Alternatives aux Moindres Carrés Ordinaires

Sabaye Fried-Junior, Boussengui François, Barré Nicolas

24 décembre 2020

Table des matières

Préambule

Dans ce document, il sera tout d'abord question de présenter des concepts mathématiques et statistiques communément utilisés en économétrie. Dans un premier temps, nous reviendrons sur un concept majeur et fondamental en économétrie à savoir la **covariance**. Il s'agira de revenir sur la définition de la covariance communément connue puis d'en proposer une définition alternative avec à l'appui une représentation graphique de celle-ci.

Ensuite, nous proposons une représentation alternative ou plus précisément une représentation diagrammatique de la méthode des moindre carrés ordinaires. Cette outi graphique nous permettra de visualiser tous les principaux résultats de la méthode des moindres carrés ordinaires.

1 Covariance

1.1 Covariance défnie en fonction du centre de gravité

Considérons deux variables aléatoires $X,\ Y$, et un échantillon simple bivarié $(x_1,\ y_1), (x_2,\ y_2), \cdots (x_N,\ y_N)$.

Hypothèses :

- Les observations sont indépendantes entre elles
- Les variables X et Y ne sont pas nécessairement indépendantes

Definition :

Un estimateur sans biais de la covariance entre 2 variables X et Y, avec N observations est donnée par :

$$Cov(x,\ y) = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

ce qui équivaut à :

$$Cov(x,\ y) = \frac{N}{N-1} (\overline{xy} - \overline{x}\ \overline{y})$$

1.2 Revue de littérature

Rares sont les auteurs qui ont proposé une alternative à la représentation de la covariance, notion centrale en économétrie.

La proposition la plus célèbre en date est celle de l'economiste Kevin Hayes dans un article intitulé : "A Geometrical Interpretation of an Alternative Formula for the Sample Covariance", datant de 2011.

1.2.1 Proposition de Kevin HAYES

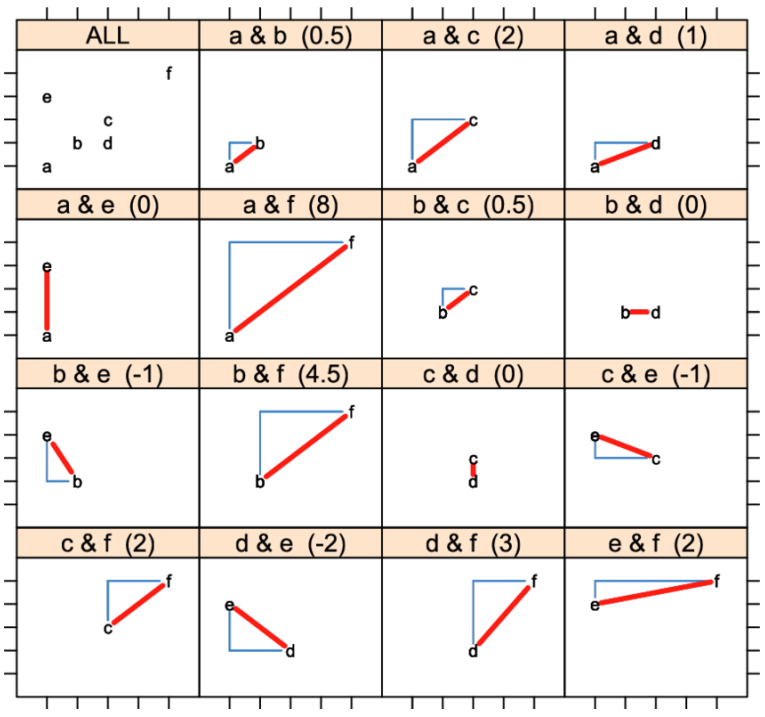
Considérons deux variables aléatoires $X,\ Y$, et un échantillon simple bivarié $(x_1,\ y_1), (x_2,\ y_2), \cdots (x_N,\ y_N)$.

Hayes choisit de représenter la covariance entre deux points (x, y) par le triangle au dessus (en-dessous) de l'hypothénuse selon que la pente entre les deux points soit positive ou négative. Cependant, il ne propose pas une représentation générale pour un nuage de points mais plutôt et uniquement une représentation pour deux points. De plus, il est important de souligner que, selon lui, lorsque nous sommes confrontés à des points exaequo, le triangle est dégénéré et est représenter par une droite.

