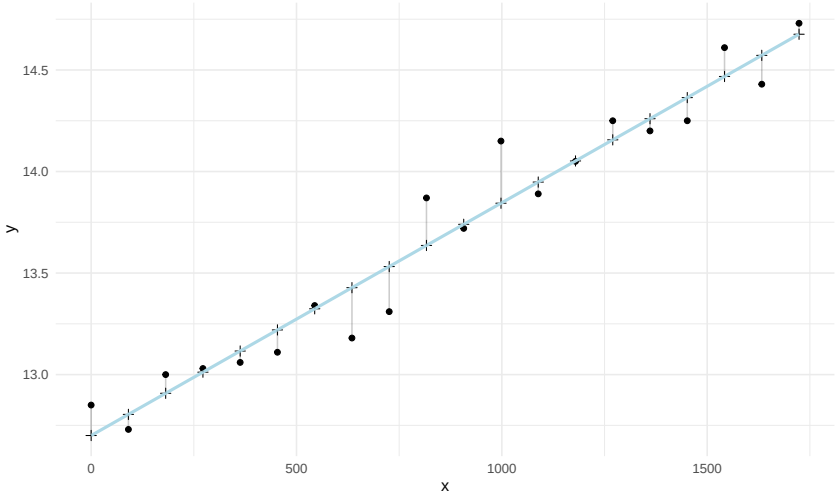
$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

c'est la différence entre les valeurs observées et ajustées de Y.

```
> prdt <- cbind(mdt, Fitted = predict(mod), Residuals = resid(mod))
> head(prdt, 4)
```

x	y	Fitted	Residuals
0.0	12.8	12.7	0.150
90.7	12.7	12.8	-0.074
181.4	13.0	12.9	0.092
272.2	13.0	13.0	0.018

Voici une représentation graphique de ces chiffres



L'examen des résidus nous permet de juger et quantifier la qualité d'ajustement du modèle aux données. Un traitement plus en détail de cette question sera développé plus tard

Prédiction

Un des buts de la régression est de proposer des prévisions pour la variable à expliquer Y. Soit x_{new} une nouvelle valeur de X (non présente dans notre échantillon). Selon notre modèle, la vraie valeur de Y qui correspond à x_{new} est donnée par

$$Y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}$$

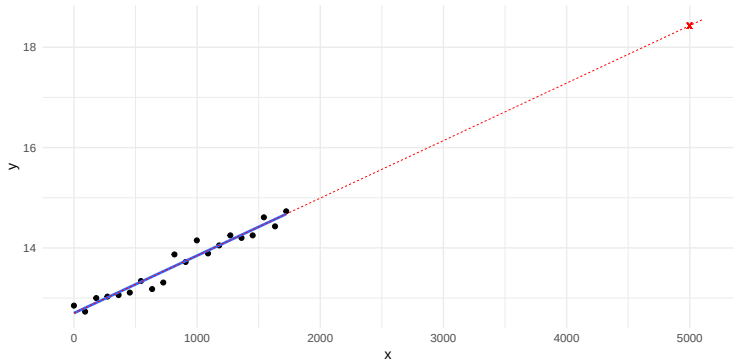
Cette dernière est une variable aléatoire dont nous pouvons estimer la moyenne par $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$. Autrement dit, la prédiction consiste simplement à prédire la valeur de Y d'un individu par la moyenne du groupe auquel cet individu appartient. En calculant par exemple $12.7 + 0.0011 \times 635.03 \approx 13.4$, on obtient une estimation de la moyenne de la longueur des ressort qui subissent une charge de 635.03 g.

Dans R, nous pouvons utiliser la fonction 'predict' pour faire des prédictions. Voici un exemple.

```
> predict(mod, new = data.frame(x = c(430, 635.03, 1300)))

   1    2    3
13.2 13.4 14.2
```

Mise en garde

Il ne faut pas extrapoler la droite au-delà des limites du domaine observé de X. Par exemple, on peut très bien calculer une prévision pour un $\texttt{Poids} = 5000$ g. En effet, l'estimation des moindres carrés pour ce poids est de 18.4 cm mais cette estimation n'est pas fiable. Il n'y a aucune raison que la linéarité se maintienne en dehors de notre domaine d'observation. Une extrapolation de notre équation de régression hors de ce domaine se fera à nos risques et périls, sans aucune garantie.



