La Régression Linéaire Multiple
Modèle biologique : estimation de l'âge des ormeaux

L3 Bio-Informatique (ISV51)

Assia Bemehdia[1]
Pauline Spinga[2]
groupe 6

17/12/2020

[1]benmehdia.assia@gmail.com
[2]spinga.pauline@gmail.com

Objectif

- Existe-il un lien entre l'âge d'un organisme et ces mesures physiques ?
- Quelle méthode statistique doit-on utilisée pour répondre à cette question biologique ?

$$\Downarrow$$

Prédire l'âge des ormeaux via un modèle de régression linéaire multiple élaboré avec le langage de programmation R.

Présentation du package

Pour ce projet nous avons construit et utilisé un package de 4 fonctions :

- Identite(coeff,hat,seuil)
- Simulation(sigma,coeff,n,seuil)
- Resultat_simu(res)
- Histog_simu(sigma_inf,coeff,n)

Plan

Introduction

La régression linéaire :

- **définition :** Un modèle de régression linéaire est un modèle qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.
- **type :**

1. La régression linéaire simple

$$y = a_0 + a_1 x_1 + \varepsilon$$

2. La régression linéaire multiple

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_p x_p + \varepsilon$$

$\longrightarrow$ plusieurs variables explicatives

Introduction

$$y = a_0 + a_1 x_1 + \varepsilon$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_p x_p + \varepsilon$$

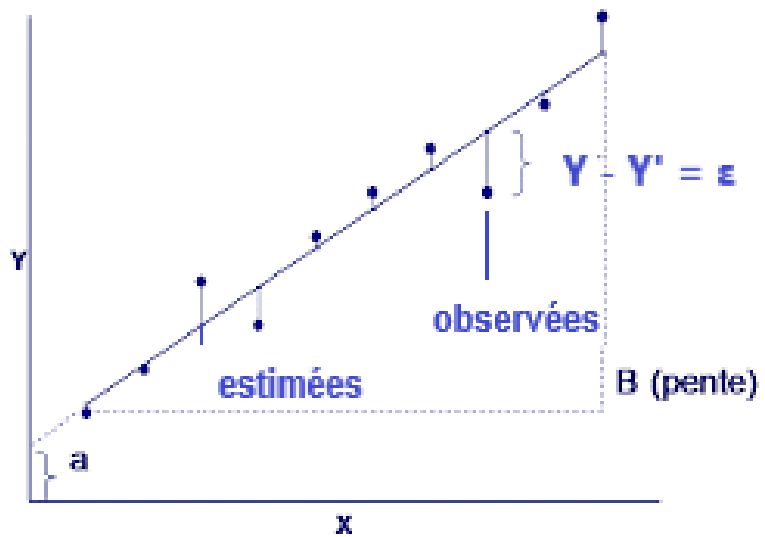
$$\downarrow$$

$$Y = Xa + \varepsilon$$

Y : variable à expliquer (indépendante)
x : variables explicatives (dépendantes)
a_i : paramètres de régression
ε : le bruit du modèle $\Longrightarrow \varepsilon \sim N(0, \sigma^2)$

La régression linéaire simple

- Une droite qui représentera mathématiquement la relation existante entre des variables $\Longrightarrow Y = a + bX$
- Quand on tient compte de l'erreur résiduelle de chaque observation $\Longrightarrow \hat{Y} = a + bX + \varepsilon$



3

[3] adapté d'aprés http://www.lacim.uqam.html

La régression linéaire multiple

Plusieurs tirages

$$\left\{\begin{array}{c} y_1 = a_0 + a_1 x_{11} + a_2 x_{12} + ... + a_p x_{1p} + \varepsilon_1 \\ y_2 = a_0 + a_1 x_{21} + a_2 x_{22} + ... + a_p x_{2p} + \varepsilon_2 \\ \vdots \\ y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} + \varepsilon_i \\ \vdots \\ y_n = a_0 + a_1 x_{n1} + a_2 x_{n2} + ... + a_p x_{np} + \varepsilon_n \end{array}\right\}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_i \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{1p} \\ 1 & x_{21} & x_{22} & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & x_{ip} \\ 1 & x_{n1} & x_{n2} & x_{np} \end{bmatrix} \times \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_i \\ \varepsilon_n \end{pmatrix}$$

