Université Rennes 2
Statistique
M1-GEO

Notes de cours

Ouvrages recommandés

Tous ces livres sont à la BU. Pour les acheter, venir au bureau A-240 ou envoyer un mail : nicolas.jegou@uhb.fr

1. Agnès Hamon et Nicolas Jegou, "Statistique descriptive. Cours et exercices corrigés.", PUR, 2008
   Pour revoir la base et s'initier à Rcmdr.
2. Jérôme Pagès, "Statistiques générales pour utilisateurs. 1-Méthodologie", PUR, 2005
   Transcription du cours donné à Agrocampus Rennes. Estimation, analyse de variance et régression puis introduction aux plans d'expérience et à l'ACP. Introduction à la statistique pratique, trés pédagogique et trés bien écrit.
3. François Husson et Jérôme Pagès, "Statistiques générales pour utilisateurs. 2-Exercices et corrigés", PUR, 2005
   Exercices et corrigés en lien avec l'ouvrage précédent. Quelques TP sur R proposés.
4. P.A.Cornillon et *al.*, "Statistiques avec R.", PUR, 2008
   Présentation du logiciel : objets, graphiques, programmation. Quinze méthodes statistiques classiques présentées avec R. Indispensable pour l'aspect logiciel.
5. J.Pagès, F.Husson, S.Lê, "Analyse de données avec R", PUR, 2009. Utile pour la seconde partie du S2 et le master 2.

1 Régression multiple

1.1 Rappels : régression linéaire simple

Dans cette section, nous reprenons brievement quelques points vus dans le cours du premier semestre. Il faut donc se référer à ce cours pour plus de développements. Nous insistons ici sur la notion de modèle et présentons le calcul du R^2 qui n'a pas été vu.

1.1.1 Modèle linéaire simple

Au S1, nous nous sommes restreints au cadre bivarié qui cherche à expliciter le lien entre deux variables : une variable explicative notée X et une variable à expliquer notée Y. Lorsque ces deux variables sont quantitatives i.e. lorsqu'elle prennent leurs valeurs dans $\mathbb{R}$, il est naturel, pour expliciter cette relation, de chercher une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $Y \approx f(X)$. On suppose ainsi qu'il existe une fonction $f : \mathbb{R} \to \mathbb{R}$ telle que $Y \approx f(X)$. On préfère écrire ce modèle sous la forme de l'égalité

$$Y = f(X) + \epsilon \tag{1}$$

où ϵ est une variable aléatoire mesurant les écarts entre Y et $f(X)$. Ces écarts aléatoires peuvent être dûs aux erreurs de mesures ou à d'autres variables qui n'ont pas été mesurées.

On dispose de mesures simultanées $(x_i, y_i)_{i=1,\cdots,n}$ des deux variables X et Y que l'on peut représenter par un nuage de points dans le plan. Le but de ce problème de régression est donc de d'estimer la fonction f inconnue dont la courbe est représentée en rouge (cf. partie gauche de la figure 1).