1.2.2 Illustration :

Pour illustrer sa théorie, Hayes, utilise un échantillon de n = 6 observations bivariées:

Point	a	b	c	d	e	f
x	1	2	3	3	1	5
y	1	2	3	2	4	5



1.3 Covariance définie en fonction de la comparaison des couples de points

1.4 Un peu de théorie

Considerons deux variables aléatoires $X,\ Y$, et un échantillon aléatoire simple bivarié $(x_1,\ y_1), (x_2,\ y_2), \cdots (x_N,\ y_N)$

Prenons deux observations au hasard k, l

$$\mathbb{E}(x_k - x_l)(y_k - y_l)$$

On a donc :

$$\mathbb{E}(x_k y_k - x_k y_l - x_l y_k + x_l y_l)$$
$$\mathbb{E}(x_k y_k) - \mathbb{E}(x_k)\mathbb{E}(y_l) - \mathbb{E}(x_l)\mathbb{E}(y_k) + \mathbb{E}(x_l y_l)$$
$$2\mathbb{E}(xy) - 2\mathbb{E}(x)\mathbb{E}(y)$$

Il est important de rappeler les hypothèses suivantes : Premièrement, les observations sont indépendantes entre elles. Cependant, ce n'est pas nécessairement le cas entre les variables X, Y. Donc pour un couple de points choisi au hasard dans l'échantillon, l'expression théorique de l'espérance mathématique de l'aire du rectangle que formé par deux points est :

$$\mathbb{E}(x_k - x_l)(y_k - y_l) = 2\ Cov(x, y)$$

Une définition équivalente de la covariance : La mesure de la covariance de l'échantillon entre deux variables aléatoires, avec N observations, est la moyenne empirique suivante :

$$\frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)$$

$$\frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j)$$

Preuve :

$$\frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j) =$$

$$\frac{1}{2N(N-1)} \sum_{i=1}^{N} Nx_i y_i - x_i N\bar{y} - y_i N\bar{x} + N\bar{x}\bar{y} =$$

