# Couples aléatoires, régression linéaire

13 décembre 2016

## 1 Simulation d'une expérience de dés

### Exercice 1 (*Min et max de deux échantillons*)

**1.** Obtenir deux échantillons `X` et `Y` de loi géométrique $\mathcal{G}(\frac{1}{10})$.

**2.** Vérifier que la minimum des variables géométriques est aussi de loi géométrique.
Est-ce aussi vrai pour le maximum ? *(syntaxe* `min(X, Y)` *et* `max(X, Y)`*.)*

### Exercice 2 (*Ranger dans l'ordre*)

**1. Pour un vecteur**

**a)** Obtenir un échantillon aléatoire `u = floor (10 * rand(1, 5))`.

**b)** Que retournent les commandes `min(u)`, et `max(u)` ?

**c)** Que retourne la commande `gsort(u)` ?

**2. Pour une matrice**

**a)** Obtenir un échantillon `v = floor (10 * rand(3, 5))`.

**b)** Que retournent les commandes `min(u)`, et `min(u,"r")`, et `min(u,"c")` ?

**c)** Que retournent les commandes `gsort(u)`, et `gsort(u,"r")`, et `gsort(u,"c")` ?

### Exercice 3 (*Une régression linéaire à l'œil nu*)

On engendre trois nombres aléatoires uniformes sur $(0\,;1)$, et on les range dans l'ordre : $X_1 \leqslant X_2 \leqslant X_3$.
On s'intéresse au couple de $M = X_2$ et $S = X_1 + X_3$.

**1.** Simuler l'expérience en obtenant un échantillon `echTrie`
*(grâce aux commandes* `rand(3,N)` *et* `gsort`*, option "r").*

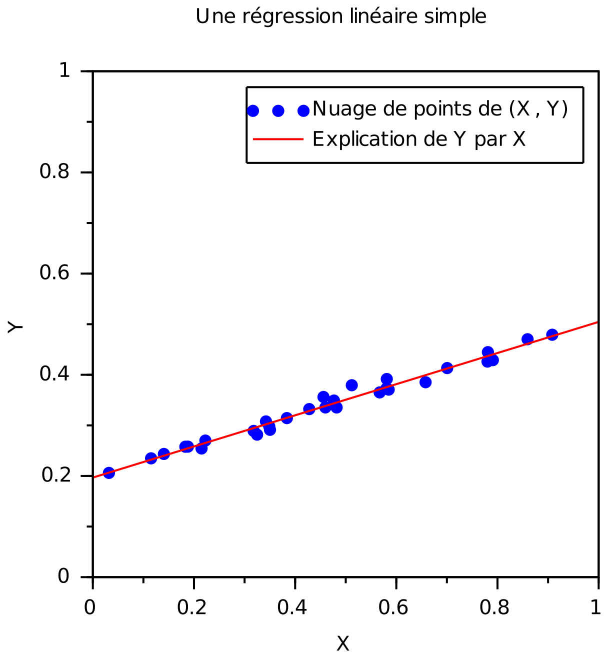
**2.** Définir les échantillons pour `M` et `S`. *(extraire les lignes avec* `M = echTrie(2,:)`*)*

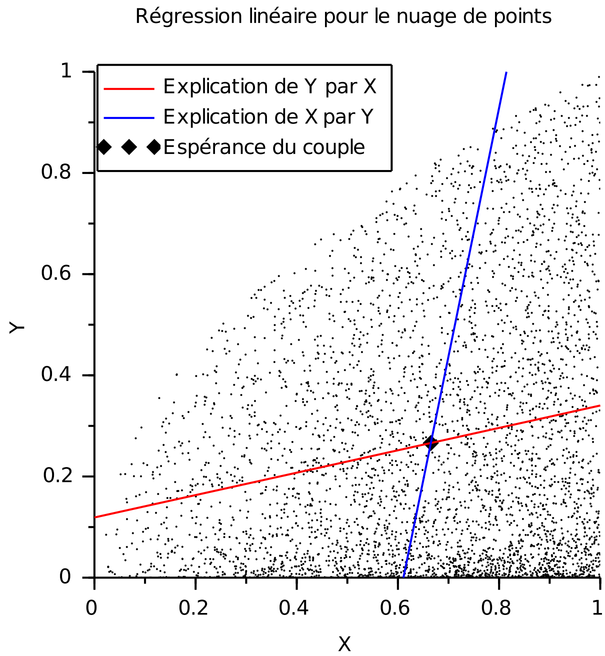
**3.** `plot`er le nuage de points avec la cosmétique adaptée.
Quelle semble être la tendance entre `M` et `S` ?

**4.** Tracer au dessus du nuage de points la droite d'équation $y = 2x$.

# 2 Un exemple de régression linéaire en macroéconomie

## 2.1 Illustration de la régression linéaire




**1.** **a)** Engendrer un échantillon $X$ de loi $\mathcal{U}[0,1]$ de 20 points

**b)** Engendrer un échantillon pour $Y = X/2 + \epsilon$ avec $\epsilon \hookrightarrow \mathcal{N}(0.3, 0.05)$.

**c)** Représenter par un nuage de points. Placer la moyenne empirique. Que constate-t-on ?

**d)** Utiliser la commande `reglin` pour faire la régression linéaire.

**e)** Représenter la droite de régression en rouge.

```
a = gca ()
a.isoview = 1
a.data_bounds = [0 , 0 ; 1 , 1 ]
a.tight_limits = "on"
```

**2.** Même questions pour $X$ , $Y = X \times U^2$, où $X, U \hookrightarrow \mathcal{U}[0,1]$ avec 5000 points,

```
// Syntaxe de la commande reglin
[a , b , sig] = reglin (x , y) //ou bien
[a , b ] = reglin (x , y)
```

## 2.2 Vérification de la loi d'Okun

On va chercher l'**élasticité** historique du taux de chômage sur la croissance économique.