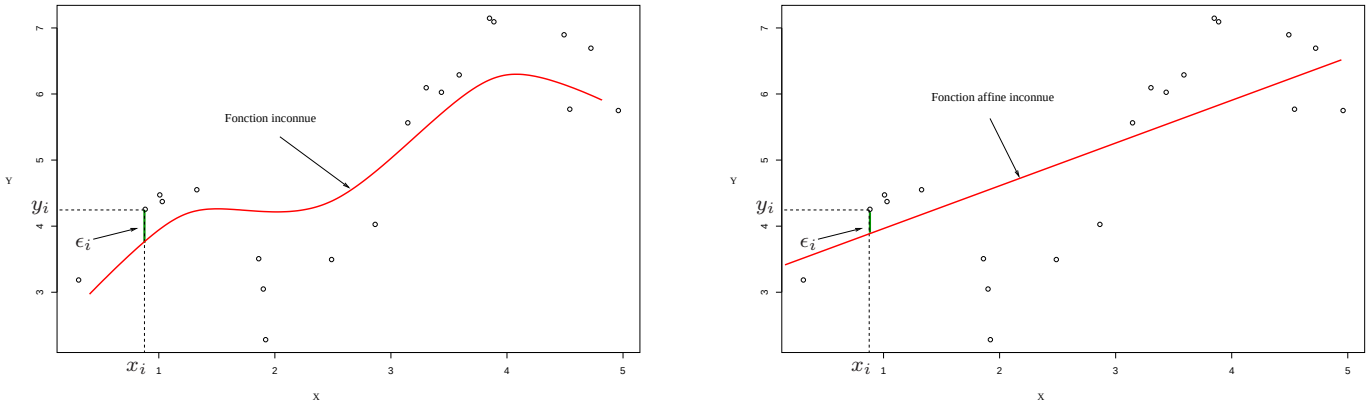

FIG. 1 – Modèles de régression simple.

Sans connaissance a priori sur la fonction f et si la représentation des points n'est pas en contradiction avec cette idée, il est courant de supposer que f est de la forme $f(x) = ax + b$. On dit alors qu'on postule un modèle linéaire entre les deux variables X et Y. Le modèle (1) devient :

$$Y = aX + b + \epsilon. \tag{2}$$

Postuler un modèle linéaire revient à formuler une hypothèse très forte sur la nature de la liaison entre les variables. Cela revient à supposer que la fonction sous-jacente est représentée par une droite (cf. partie droite de la figure 1). Cependant, outre le fait que cette simplification n'apparait pas forcément moins en accord avec les données que d'envisager une forme plus sophistiquée, elle offre une interprétation simple des paramètres a et b.

Une fois ce modèle linéaire postulé, le problème consistant à estimer f est ramené à celui de l'estimation des paramètres a et b. Il faut comprendre que, même si la relation entre X et Y est en effet linéaire, déterminer la vraie valeur des paramètres a et b est impossible. Pour ce faire, il faudrait en effet pouvoir observer la totalité des couples (x_i, y_i) dans la population mère ce qui est bien sûr impossible. Par contre, on dispose d'un échantillon de n observations de ces couples : $(x_i, y_i)_{i=1,\cdots,n}$ et ces données nous permettent de faire une estimation $\hat{a}$ et $\hat{b}$ des paramètres a et b. Donner une "bonne" estimation de a et b revient à tracer une droite qui passe "près" des points. Le critère classique de choix est celui des moindres carrés à savoir que l'on retient le couple $(\hat{a}, \hat{b})$ qui minimise la somme des carrés des résidus :

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2 . \tag{3}$$

Après avoir postulé un modèle linéaire, on calcule donc les estimations $\hat{a}$ et $\hat{b}$. La droite ajustée aux observations par le critère des moindres carrés a ainsi pour équation

$$y = \hat{a}x + \hat{b}.$$

Pour les x_i observés, on peut calculer l'ensembles des valeurs ajustées par le modèle en calculant, pour $i = 1, \cdots, n$:

$$\hat{y}_i = \hat{a}x_i + \hat{b}.$$

On peut représenter ces points en considérant les points situés sur la droite de régression et dont les abscisses sont les valeurs x_i (cf. figure 2). On appelle résidus, l'ensemble des valeurs notées $\hat{\epsilon}_i$ et définies par :