$$\frac{2N^2}{2N(N-1)} (\bar{xy} - \bar{x}\bar{y}) =$$

$$\frac{N}{N-1} (\bar{xy} - \bar{x}\bar{y}) = Cov(x, y)$$

1.5 Construction graphique

$$

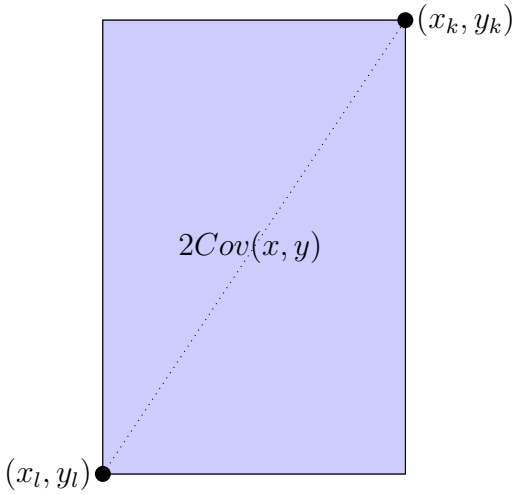


FIGURE 1 – Case 1 : Positive correlation

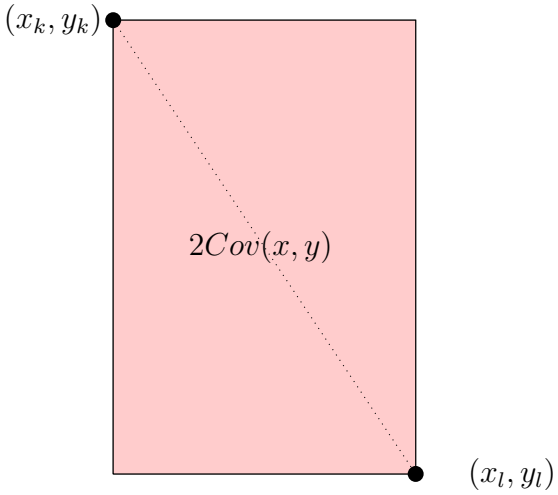
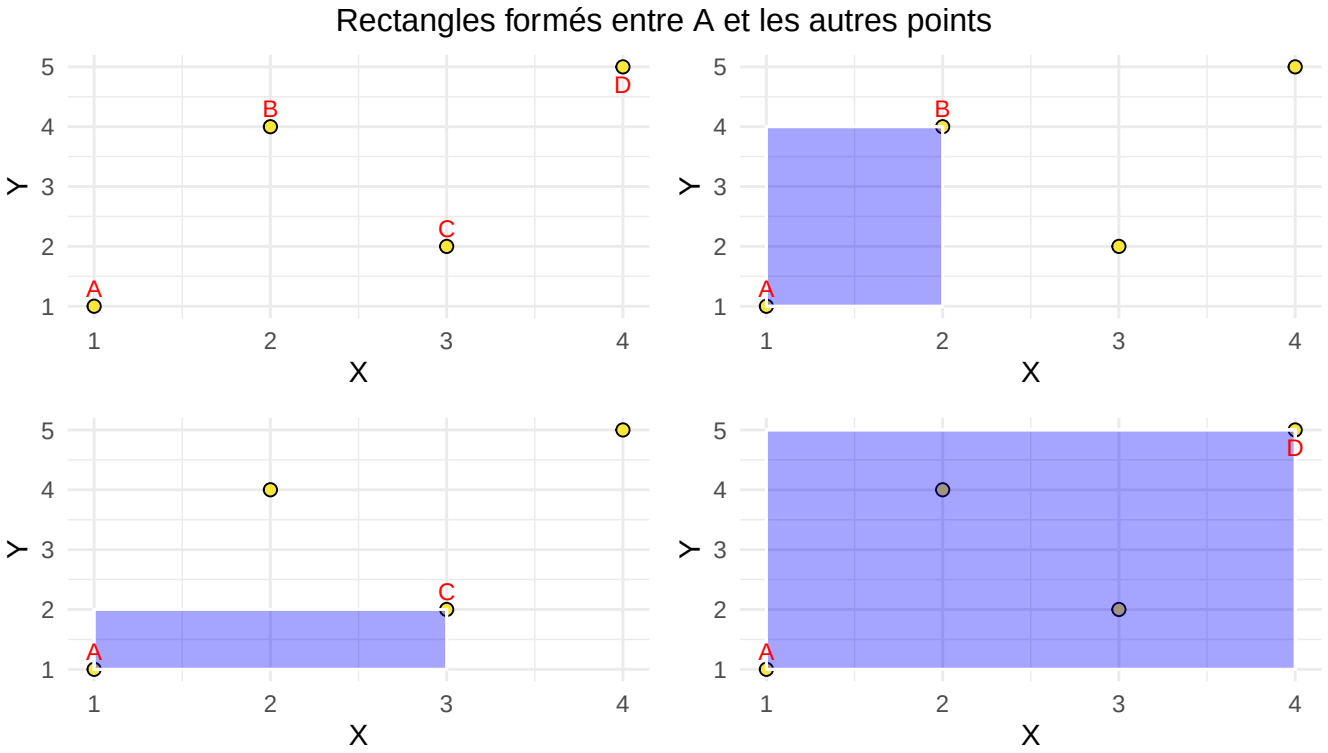
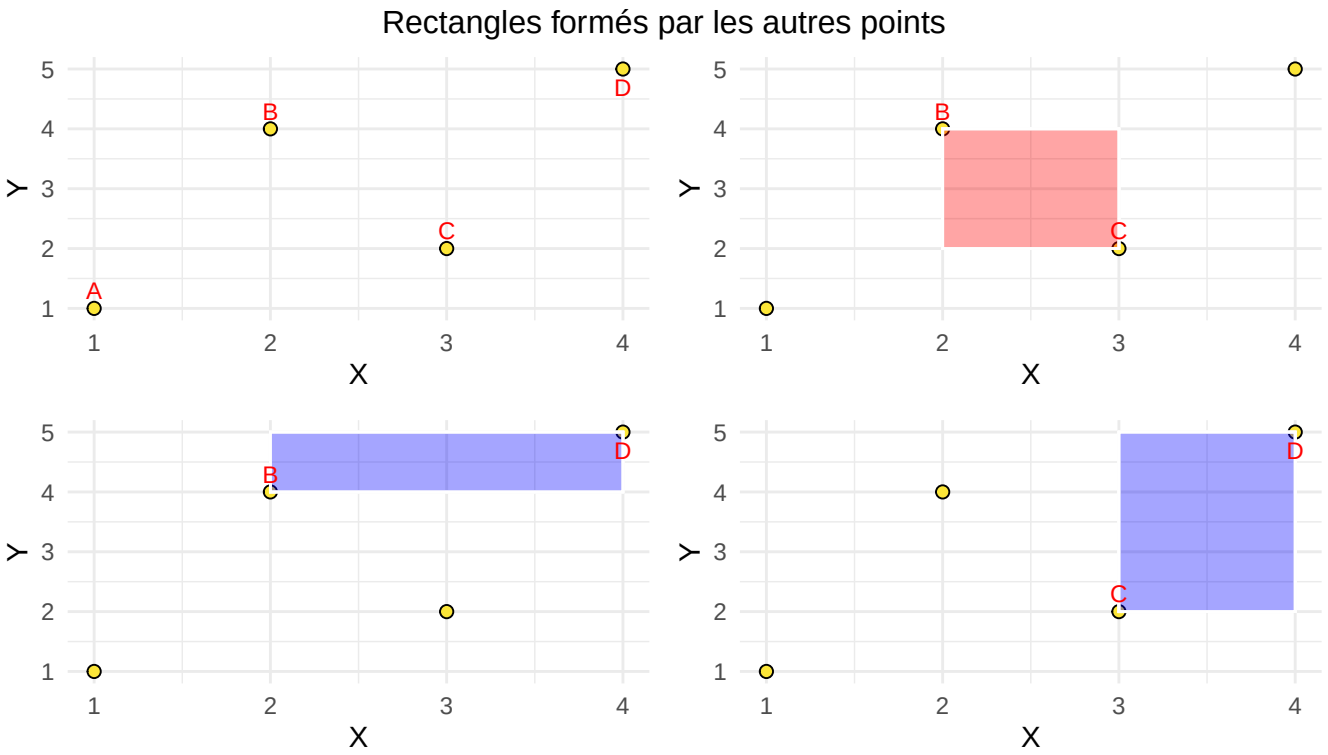