Estimation du modèle de la régression linéaire multiple

$$Y = Xa + \varepsilon \Longrightarrow \hat{Y} = X\hat{a} + e$$

Avec :

- $\hat{Y}, \hat{a}$: Les paramètres de regression estimés
- $e = Y - \hat{Y}$

Pour calculer les valeurs des paramètres estimés on utilise **la méthode des moindres Carrés** qui permet la minimalisation de la somme des Carrés des erreurs.

$$\Sigma e_i^2 = (Y - \hat{Y})(Y - \hat{Y}) \longrightarrow (Y - X\hat{a})(Y - X\hat{a})$$

$$\boxed{\hat{a} = (X^T \times X)^{-1} \times X^T \times Y}$$

Simulation du modèle de régression linéaire multiple

1. Exemple 1 une seule simulation avec Sigma fixé
2. Exemple 2 plusieurs simulation
   - Avec un sigma fixé
   - Avec un sigma varié (min ou max)

Simulation du modèle de régression linéaire multiple

Exemple 1: une seule simulation avec Sigma fixé

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_p x_p + \varepsilon$$

- On tire $\varepsilon \Longrightarrow \varepsilon \sim N(0, \sigma^2)$
- On fixe $\sigma = 22$: représente le bruit
- On fixe le nombre de tirage $n = 1000$
- Initialisation du nombre de paramètres de régression $p = 3$ avec $a_1 = 5$ et a_2, a_3 sont proportionnels à a_1
- Initialisation des variables explicatives $x_1, x_2, x_3 \Longrightarrow$ loi uniforme
- générer une matrice

$$\Downarrow$$

Estimer la valeur des coefficients

étudier l'ordre d'importance des paramètres

université évry Val d'Essonne | université PARIS-SACLAY

Simulation du modèle de régression linéaire multiple

Exemple 1: une seule simulation avec Sigma fixé

```
## [1] "Les paramètres estimés"

##              [,1]
## [1,]  8.333593
## [2,]  6.355068
## [3,] 18.494502

## [1] "Les paramètres fixés au départ"

## [1]  5 10 15
```

Comparaison entre paramètres observés et les paramètres estimés

- Vérification de la bonne estimation des paramètres selon un seuil fixé arbitrairement à 2
  Si $a - \hat{a} < 2 \longrightarrow$ **bonne estimation**

```
## [1] "0 = nombre de paramètres bien estimés"
```

```
## [1] "La proportion de paramètres bien estimés"
```

```
## [1] 0
```

- Vérification de la conservation de l'ordre des paramètres

```
## [1] "la proportion des p.estimés dans le bon ordre"
```

```
## [1] 0.3333333
```