Qualité d'ajustement

Qualité d'ajustement

Coefficient de détermination

Avant d'utiliser le modèle, pour, par exemple, réaliser des prévisions, il faut d'abord s'assurer de sa qualité. Cette tache est nécessaire, car, en réalité, on se trouve souvent confronter à plusieurs modèles possibles et il n'est pas toujours facile de les séparer et d'en choisir le meilleur.

Par exemple, comment choisir entre les modèles suivants:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad Y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i, \quad Y_i = \beta_0 + \beta_1 \sqrt{x_i} + \epsilon_i,$$
$$\sqrt{Y_i} = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$$

Voici quelques éléments de réflexion :

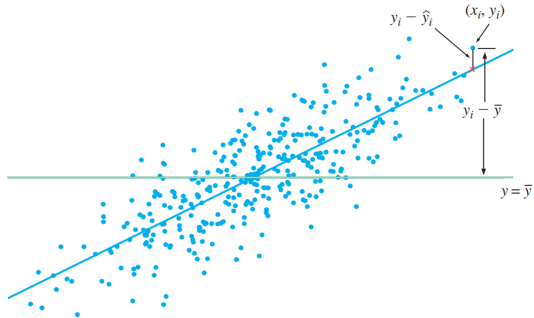
- Un modèle linéaire s'ajuste bien s'il existe une forte relation linéaire entre la réponse et la varaible explicative.
- Précédemment, nous avons vu que le coefficient de corrélation r mesure la force de cette relation. Par conséquent, r est une mesure de la qualité de l'ajustement d'un modèle linéaire.
- Mais quelle information ce coefficient nous fournit-il concernant la qualité du modèle? Autrement dit, comment interpréter r concrètement ?

Commençons tout d'abord par la relation, triviale, suivante:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

On peut montrer que

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$



Autrement dit, SCT $= \sum_i (y_i - \bar{y})^2$, *la Somme des Carrés des écarts Totaux*, càd la variation des y_i autour de leur moyenne $\bar{y}$, peut-être décomposée en deux parties

$$\text{SCT} = \text{SCR} + \text{SCE}, \text{ où}$$

- SCR $= \sum_i (y_i - \hat{y}_i)^2$: Somme des Carrés des écarts Résiduels, c'est la variation des y_i autour de la droite de régression (variance non expliquée par le modèle),
- SCE $= \sum_i (\hat{y}_i - \bar{y})^2$: Somme des Carrés des écarts Expliquées par le modèle, c'est la variation des $\hat{y}_i$ autour de leur moyenne $\bar{y}$.

Pour mieux comprendre la signification de chacun de ces termes, notez que

- SCT ne dépend que des données et non pas du modèle choisi.

$$SCT = 0 \Leftrightarrow y_i = \bar{y}\ \forall i \Leftrightarrow \text{ les } y_i \text{ ne varient pas}$$

- Le modèle s'ajuste parfaitement aux données ssi

$$\hat{y}_i = y_i\ \forall i \Leftrightarrow SCR = 0 \Leftrightarrow SCT = SCE$$

- Le modèle n'explique rien (aucune variation) ssi

$$\hat{y}_i = \bar{y}\ \forall i \Leftrightarrow SCE = 0 \Leftrightarrow SCT = SCR$$

La quantité

$$\frac{SCE}{SCT} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2},$$

qui est comprise entre 0 et 1, est la part de la variation en Y expliquée par le modèle. Cette quantité, très souvent exprimée en %, est appelée coefficient de détermination.

Il a été prouvé que le $\boxed{\text{coefficient de détermination} = r^2}$

Dans notre exemple avec le ressort, on a vu que $r = 0.974$, donc le coefficient de détermination est $r^2 = 0.95$, qu'on peu aussi calculer directement à l'aide de la commande suivante:

```
> summary(mod)$r.squared

[1] 0.949
```

On peut donc conclure que 95% de la variation de la longueur du ressort peut être expliqué par le poids de la masse ou, plus précisément, par la relation linéaire ($\hat{y}_i = 12.7 + 0.0011x_i$) qui relie ces deux variables.