FIGURE 1 – Case 2 : Negative correlation

$$

1.5.1 Principe de supperposition

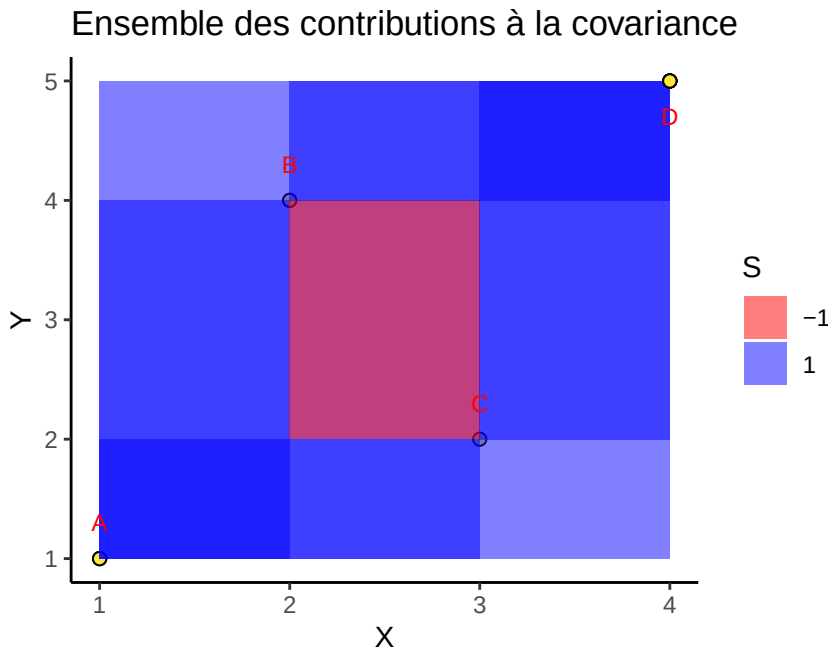




Remarques :

- On a vu précedemment que la moyenne de toutes les aires représentées était égale à $2Cov(X, Y)$
- Nous avons représenté la covariance entre chaque couple de points, le plan x, y a ainsi été séparé en plusieurs quadrillages en fonction des différentes covariances susmentionnées
- L'étape suivante consiste à superposer toutes les covariances