Génération d'un modèle linéaire lm() pour notre exemple

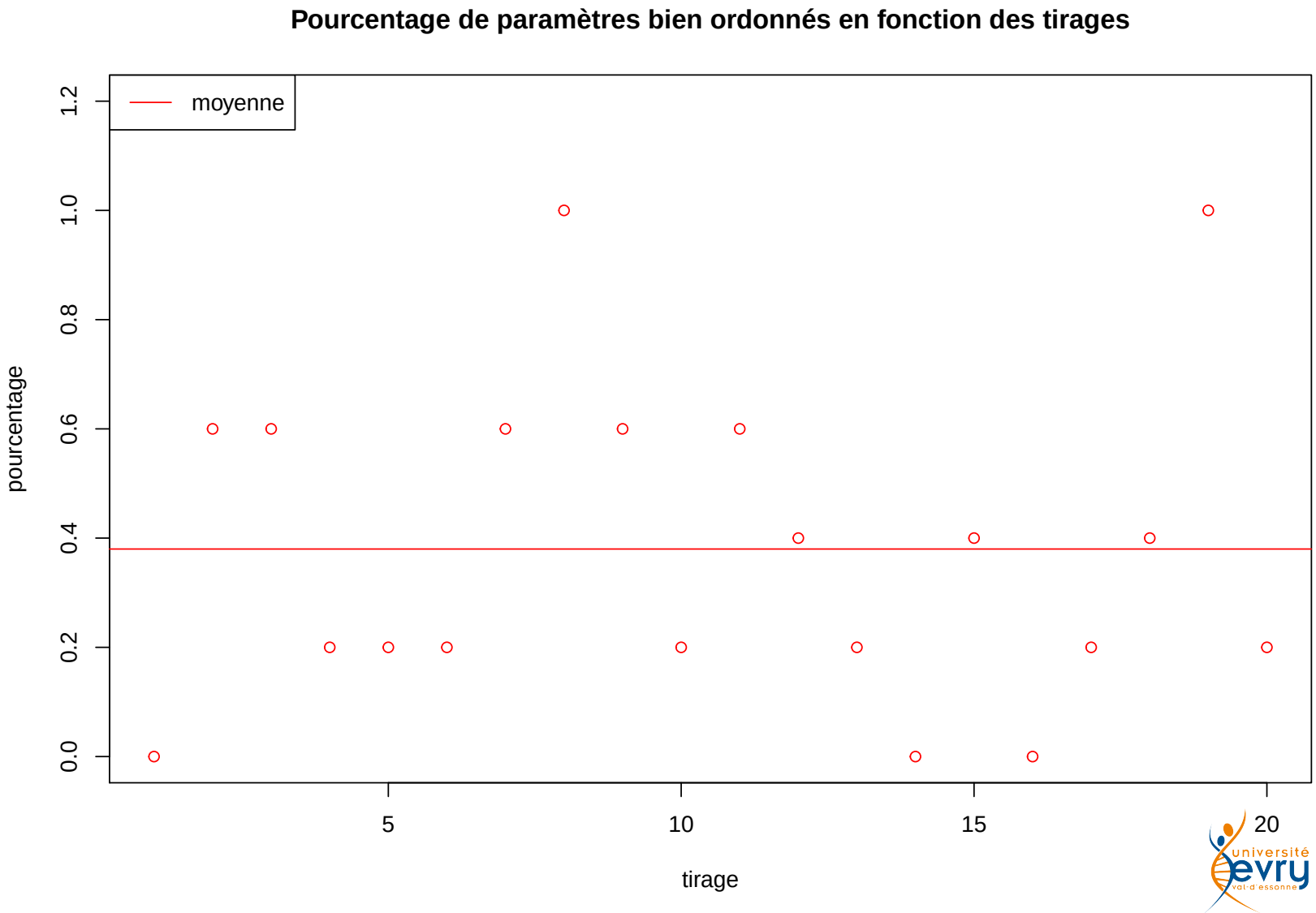
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.846 -14.259   0.259  14.639  66.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.649      2.273   0.726  0.46829
## X1             7.325      2.430   3.015  0.00264 **
## X2             5.420      2.424   2.236  0.02559 *
## X3            17.472      2.457   7.110 2.22e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.35 on 996 degrees of freedom
## Multiple R-squared:  0.05875,    Adjusted R-squared:  0.05591
## F-statistic: 20.72 on 3 and 996 DF,  p-value: 4.993e-13
```