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

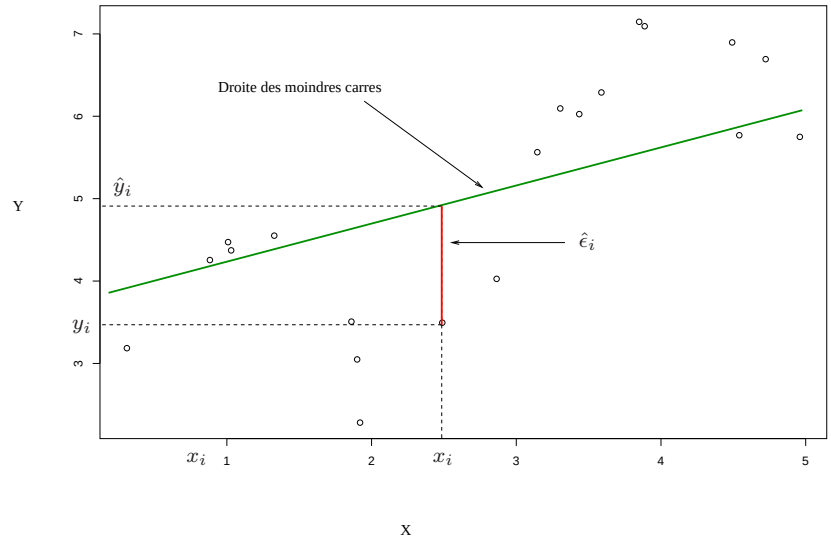

FIG. 2 – Résidus.

1.1.2 Une mesure de la qualité de l'ajustement : le R^2

Le R^2 est une mesure de la qualité de l'ajustement fondée sur la décomposition de la variabilité totale des observations. La quantité suivante, qui mesure la somme des carrés des écarts à la moyenne, est une mesure globale de la variabilité des y_i autour de la moyenne :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{4}$$

Notons qu'au coefficient $1/n$ près, elle correspond à la variance des y_i : σ_Y^2. Pour faciliter l'interprétation, on peut voir une mesure de variabilité de données comme une mesure de l'information que ces données contiennent. Notre problème consistant à expliquer Y, on peut donc considérer la quantité (4) comme la variabilité à expliquer.
On peut montrer l'égalité suivante qui en donne une décomposition :

$$\begin{aligned} \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\epsilon}_i^2. \end{aligned} \tag{5}$$

Nous avons deux termes. Le premier est la somme des carrés des écarts des valeurs ajustées à la moyenne. Si l'on considère qu'ajuster un modèle linéaire aux données conduit à subsituer aux données réelles (x_i, y_i) les données simplifiées $(x_i, \hat{y}_i)$, ce premier terme apparaît comme la part de la variabilité totale expliquée par cette simplification c'est-à-dire par le modèle. L'autre terme, la somme des carrés des résidus, peut dès lors être vue comme la part de la variabilité que le modèle n'explique pas et qu'on l'on attribue par conséquent au hasard. Graphiquement, la variabilité totale correspond à la somme des carrés des longueurs représentées en rouge en figure 3 et la part expliquée par le modèle à la somme des carrés des longueurs représentées en vert.
Par définition, le R^2 est le rapport :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{6}$$

C'est donc le rapport entre la variabilité des données expliquée par l'ajustement linéaire et la variabilité totale des données à expliquer. D'après (5), il est clair que c'est un rapport compris entre 0 et 1. On peut donc l'interpréter en termes de pourcentages. Plus il est proche de 1 et plus la part de la variabilité expliquée par le modèle est proche de la variabilité réelle. Au contraire, plus il est proche de 0 et moins le modèle explique la variabilité des données.