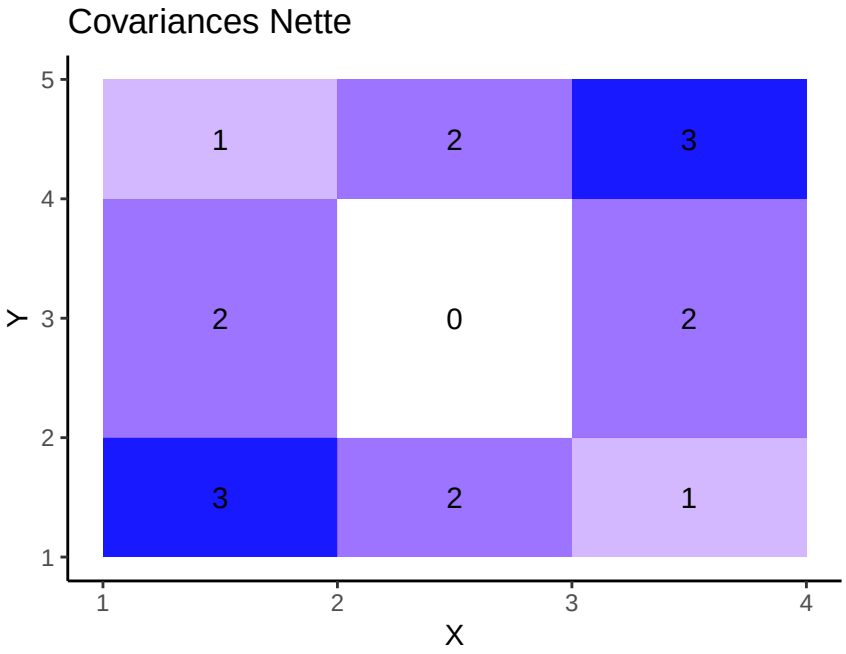
1.6 Résultat



1.7 Nouvelle représentation de la covariance

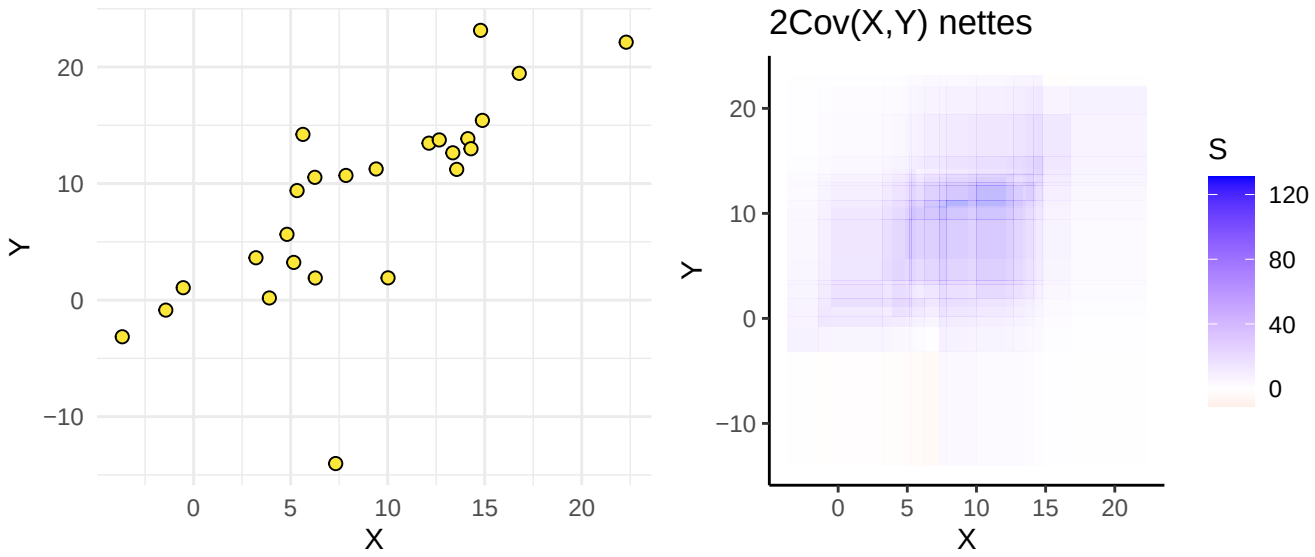
Étant donné que les différentes doubles covariances se supperposent dans le zonage, l'idée est de calculer la somme des couches qui se supperposent dans chaque pavage et d'attribuer une couleur à chacun d'eux en fonction du résultat.

Remarque : Lorsque dans un pavage il y a autant de bleu que de rouge, la somme des double covariances donne 0.



1.8 Extension

Avec 25 points



2 Diagrammes de Régression

2.1 Rappel

Il a été présenté dans la première partie qu'une façon intuitive de représenter la relation que caractérise deux variables aléatoires X et Y d'un échantillon simple bivarié $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ est de choisir deux observations au hasard, (x_k, y_k) et (x_l, y_l), et de calculer alors l'espérance de l'aire du rectangle formé par ces deux points, défini comme :

$$\mathbb{E}[(x_k - x_l)(y_k - y_l)]$$

Sous l'hypothèse d'indépendance des observations k et l, on rapelle qu'on peut démontrer que :

$$\begin{aligned}\mathbb{E}[(x_k - x_l)(y_k - y_l)] &= \mathbb{E}[x_k y_k - x_k y_l - x_l y_k + x_l y_l] \\ &= \mathbb{E}(x_k y_k) - \mathbb{E}(x_k y_l) - \mathbb{E}(x_l y_k) + \mathbb{E}(x_l y_l) \\ &= 2\mathbb{E}(xy) - 2\mathbb{E}(x)\mathbb{E}(y) \\ &= 2Cov(x, y)\end{aligned}$$

Cette démonstration indique que l'aire moyenne des $\frac{N(N-1)}{2}$ rectangles construits à partir de données est un estimateur du double de la covariance entre les variables X et Y. Cette construction à été nommée Average Rectangle Tool.

Remarques :

— Les mêmes principes s'appliquent dans le cas particulier de la variance:

$$\mathbb{E}[(x_k - x_l)^2] = 2\ Cov(X, X) = 2\ Var(X).$$

— On notera ρ le coefficient de correlation linéaire de Pearson défini comme $\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$.

2.2 Le modèle de Régression Linéaire Simple

Considérons le modèle suivant :

$$y = \alpha + \beta x + \epsilon$$

Sous les hypothèses usuelles, on a l'identité $Cov(X, Y) = Cov(x, \alpha + \beta x + \epsilon) = \beta Var(x)$. Opérer une régression de y sur x signifie que la covariance pourrait être représentée par un rectangle (une forme représentative des $\frac{N(N-1)}{2}$ rectangles), et les variances de ces variables par des carrés. C'est la première hypothèse que nous devons poser pour construire ce diagramme. Le double de la valeur empirique de la covariance donne l'aire du rectangle, et nous imposerons que la valeur empirique de l'écart-type de la variable explicative est égale à la taille de l'un des côtés de ce rectangle, pour des raisons d'identification. Ce sera la deuxième et dernière hypothèse pour cette partie.