ATTENTION: Le fait d'obtenir une très petite valeur de r^2 n'implique pas l'absence de relation entre les deux variables. En effet, le lien qui relie X et Y peut-être trés très fort, mais non linéaire. Aussi, le fait d'obtenir une très grand valeur de r^2 n'est pas une garantie absolue de la pertinence du modèle choisi.

Qualité d'ajustement

Analyse des résidus

L'analyse des résidus est une étape primordiale de la régression, car elle nous permet de juger facilement de la qualité d'un modèle en utilisant une approche simple basée, essentiellement, sur des graphiques.

Commençons par rappeler la définition des résidus

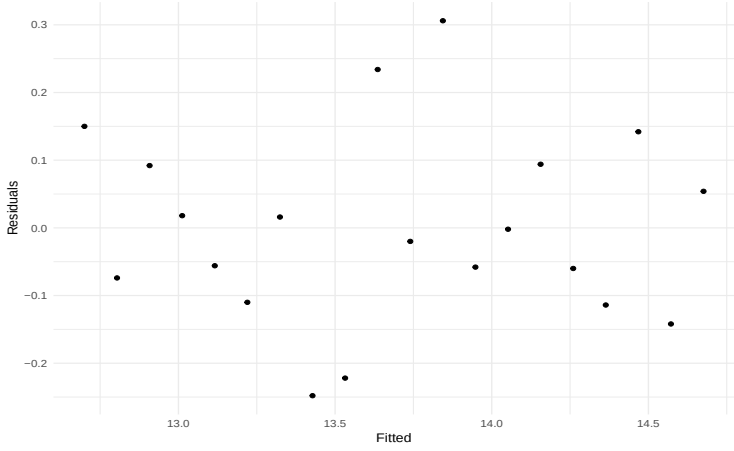
$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

Il est recommandé de tracer ces résidus en fonction des valeurs ajustées $\hat{y}_i$ càd tracer les couples de points $(\hat{y}_i, \hat{\epsilon}_i)$.

> Si ce graphique ne montre aucune structure, tendance ou courbature substantielle alors il est probable (mais pas certain) que l'hypothèse de la linéarité soit vérifiée.

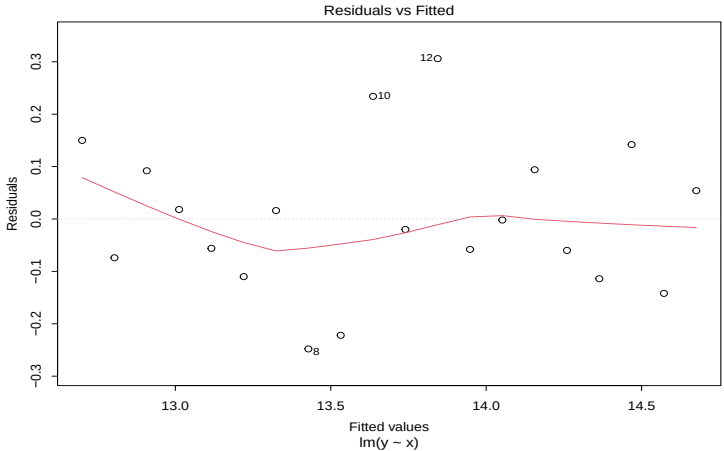
Voyons cela à l'aide de quelques exemples.

```
> plot(Residuals ~ Fitted, data = prdt)
```



On peut aussi “lisser” (càd ajuster une courbe de tendance) les résidus pour détecter plus facilement la présence d'une tendance ou forme particulière (signe d'une relation non linéaire).

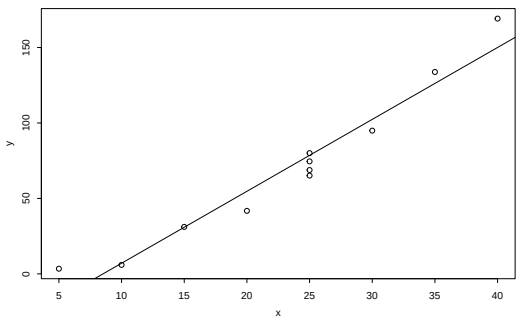
```
> plot(mod, 1)
```

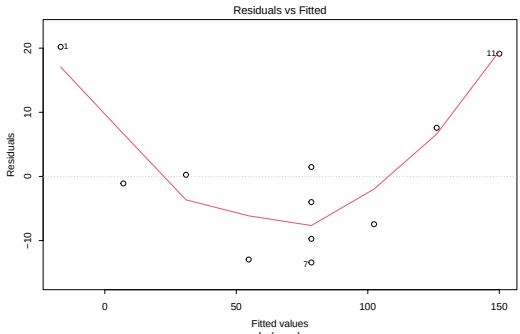


Dans ce cas, nous n'avons pas de raison de mettre en doute l'hypothèse de la linéarité

