# TP Analyse de données

## Séance 2 – La régression linéaire

**Rappel du cours:**
On a mesuré deux variables $x$ et $y$ pour une population de $n$ individus. Le coefficient de corrélation de $x$ et $y$ est donné par

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} \ .$$

La droite de régression linéaire est la droite d'équation $y = ax + b$, où

$$a = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \ , \qquad b = \bar{y} - a\bar{x} \ .$$

### 1. Calcul "manuel"

1. Créer deux vecteurs $x$ et $y$ tels que
$$x = (1.2, 7.4, 3.6, 5.8, 2.2, 9.1, 6.4, 5.6)$$
$$y = (9, 78, 40, 51, 29, 88, 53, 59)$$
2. Calculer le coefficient de corrélation de $r$ de $x$ et $y$ (la racine est implémentée par la commande `sqrt()`). Les données sont-elles indépendantes?
3. Calculer les coefficients de régression linéaire $a$ et $b$ et stocker leur valeur dans deux variables `a` et `b`.
4. Visualiser le résultat graphiquement à l'aide des commandes
```
plot(x,y)
abline(b,a)
```

### 2. La fonction `lm`

La fonction `lm` de `R` permet d'automatiser le calcul de régressions linéaires (et de bien d'autres types d'approximations).

1. Commençons par transformer nos données pour les mettre dans un tableau, à l'aide de la commande
```
tableau <- data.frame(x,y)
```
Visualiser le contenu de la variable `tableau`.
2. On adopte le modèle de régression linéaire, dénoté dans `R` par `y ~ x`. Créer une variable `regression` contenant le résultat de la commande
```
lm(y ~ x, data=tableau)
```
3. Représenter graphiquement le tableau à l'aide de la commande `plot`, puis ajouter la droite de régression à l'aide de la commande `abline(regression)`.

### 3. Applications

`R` contient un certain nombre de données préenregistrées. On peut en visualiser la liste à l'aide de la commande `data` (appuyer sur la barre d'espace pour continuer, et sur `q` pour quitter).

Effectuer une régression linéaire sur quelques-unes de ces données, par exemple `cars` et `faithful`.