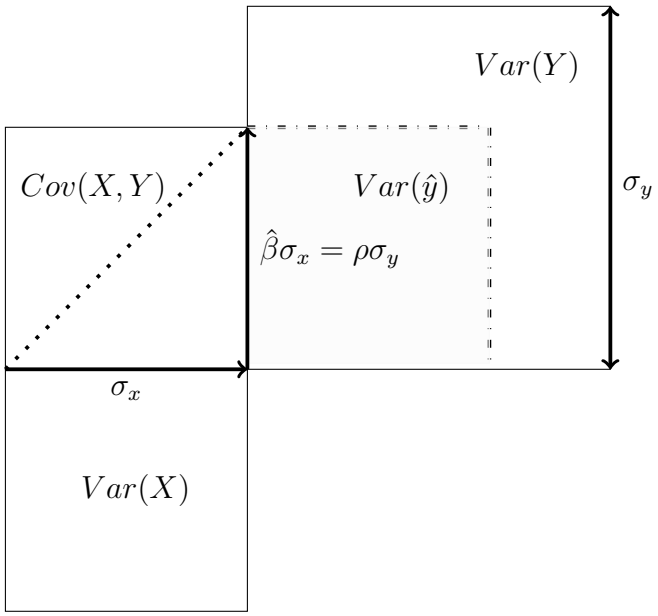
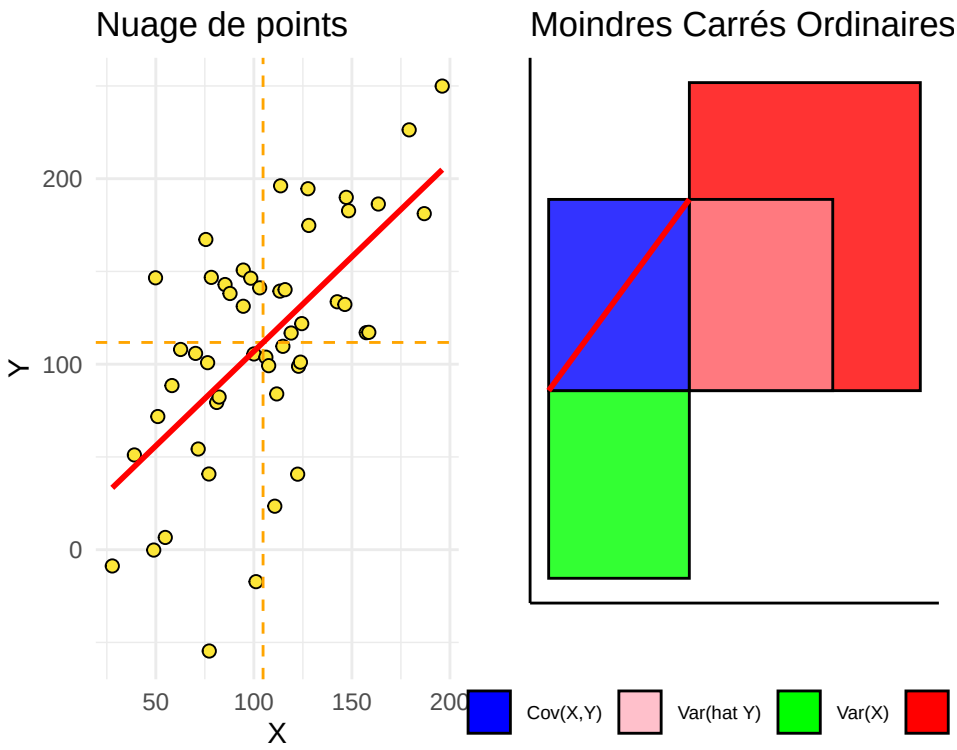


FIGURE 2 – régression de Y sur X

Cette visualisation montre le lien entre l'estimateur des Moindres Carrés Ordinaires $\hat{\beta}$ et l'Average Rectangle Tool. Le rectangle de covariance à été dessinée tel que ses côtés de longeurs $(\sigma_x, \rho\sigma_y)$ vérifient que l'aire de ce rectangle soit $Cov(X, Y)$. Ainsi, la diagonale tracée dans ce rectangle est de pente $\rho\frac{\sigma_y}{\sigma_x} = \hat{\beta}$, soit précisément l'effet marginal de X sur Y, toutes choses égales par ailleurs. En outre, ce dessin permet de représenter le R^2 comme une proportion de la variance expliquée $Var(\hat{Y})$ dans la variance totale $Var(Y)$.

2.2.1 Exemple

Voici une représentation diagrammatique d'un modèle de régression linéaire simple, réalisé sur des données simulées.