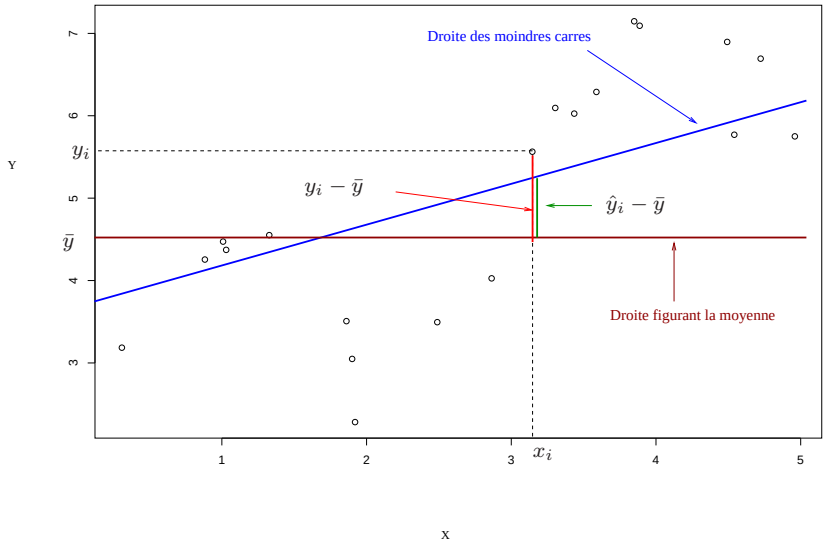

FIG. 3 – Calcul du R^2.

1.1.3 Exemple

Nous disposons de données receuillies dans une maternité. Le fichier comporte 27 variables mesurées sur 498 individus. Nous représentons en figure 4 le nuage croisant les observations des variables PoidsBB et Nbsem qui donnent le poids du bébé à la naissance et le nombre de semaine de grossesse. La variable à expliquer est bien sûr le poids du bébé.
Nous ajustons un modèle linéaire aux observations :

```
> RegModel.1 <- lm(PoidsBB~Nbsem, data=Dataset)

> summary(RegModel.1)

Call:
lm(formula = PoidsBB ~ Nbsem, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-1417.20  -244.03    -8.54   256.28  1802.80

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2337.40     523.03  -4.469 9.75e-06 ***
Nbsem         142.68      13.22  10.791  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 395.8 on 496 degrees of freedom
Multiple R-squared: 0.1901,Adjusted R-squared: 0.1885
F-statistic: 116.4 on 1 and 496 DF,  p-value: < 2.2e-16
```

Nous lisons $\hat{a} = 142.68$, $\hat{b} = -2337.40$. L'hypothèse $H_0 : a = 0$ est rejetée puisque la probabilité critique associée est très faible. Nous avons par ailleurs $R^2 = 0.1901$ (Multiple R-squared : 0.1901) : le modèle explique 19% de la variabilité des données. Notons que les sorties logiciel renvoient aussi le résultat du test de significativité de R^2 :

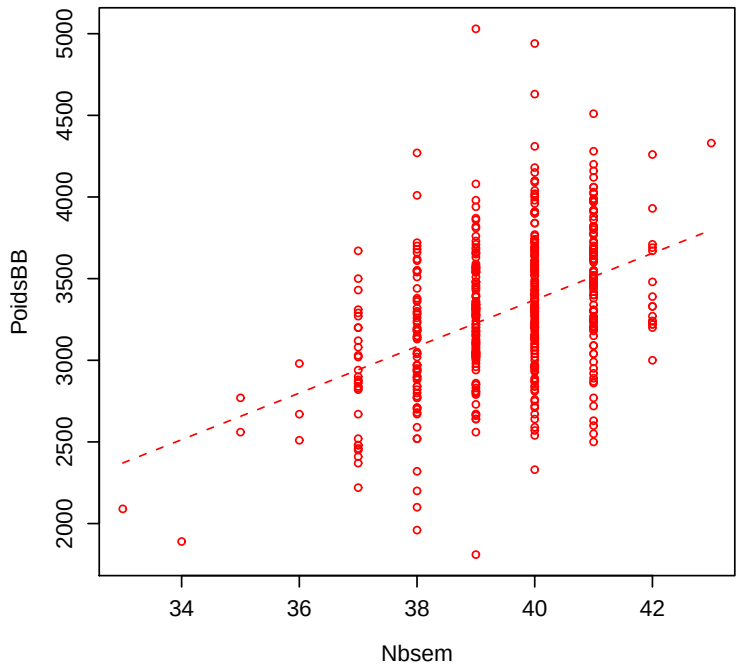

FIG. 4 – Croisement du poids du bébé et du nombre de semaines de grossesse.

```
F-statistic: 116.4 on 1 and 496 DF,  p-value: < 2.2e-16
```

On peut donc considérer que, si la part de variabilité expliquée par le modèle n'est pas considérable (19%), cette valeur est quand même très significativement différente de 0.