Voici un exemple d'un modèle linéaire dont les résidus semble indiquer un manque d'ajustement.

```
> x <- c(5, 10, 15, 20, 25, 25, 25, 25, 30, 35, 40)
> y <- c(3.42, 5.96, 31.14, 41.76, 74.54, 68.81, 65.13, 80.01, 94.92, 133.78, 169.16)
> dat <- data.frame(x = x, y = y)
> mod1 <- lm(y ~ x, data = dat)
```





Ce graphique montre une forte courbure. Et ceci malgré un coefficient de détermination de 95%.

Remédier aux problèmes

Si un modèle de régression linéaire simple s'avère non adéquat, alors on peut considérer l'une des deux options suivantes :

1. Abandonner ce modèle et développer un modèle plus complexe (régression pondérée, régression multiple, régression non-linéaire, régression non-paramétrique, ...) et plus approprié aux données.
2. Employer une certaine transformation sur les données (X, Y ou les deux) afin que le modèle de régression simple s'ajuste mieux aux données transformées. Exemple de transformations : $\log(X)$, $\sqrt{X}$, X^2, $\sin(X)$, ...

Ici nous allons considérer uniquement la dernière option.

Souvent le digramme de dispersion des données et le graphe des résidus contiennent des informations utiles et des indications précieuses qu'il convient de considérer pour choisir la transformation à appliquer. Cette approche est illustré par la suite.

Le graphique de notre modèle *mod1* montre une forte courbure, nous allons essayer une régression quadratique, càd une régression avec, en plus de X, X^2.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

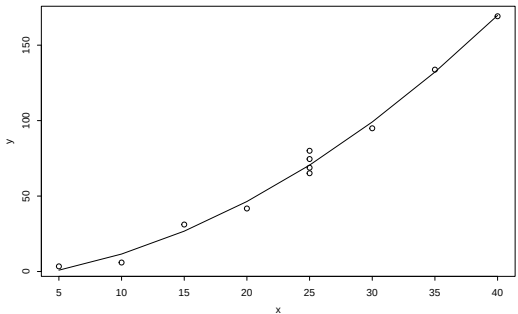
Pour cela nous ne devons pas nécessairement créer et sauvegarder la nouvelle variable X^2, mais l'ajouter simplement dans la formule du modèle, de la manière suivante

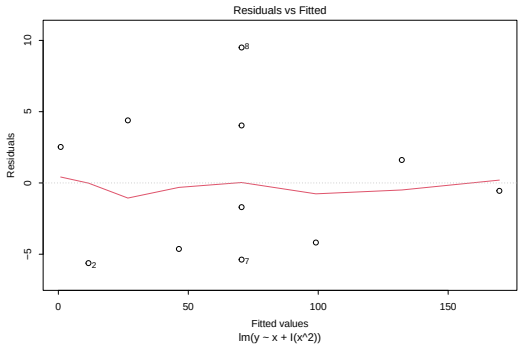
```
> mod2 <- lm(y ~ x + I(x^2), data = dat)
> mod2

Coefficients:
(Intercept)            x       I(x^2)
    -5.3185       0.7951       0.0895
```

Notez l'utilisation de la fonction "I", dans le code ci-dessus, sans quoi l'opérateur de puissance "^" risque d'être mal interprété par R (taper `?formula` pour plus d'info.).

L'équation du modèle *mod2* est $\hat{Y} = -5.318 + 0.795x + 0.090x^2$. Son coefficient de détermination est de 99%.

Voici le graphique des données avec la courbe de régression ainsi que le graphique des résidus.





Ce graphique ne montre aucune structure ou courbure particulière $\rightarrow$ nous choisissons donc le modèle *mod2*.