2.3 La régression des Moindres Rectangles

Il est possible d'étendre les principes exposés dans la partie consacrée aux Moindres Carrés Ordinaires à d'autres techniques d'estimation telle que la régression des moindres rectangles.

$$\underset{\{\alpha,\beta\}}{\text{Min}} \sum_{i=1}^{N} |Y_i - \alpha - \beta X_i| . |X_i - \frac{1}{\beta}(Y_i - \alpha)|$$

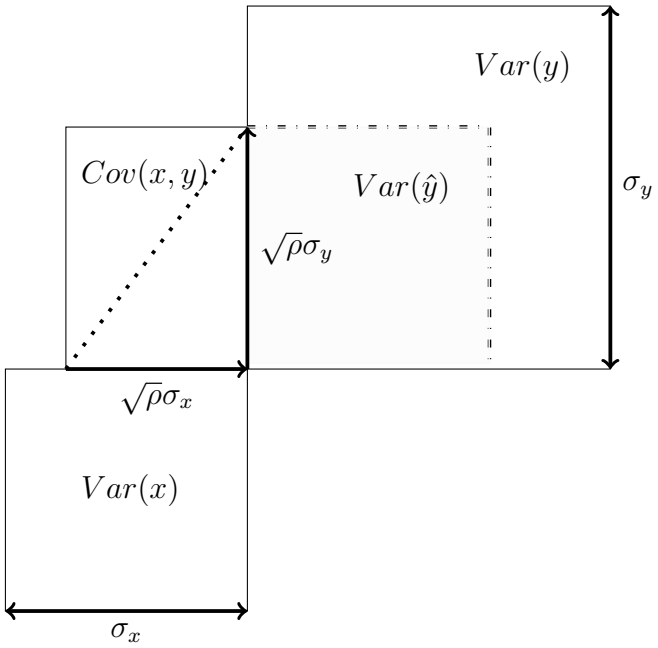
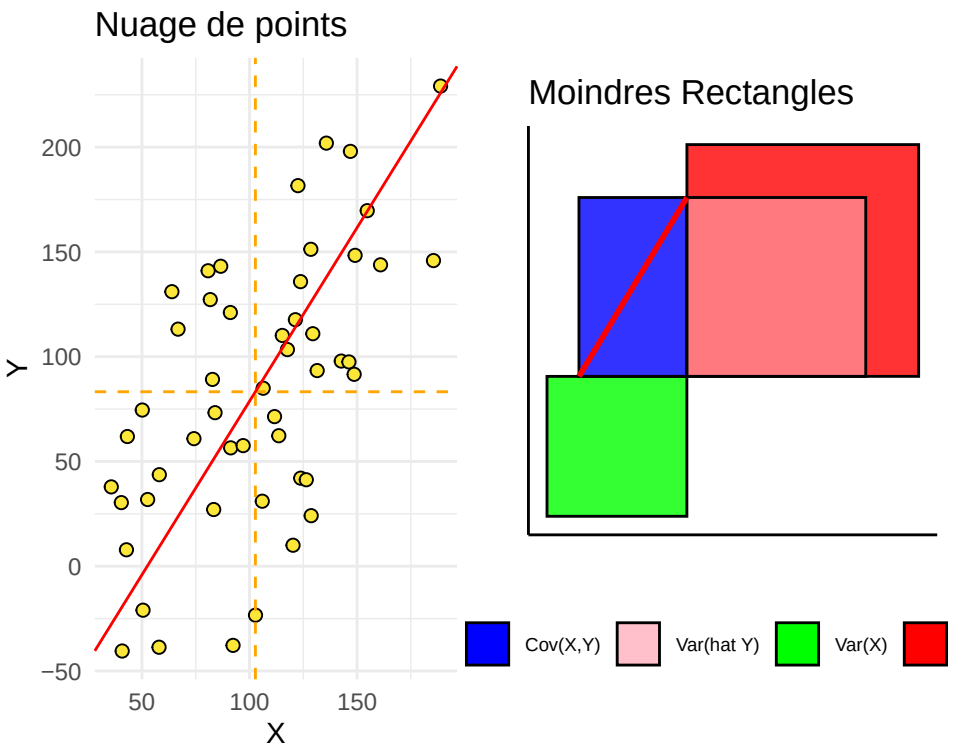


FIGURE 3 – Régression par Moindres Rectangles

Dans cette figure, le rectangle représentant la covariance a ses côtés fixés respectivement à $\sqrt{\rho}\ \sigma_x$ et $\sqrt{\rho}\ \sigma_y$. La diagonale tracée dans ce rectangle est de pente $\frac{\sigma_y}{\sigma_x}$.

2.3.1 Exemple

Voici un exemple de repésentation diagrammatique des résultats d'une régression par Moindres Rectangles appliqué à des données simulées.



2.4 Automatisation dans R

L'équipe en charge du sujet a pu commencer à développer des outils sous le logiciel R qui assurent la production visuelle de cette construction avec des vraies données. Un détail du code constituant l'algorithme générant des diagrammes de régression sera fourni en annexe.