- Le fichier `statFrance.sce` contient les données de l'INSEE pour le PIB et le taux de chômage pour la France
- Le fichier `statUSA.sce` contient les données de l'OCDE pour le PIB et le taux de chômage pour les États-Unis d'Amérique.

# 3 Le principe de la régression linéaire

## 3.1 Traitement mathématique

⋆) *Variables :* On se donne deux variables $X, Y$ : ▸ $X$ est la **variable explicative**
▸ $Y$ est la **variable expliquée**

⋆) *Ajustement affine :* Pour $\alpha, \beta$ des constantes déterministes, on pose $Y' = \alpha X + \beta$.

⋆) *Comparaison :* On compare $Y$ à $Y' = \alpha X + \beta$. L'erreur $\epsilon = Y - Y'$ s'appelle le **résidu**.

⋆) *Optimisation :* On cherche le couple $\alpha, \beta$ qui rende le résidu $\epsilon$ « aussi petit que possible »

⋆) *Moindres carrés :* On minimisera ici $\mathbb{E}[\epsilon^2] = \mathbb{E}[(Y - Y')^2]$ ***(erreur quadratique moyenne)***

**Proposition 1**

▸ **Variables centrées réduites** On suppose ici $\begin{cases} \mathbb{E}[X] = \mathbb{E}[Y] = 0, \\ \mathrm{Var}(X) = \mathrm{Var}(Y) = 1. \end{cases}$

Alors **l'optimisation au sens des moindres carrés** est réalisée pour l'ajustement affine $Y' = \alpha X + \beta$ avec

$$\alpha = \rho(X, Y) \qquad\qquad \beta = 0.$$

On note *(temporairement)* : $Y \approx Y' = \rho X$

▸ **Formule générale** Pour les versions centrées réduites : $\frac{Y - m_Y}{\sigma_Y} = Y^* \approx \rho X^* = \rho \frac{X - m_X}{\sigma_X}$.

On développe et on trouve l'ajustement affine

$$\begin{aligned} Y' &= \rho \frac{\sigma_Y}{\sigma_X}(X - m_X) + m_Y \\ &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - m_X) + m_Y \end{aligned}$$

**Remarques**

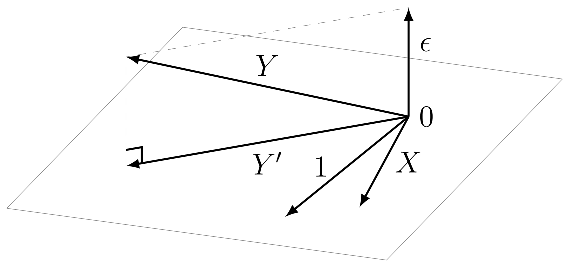
▸ **Espérance, covariance** L'ajustement affine $Y' = \rho\frac{\sigma_Y}{\sigma_X}(X - m_x) + m_y$ est le seul qui satisfasse simultanément :

$$\mathbb{E}[Y'] = \mathbb{E}[Y] \qquad \mathrm{Cov}(X, Y') = \mathrm{Cov}(X, Y)$$

▸ **Orthogonalité du résidu** En écrivant $\epsilon = Y - Y'$, on trouve une relation « à la Pythagore » pour $Y = Y' + \epsilon$ :

$$\mathrm{Cov}(Y', \epsilon) = 0 \qquad \sigma_Y^2 = \sigma_{Y'}^2 + \sigma_\epsilon^2$$

Ainsi $Y'$ s'interprète géométriquement comme la **projection orthogonale** de $Y$ sur le plan $\mathrm{Vect}(1, X)$ ; c'est comme une ombre portée par le vecteur $Y$ *(vide contra)*.



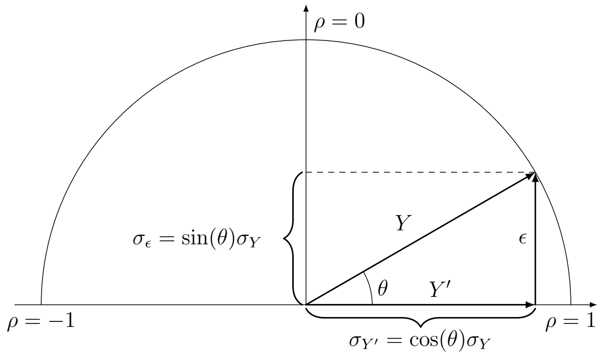

**Définition 2 (*Le coefficient de détermination*)**

Ici, c'est simplement le coefficient de corrélation, noté $R = \rho(X, Y)$. Il s'interprète comme un **indicateur de la qualité** de $X$ comme variable explicative affine de $Y$.

En effet, dans l'écriture pour $Y = Y' + \epsilon$, on a : $\sigma_Y^2 = \sigma_{Y'}^2 + \sigma_\epsilon^2$. La variance de $Y$ se décompose donc en deux parts, dont les écarts-types associés sont :

- **la part expliquée** $\frac{\sigma_{Y'}}{\sigma_Y} = R = \cos(\theta)$
- **la part résiduelle**

$$\frac{\sigma_\epsilon}{\sigma_Y} = \sqrt{1 - R^2} = \sin(\theta)$$




Par exemple ci-contre, la variable $X$ :

- explique correctement $Y_1$ et $Y_4$,
- explique imparfaitement $Y_2$
- explique mal $Y_3$
  *($X$ et $Y_3$ sont presque décorrélées)*