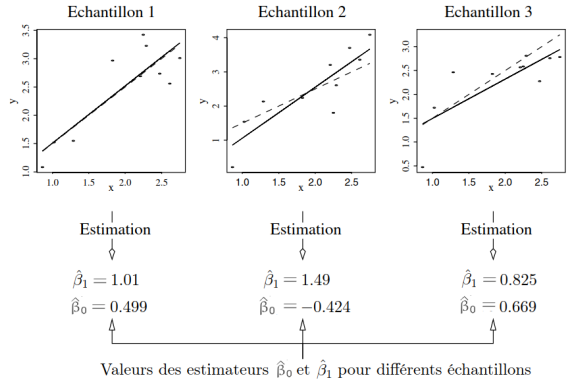
Inférence

Il est important de comprendre que les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$, qui sont calculées à partir d'un échantillon, changent avec les observations collectées.

Si le modèle linéaire est vrai, càd si il existe des constantes β_0 et β_1 tel que $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\forall i$, alors nous pouvons utiliser les estimations comme approximations pour les vraies valeurs. Dans ce cas, $(\widehat{\beta}_0, \widehat{\beta}_1) \xrightarrow[n\to\infty]{} (\beta_0, \beta_1)$.

Dans tous les cas, la méthode des moindres carrés fournie la meilleure approximation linéaire possible de Y basée sur X au sens que

$$(\widehat{\beta}_0, \widehat{\beta}_1) \xrightarrow[n\to\infty]{} \arg\min_{\beta_0,\beta_1} E(Y-\beta_0-\beta_1 X)^2$$



Le trait en pointillé représente la vraie droite de régression $Y = X + 0.5 + \varepsilon$ et le trait plein son estimation.

En utilisant une démarche similaire à celle vue dans le chapitre précédent, nous pouvons construire des intervalles de confiances pour β_0 et β_1, et de là, construire des intervalles de confiances pour les prédictions $\beta_0 + \beta_1 x$.

Voici comme calculer cela dans R. Nous reprenons ici l'exemple du ressort étiré.

```
> confint(mod, level = 0.95)

               2.5 %   97.5 %
(Intercept) 12.56778 12.83222
x            0.00102  0.00128
```

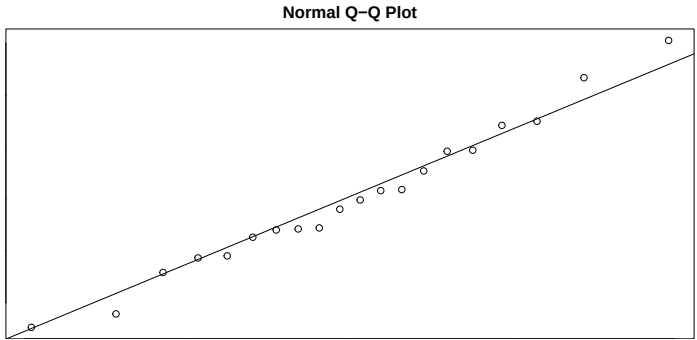
À 95%, nous pouvons déclarer que la vraie pente se situe dans l'intervalle $[0.0010, 0.0013]$

```
> predict(mod, new = data.frame(x = c(430, 635.03, 1300)), interval = "confidence",
      level = 0.95)

   fit  lwr  upr
1 13.2 13.1 13.3
2 13.4 13.4 13.5
3 14.2 14.1 14.3
```

À 95%, nous pouvons, par exemple, déclarer que la longueur moyenne des ressort qui subissent une charge de 635.03 g est située dans l'intervalle $[13.4, 13.5]$.

Notez que la validité de ces intervalles suppose, entre autres, que le modèle linéaire stipulé est vrai, et l'erreur ϵ est normalement distribué. Cette dernier hypothèse peut être vérifier à l'aide d'un qq-plot.

```
> qqnorm(prdt$Residuals)
> qqline(prdt$Residuals)
```



Si la normalité est rejetée, alors il est conseillé de transformer la réponse Y et réessayer.

Commandes R les plus utiles vues dans ce chapitre

- data.frame, cbind, et head
- plot, abline, qqnorm et qqline
- lm, I et cor
- predict et resid
- confint et summary