L'objectif est de travailler avec Shiny pour fournir une application interactive qui permettra à l'utilisateur de visualiser les résultats principaux en faisant varier des paramètres sensibles, tels que le coefficient de correlation linéaire, la taille de l'échantillon, entre autres. Le travail de cette partie consistera en l'adaptation et le développement d'objets R dans une démarche d'aide à la visualisation des mécanismes latents en économétrie basique, et à la traitance de cas particuliers ou dégénérés pouvant émerger dans le programme.

3 Annexe

3.1 Références

— Bousquet, A. and Concettini, S. (2020), "ART in OLS"
— Hayes, K. (2011) "A Geometrical Interpretation of an Alternative Formula for the Sample Covariance" , The American Statistician, 65 (2), 110-112
— Li, C.C. (1964) "Two additional views of Linear Regression coefficients", The American Statistician, The Teachers' Corner, 18 (4), 27-28
— Peter M. Hefferman (1988) "New Measures of Spread and a Simpler Formula for the Normal Distribution", The American Statistician, Vol. 42, No. 2 (May, 1988), pp 100-102
— A Giloni, M. PADBERG (2002), "Alternative Methods of Linear Regression", Mathematical and Computer Modelling 35, p. 361-374

3.2 Code

L'algorithme permettant de générer les représentations diagrammatiques, prenant en compte la méthode choisie :

```
Diag <- function(X,Y, method = "ols", norm = F, label = ""){
  Out <- NULL
  if(norm == TRUE){
    X = (X-mean(X))/sd(X)
    Y = (Y-mean(Y))/sd(Y)
  }

  if(method == "ols"){
    dols=data.frame(x1=c(0,0,sd(X),sd(X)),
                    x2=c(sd(X),sd(X),sd(X)+sd(Y),sd(X)+abs(cov(X,Y))/sd(X)),
                    y1=c(0,sd(X),sd(X),sd(X)),
                    y2=c(sd(X),sd(X)+abs(cov(X,Y))/sd(X),sd(X)+sd(Y),sd(X)
                         +abs(cov(X,Y))/sd(X)),
                  t=c('Var(X)','Cov(X,Y)','Var(Y)','Var(hat Y)'), r=c(1,2,3,4))
    Out<- ggplot() +
      geom_rect(data=dols, mapping=aes(xmin=x1, xmax=x2, ymin=y1, ymax=y2, fill=t),
                color="black", alpha=c(0.8,0.8,0.8,0.5))+
      scale_fill_manual(values=c('blue','pink','green','red'))+
      geom_segment(aes(x = 0, y = sd(X), xend = sd(X), yend = sd(X)+abs(cov(X,Y))/sd(X)),
                   colour="red",size=1)+
      #geom_rect_pattern(aes(),pattern = "stripe" , fill = "green", colour = "black")+
      theme(panel.border = element_blank(), panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"))+
      theme(legend.title=element_blank(),legend.text = element_text(size = 7))+
      theme(legend.position="bottom")+
      xlab(" ")+
      ylab(" ")+
      ggtitle(label=label)
```

```
  }

  if(method == "lr"){
    p = (abs(cov(X,Y)))/(sd(Y)*sd(X))
    dols=data.frame(x1=c(0,sd(X) - sqrt(p)*sd(X),sd(X),sd(X)),
                    x2=c(sd(X),sd(X),sd(X)+sd(Y),sd(X)+sqrt(p)*sd(Y)),
                    y1=c(0,sd(X),sd(X),sd(X)),
                    y2=c(sd(X),sd(X)+sqrt(p)*sd(Y),sd(X)+sd(Y),sd(X)+sqrt(p)*sd(Y)),
                  t=c('Var(X)','Cov(X,Y)','Var(Y)','Var(hat Y)'), r=c(1,2,3,4))

    Out<- ggplot() +
      geom_rect(data=dols, mapping=aes(xmin=x1, xmax=x2, ymin=y1, ymax=y2, fill=t),
                color="black", alpha=c(0.8,0.8,0.8,0.5))+
      geom_segment(aes(x = sd(X) - sqrt(p)*sd(X), y = sd(X),
                       xend = sd(X), yend = sd(X)+sqrt(p)*sd(Y)),
                   colour="red",size=1)+
      scale_fill_manual(values=c('blue','pink','green','red'))+
      theme(panel.border = element_blank(), panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),aspect.ratio = 1)+
      theme(legend.title=element_blank(),legend.text = element_text(size = 7))+
      theme(legend.position="bottom")+
      xlab(" ")+
      ylab(" ")+
      ggtitle(label=label)
  }
  return(Out)
}
```

Cette fonction est un prototype d'un des composants de ce qui servira à la création d'un package de visualisation reprennant les éléments cités au début de cette partie.