Remarque 1 *Le test de nullité de a et le test de nullité de R^2 conduisent aux mêmes probabilités critiques : c'est toujours le cas en régression simple.*

1.2 Modèle de régression multiple

1.2.1 Modèle de régression linéaire multiple

On dispose fréquement de plusieurs variables explicatives. Nous considérons dans cette section que toutes ces variables sont quantitatives. Nous les notons $X_1, \cdots, X_p$. La variable à expliquer Y est toujours quantitative. Expliciter la liaison entre Y et les X_i revient alors alors chercher une fonction f à p variables telle que $Y \approx f(X_1, \cdots, X_p)$. On considère donc le modèle

$$Y = f(X_1, X_2, \cdots, X_p) + \epsilon \tag{7}$$

où ϵ est une variable aléatoire et $f : \mathbb{R}^p \to \mathbb{R}$ la fonction à déterminer.

Il est difficile voire impossible de représenter la situation de manière simple. C'est encore possible lorsque l'on ne considère que deux variables explicatives car on peut représenter les fonctions à deux variables en dimension 3. En effet, pour représenter une fonction f de la forme

$$f \left\{ \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R} \\ (x_1, x_2) & \mapsto & y = f(x_1, x_2) \end{array} \right.$$

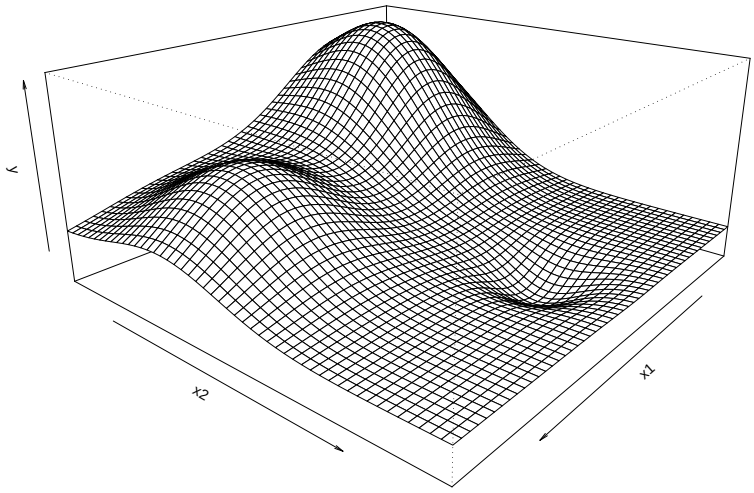

FIG. 5 – Représentation d'une fonction à deux variables.

un premier axe est consacré à la représentation des valeurs de x_1, un second à la représentation des valeurs de x_2, le troisième axe représentant les valeurs images $f(x_1, x_2)$. On obtient ainsi une surface comme représenté en figure 5.
Les fonctions multivariées les plus simples sont les fonctions affines de la forme

$$f(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

La représentation graphique d'une fonction affine à deux variables est un plan en dimension 3 (cf. figure 6). Comme en régression simple, on postule souvent que la relation entre Y et les X_j est linéaire c'est-à-dire que le modèle (7) s'écrit sous la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \tag{8}$$

Ce faisant, on fait une hypothèse forte sur la nature de la relation mais il s'avère que c'est souvent une première approximation satisfaisante de la situation. D'autre part, on verra en exemple que cette simplification permet d'interpréter de façon simple l'effet de chaque variable.