Simulation du modèle de régression linéaire multiple

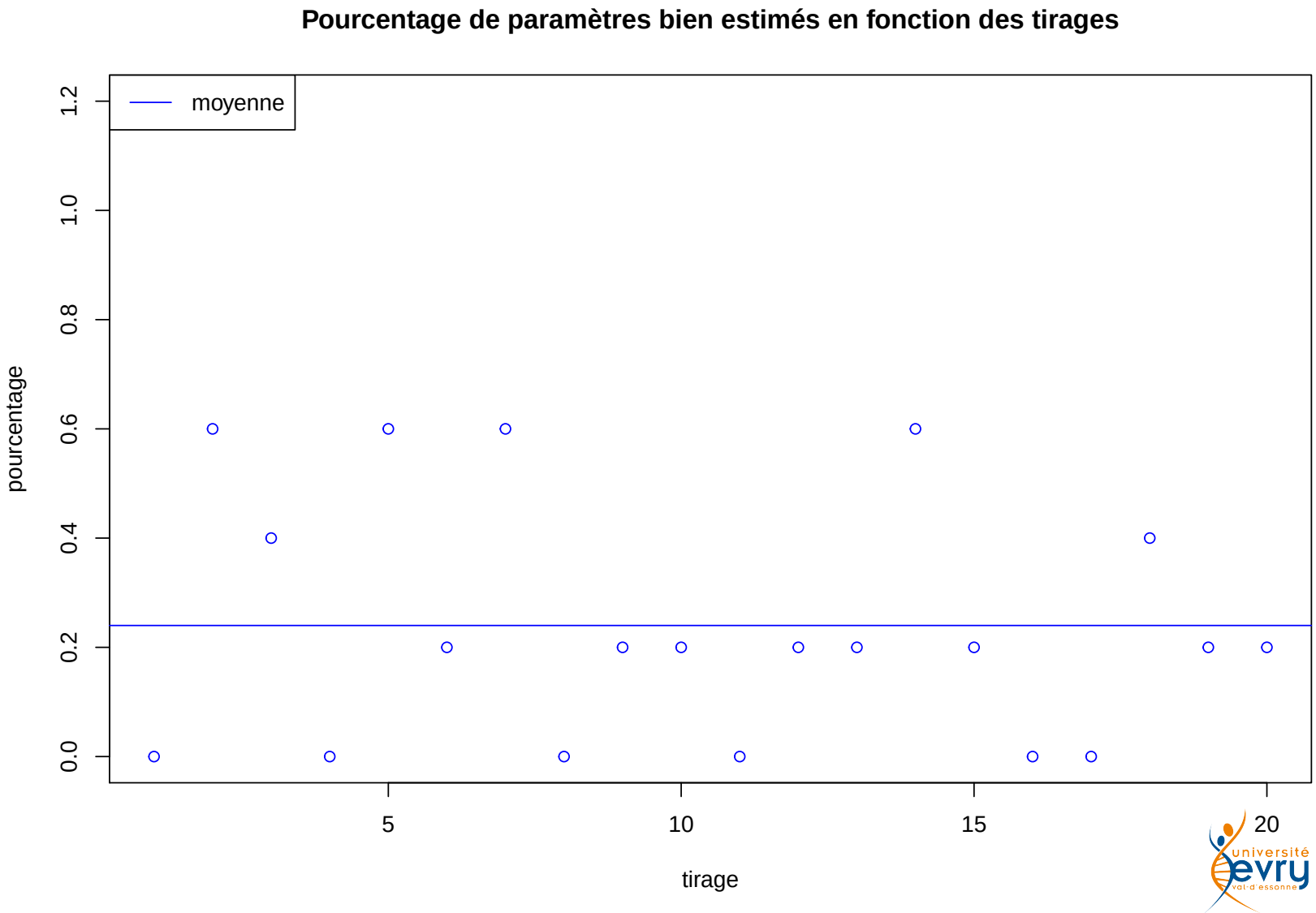
Exemple 2: plusieurs simulations avec un sigma fixé

- replicate() de la fonction simulation
- Affichage des résultats

Simulation du modèle de régression linéaire multiple



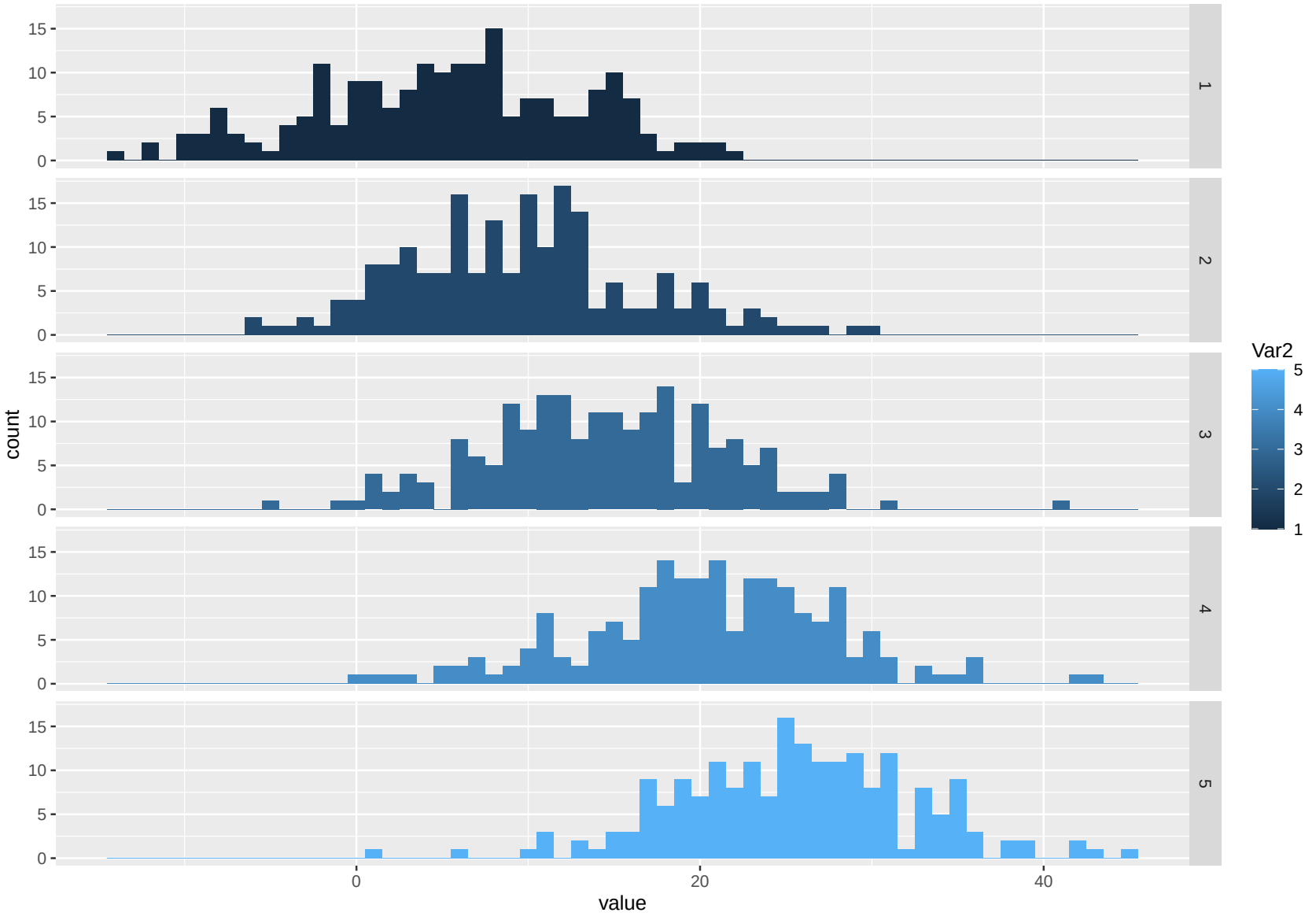
Simulation du modèle de régression linéaire multiple



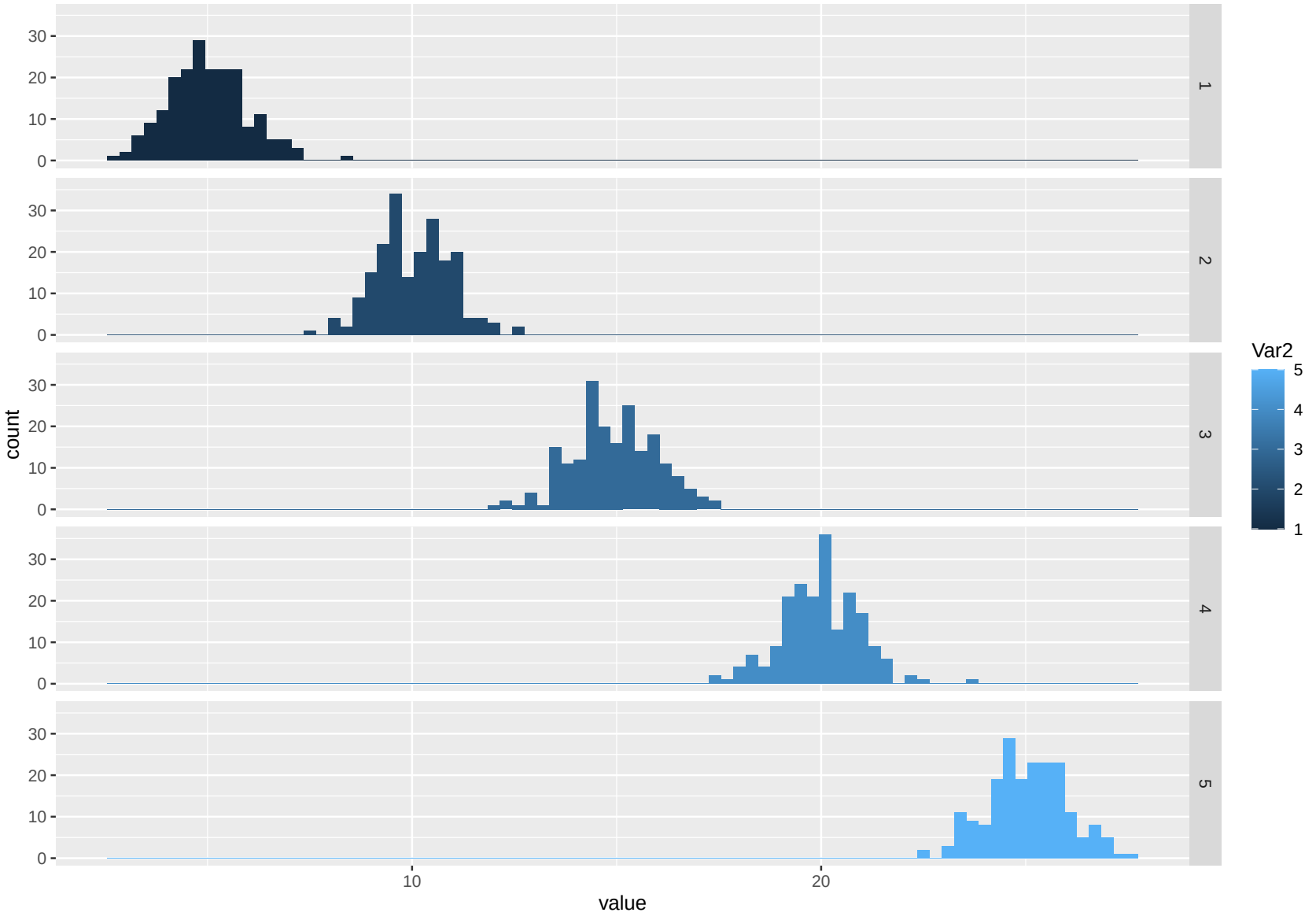
pourquoi nous avons obtenus ces résulatats ?

Les coefficients a_i sont des estimateurs. Ils ont une distribution centrée autour de leur valeur théorique.