1.2.2 Estimateur des moindres carrés

On considère que l'on dispose de n observations de la variable à expliquer $y_1, \cdots, y_n$. On dispose également de n observations simultanées de chaque variable X_j. Pour la variable X_j, on note ces observations $x_{1j}, x_{2j}, \cdots, x_{nj}$. En postulant le modèle linéaire (8), on suppose donc qu'elles sont de la forme :

$$\begin{array}{lll} y_1 & = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} & +\epsilon_1 \\ y_2 & = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} & +\epsilon_2 \\ \cdots & & \\ y_n & = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} & +\epsilon_n \end{array} \tag{9}$$

On notera de manière condensée : pour $i = 1, \cdots, n$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i. \tag{10}$$

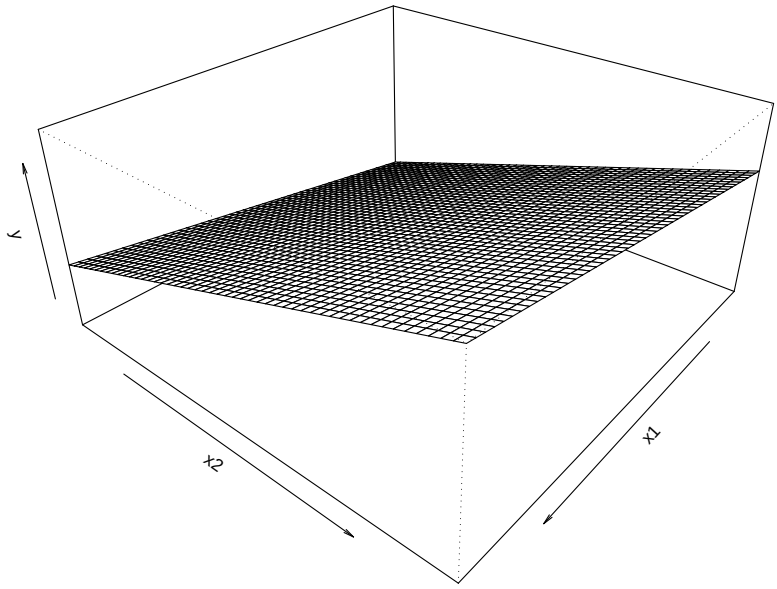

FIG. 6 – Représentation d'une fonction affine à deux variables.

On ne sait pas si la relation entre Y et les X_i est effectivement linéaire mais si en effet elle l'est, les paramètres $\beta_0, \beta_1, \beta_p$ sont et resteront inconnus. Les données $(y_i, x_{i1}, \cdots, x_{ip})$ nous permettent cependant d'en faire une estimation. Comme en régression linéaire simple, c'est le critère des moindres carrés que l'on utilise le plus souvent. Le principe est de retenir comme estimation de $(\beta_0, \beta_1, \cdots, \beta_p)$ le $(p+1)$-uplet $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ qui minimise la somme des carrés des résidus :

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}))^2 .$$

Ce qui vient d'être écrit définit l'estimateur dit des moindres carrés.

On peut interpréter graphiquement la situation en dimension 3 pour deux variables explicatives (cf. figure 7). Les paramètres $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ sont ceux, parmi tous les triplets $(\beta_0, \beta_1, \beta_2)$, qui définissent le plan le plus proche des points dans le sens où c'est le plan tel que la somme des carrés des longueurs rouge est minimum.

1.3 Exemple

1.3.1 Modèle linéaire complet

Nous poursuivons l'exploration des données du fichier bebe.txt. Nous commençons par enlever tous les cas contenant des valeurs manquantes :

```
Dataset <-
  read.table("/home/jegou/ENSEIGNEMENT/GEO/MASTER/COURS-TD/S2/bebe.txt",
  header=TRUE, sep=";", na.strings="NA", dec=".", strip.white=TRUE)
Dataset <- na.omit(Dataset)
```

Nous considérons les 11 variables suivantes pour expliquer la variable poidsBB :