Affichage des résultats des simulations sous forme d'histogrammes



Affichage des résultats des simulations sous forme d'histogrammes



Application du modèle sur des données réelles

Règne : Animalia
Embranchement : Mollusca
Famille : Haliotididae
Genre : Haliotis

- Mollusques marins à coquille unique, qu'on trouve dans les eaux peu profondes du littoral accrochés aux rochers.

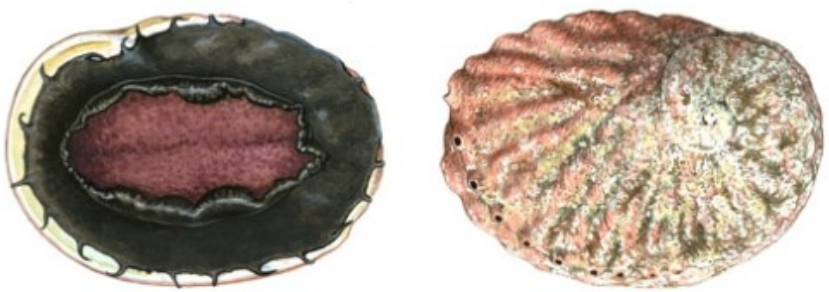


Figure 1: abalone [4]

[4] https://dpipwe.tas.gov.au/

Application du modèle sur des données réelles

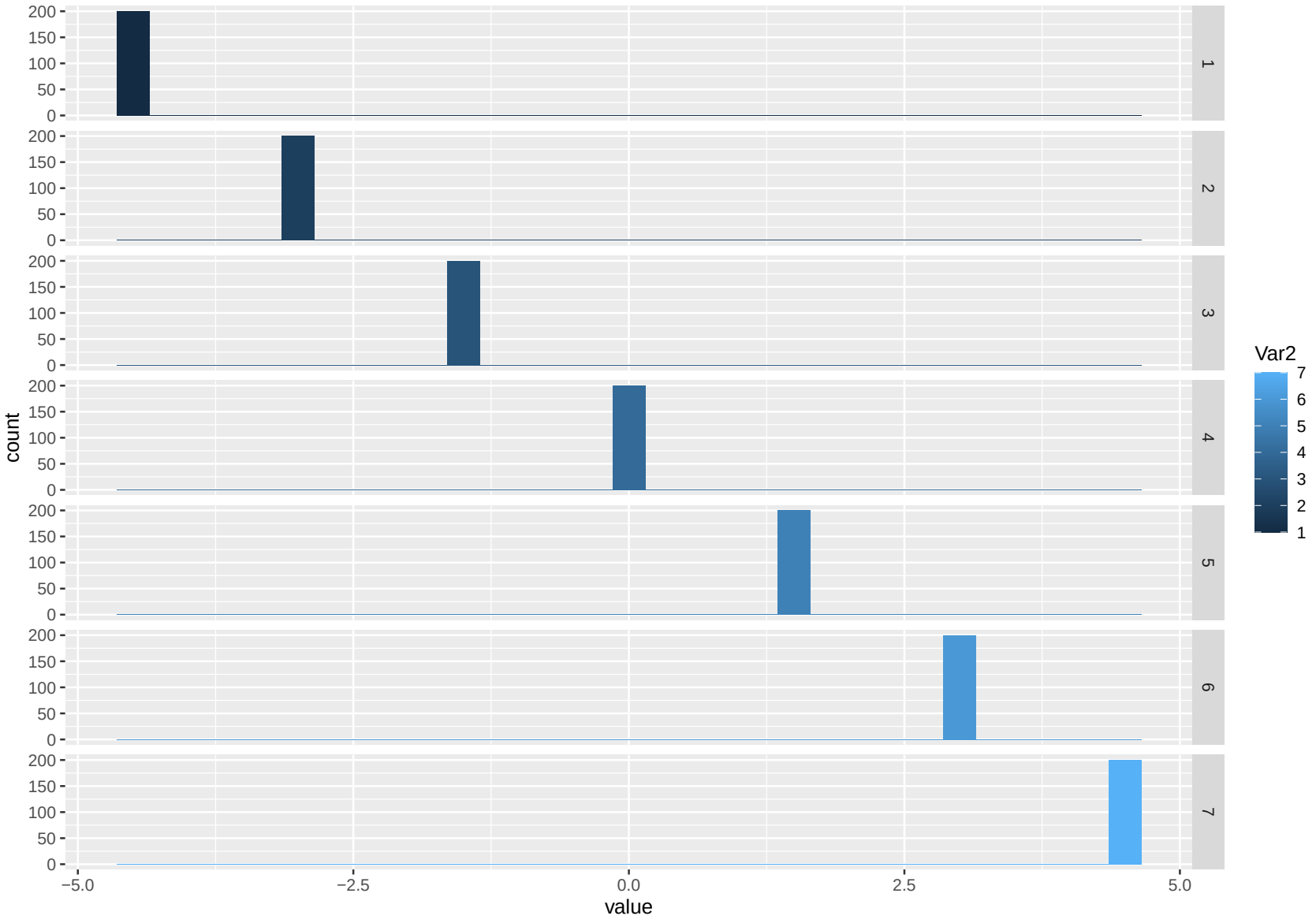
```