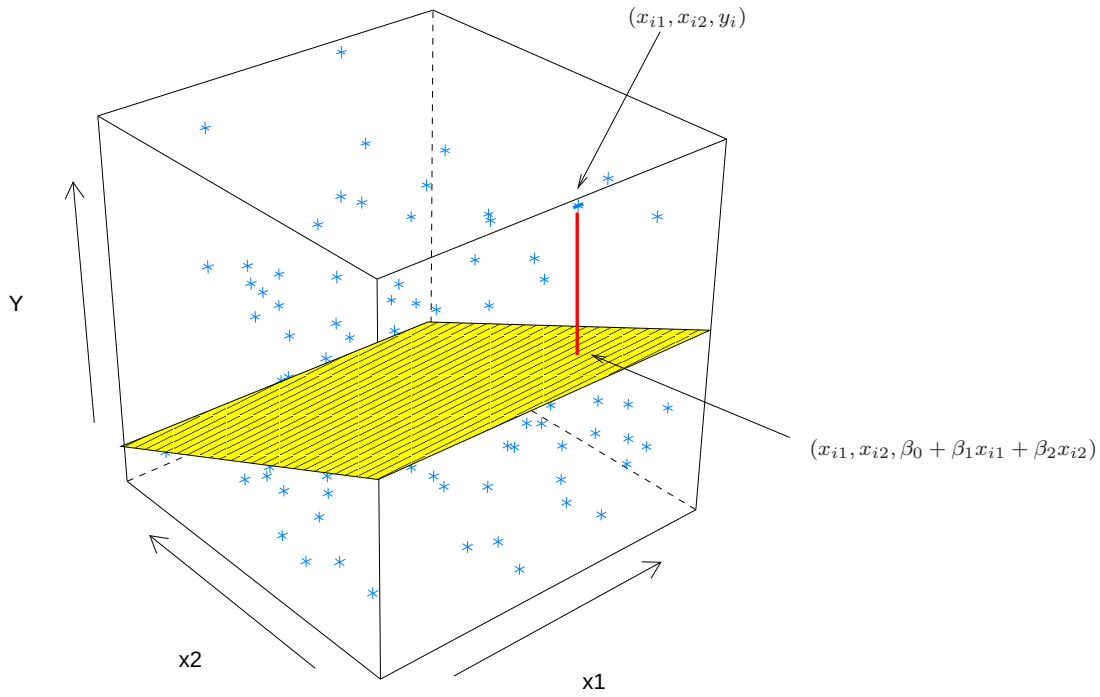

FIG. 7 – Moindres carrés.

- Nbsem
- TailleBB
- PoidsPlacenta
- AgedelaMere
- TailMere
- PoidsMere
- Agedupere
- TailPere
- PoidsPere
- NbGrossess
- NbEnfants

Ci-dessous, nous définissons le modèle linéaire correspondant avec R. Nous obtenons le résumé suivant où l'on peut en particulier lire l'estimation de chacun des coefficients $(\beta_0, \beta_1, \cdots, \beta_{11})$:

```
> LinearModel.1 <- lm(PoidsBB ~ Nbsem +TailleBB +PoidsPlacenta +AgedelaMere
+   +TailMere +PoidsMere +Agedupere +TailPere +PoidsPere +NbGrossess +NbEnfants,
+    data=Dataset)

> summary(LinearModel.1)

Call:
lm(formula = PoidsBB ~ Nbsem + TailleBB + PoidsPlacenta + AgedelaMere +
    TailMere + PoidsMere + Agedupere + TailPere + PoidsPere +
    NbGrossess + NbEnfants, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-778.226 -167.590    4.189  157.998  730.712

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   -5168.2579   592.4632  -8.723  < 2e-16 ***
Nbsem            51.4677    10.8161   4.758 2.85e-06 ***
TailleBB        141.5900     7.5529  18.746  < 2e-16 ***
PoidsPlacenta     0.6337     0.1068   5.932 7.16e-09 ***
AgedelaMere       0.7917     4.1453   0.191   0.8486
TailMere         -2.0962     2.4039  -0.872   0.3838
PoidsMere         3.0538     1.4292   2.137   0.0333 *
Agedupere        -0.1387     3.0100  -0.046   0.9633
TailPere         -4.3501     2.1889  -1.987   0.0477 *
PoidsPere        -0.7367     0.3693  -1.995   0.0468 *
NbGrossess       -0.6666    18.5825  -0.036   0.9714
NbEnfants        22.7136    25.3603   0.896   0.3711
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 242 on 353 degrees of freedom
Multiple R-squared: 0.6705,Adjusted R-squared: 0.6602
F-statistic:  65.3 on 11 and 353 DF,  p-value: < 2.2e-16
```