##
## Call:
## lm(formula = abalone$Rings ~ matrice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1632  -1.3613  -0.3885   0.9054  13.7440
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.93368    0.03432 289.481  < 2e-16 ***
## matriceLongestShell   -0.18877    0.21914  -0.861    0.389
## matriceDiameter        1.32594    0.22201   5.972 2.53e-09 ***
## matriceHeight          0.49465    0.06475   7.639 2.70e-14 ***
## matriceWholeWeight     4.53483    0.35928  12.622  < 2e-16 ***
## matriceShuckedWeight -4.48674    0.18274 -24.552  < 2e-16 ***
## matriceVisceraWeight -1.07747    0.14294  -7.538 5.82e-14 ***
## matriceShellWeight     1.19384    0.15824   7.545 5.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.218 on 4169 degrees of freedom
## Multiple R-squared:  0.5276, Adjusted R-squared:  0.5268
## F-statistic: 665.2 on 7 and 4169 DF,  p-value: < 2.2e-16
```

Détermination des données nécéssaires à la simulation

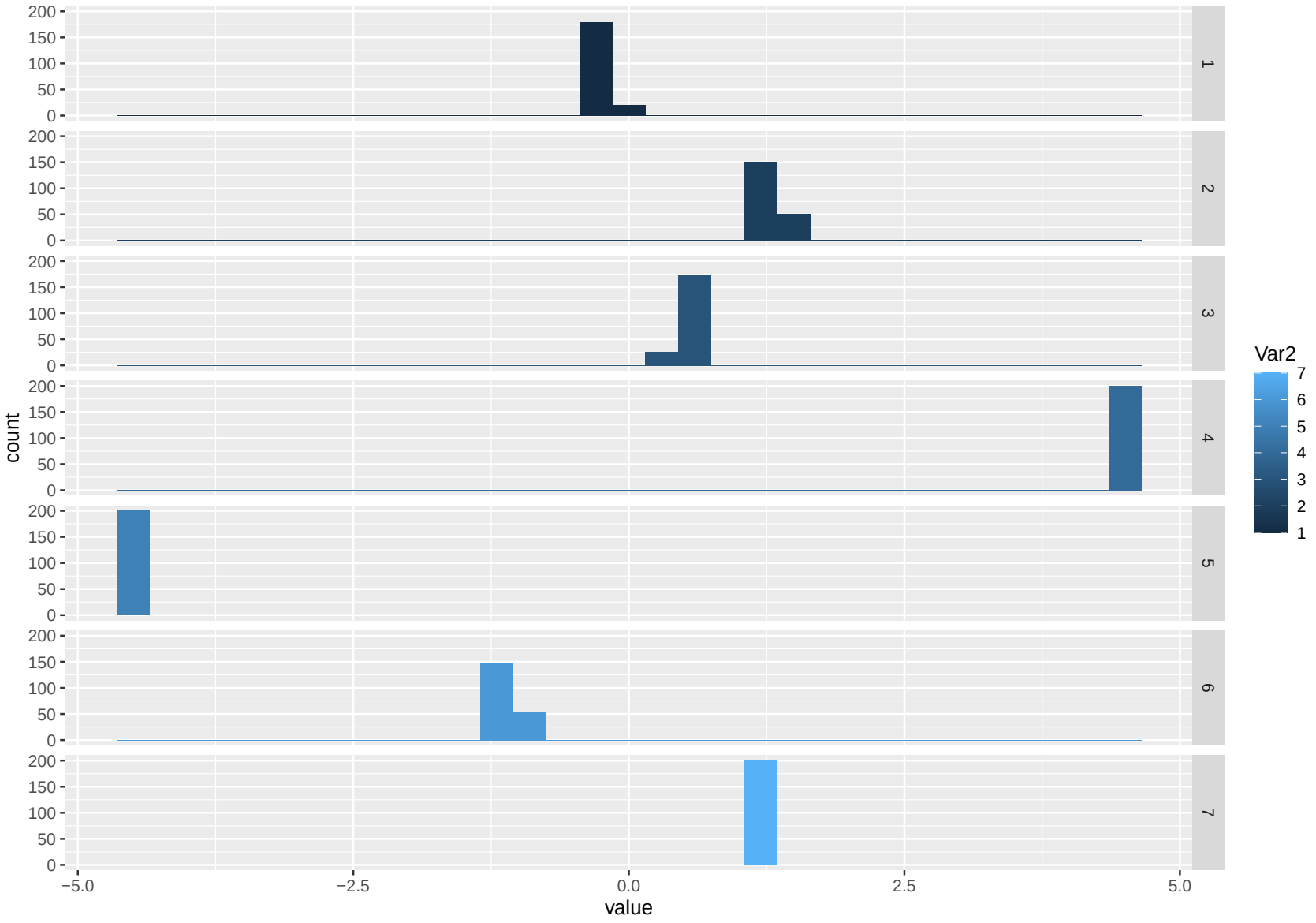
- A l'aide du modèle linéaire on détermine la valeur des coefficients associés à chaque paramètre.
- Calcul de $\varepsilon = Y - X(X^T X)^{-1} X^T Y$.
- Calcul de σ

```
## [1] 0.6872918
```

Histogramme de l'estimation des coefficients $\hat{a}_i$



Comparaison avec les vraies valeurs coeff_abalone



Conclusion

- La précision des outils de mesures.
- La perte d'information au cours des mesures.
- Des variables explicatives non exhaustives.
- Un modèle de régression pas assez précis.

Perspectives

- Compréhension du sujet.
- Utilisation de ggplot.
- Utilisation du Latex.
- Github.
- Géneration des package.

Reférences

- https://web.maths.unsw.edu.au/~lafaye/textes/livreR2014.pdf
- https://tice.agroparistech.fr/coursenligne/courses/STAV/document/Poly/ModLin_CoursExemples_R.pdf?cidReq=STAV
- https://stackoverflow.com/
- https://www.rdocumentation.org