1.3.2 Choix de modèle

Certaines variables considérées à l'instant ne sont sans doute pas pertinentes. On observe d'ailleurs que de nombreux paramètres ne sont pas significativement différents de 0. Il est préférable de se restreindre à un modèle impliquant moins de variables explicatives. En effet, tout d'abord il est vraisemblable que certaines variables ne soient pas significatives auquel cas il n'y a aucun intérêt à les garder dans le modèle. Par ailleurs, il est probable que certaines variables explicatives soient liées entre elles : NbGrossess et NbEnfants ou peut-être TailMere et PoidsMere par exemple. Ne retenir que quelques variables réellement pertinentes permettra une meilleure estimation des coefficients associés et on peut espérer que le modèle aura ainsi de meilleures vertues prédictrices.

Il existe plusieurs façon de procéder à un choix de modèles. Commençons par écarter deux mauvaises idées. On pourrait penser qu'il est judicieux de se baser sur la valeur du R^2 comme en régression simple. Malheureusement, ce n'est pas le cas. Il faut en effet savoir que le fait d'ajouter une variable dans un modèle augmentera toujours la part de variabilité expliquée par celui-ci et donc la valeur de R^2. De même, enlever une variable du modèle fera toujours baisser la valeur de R^2.
On pourrait avoir l'idée d'enlever toutes les variables dont le coefficient n'est pas significatif à un seuil donné. Là encore, l'idée n'est pas bonne. En effet, le fait de considérer un grand nombre de variable a tendance à rendre l'estimation individuelle des coefficients moins précise. Il est possible qu'en enlevant brutalement toutes les variables dont la probabilité associée au test de nullité du coefficient est inférieure à un seuil donné, on ne retienne pas dans le modèle final une variable qui n'aurait pas été éliminée avec un modèle initial en comportant moins (un exemple concret illustrant cette explication un peu nébuleuse est donné en TD).

Il existe des méthodes rigoureuses de choix de modèles. Le problème est que la théorie qui les accompagne est assez lourde. Nous en proposons une ici qui n'a pas leur rigueur mais qui présente l'avantage d'être simple et de rester cohérente. Voici la démarche :

1. Estimer les paramètres du modèle complet.
2. Enlever du modèle précédent la variable la moins significative et ré-estimer les paramètres.
3. Itérer les deux points précédents jusqu'à ce qu'il ne reste que des variables significatives.

Avec cette procédure, on retient finalement le modèle suivant :

```
> LinearModel.final <- lm(PoidsBB ~ Nbsem +TailleBB +PoidsPlacenta
+    +PoidsMere   +PoidsPere  ,
```

```
+    data=Dataset)

> summary(LinearModel.final)

Call:
lm(formula = PoidsBB ~ Nbsem + TailleBB + PoidsPlacenta + PoidsMere +
    PoidsPere, data = Dataset)

Residuals:
      Min        1Q    Median        3Q       Max
-723.5544 -158.1396   -0.5643  164.0363  770.5992

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5923.2091   440.8028 -13.437  < 2e-16 ***
Nbsem            47.6987    10.7302   4.445 1.17e-05 ***
TailleBB        138.9664     7.3841  18.820  < 2e-16 ***
PoidsPlacenta     0.6413     0.1046   6.130 2.31e-09 ***
PoidsMere         2.8369     1.3522   2.098   0.0366 *
PoidsPere        -0.8250     0.3686  -2.238   0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 243.1 on 359 degrees of freedom
Multiple R-squared: 0.6617,Adjusted R-squared: 0.657
F-statistic: 140.4 on 5 and 359 DF,  p-value: < 2.2e-16
```