Calcul du coefficient de corrélation linéaire dans une régression simple

www.adjectif.net/spip/spip.php

mercredi 19 mars 2014 par Mehdi Khaneboubi

Pour citer cet article :

Khaneboubi Mehdi (2014). Calcul du coefficient de corrélation linéaire dans une régression simple. *Adjectif.net* Mis en ligne mercredi 19 mars 2014 [En ligne] http://www.adjectif.net/spip/spip.php?article280

Résumé :

Cet article vient en complément d'un précédent http://www.adjectif.net/spip/spip.p... sur la droite d'ajustement d'un nuage de points à 2 dimensions.

Nous avons vu dans un billet précédent comment obtenir l'équation d'une droite résumant un nuage de point avec le logiciel *R*. Comment savoir si cette droite est un bon résumé du nuage de points ? L'interprétation du coefficient de régression linéaire est un premier élément de réponse.

Mots clés :

Logiciel *R*

D'après l'article de Wikipédia le coefficient de régression « renseigne sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte […]. Une corrélation égale à 0 signifie que les variables ne sont pas corrélées. » Il s'agit donc d'un indicateur de dépendance entre deux variables quantitatives. En somme lorsqu'il est proche de 0, on interprète qu'il n'y a pas de relation linéaire entre les deux variables. Lorsqu'il est proche de 1, il existe une relation croissante et lorsqu'il est proche de -1 il y a une relation décroissante.

Obtention à partir de R

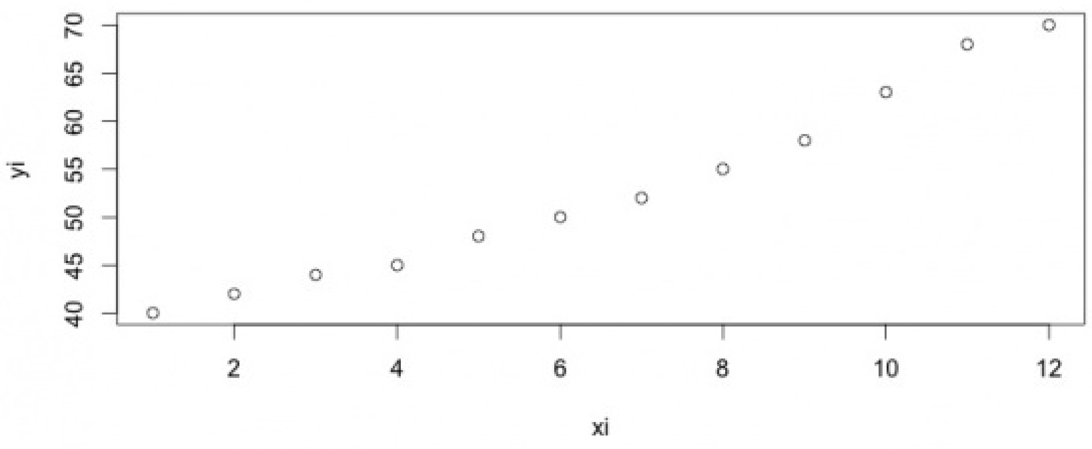
D'abord, importons deux distributions (deux variables quantitatives) dans deux objets que l'on appelle xi et yi :

```
xi <-
c(1:12)

yi <- c(40, 42, 44, 45, 48, 50, 52, 55, 58, 63, 68,
70)
```

La représentation graphique de yi en fonction de yi montre que les points sont à peu près alignés

```
>plot(xi,yi)
```



La commande cor() permet d'afficher directement le coefficient de corrélation linéaire des distributions entre elles. Sans surprise, on obtient un résultat très proche de 1

```
cor(xi, yi)
## [1]
0.9839
```

On voit ici qu'il est très proche de 1. Comme il est positif, lorsqu'une variable augmente, l'autre aussi.

Calcul pas-à-pas

Voyons maintenant comment le calculer pas à pas. L'intérêt de ces calculs réside d'abord dans la familiarisation avec la manipulation d'objets dans *R* mais constitue aussi un moyen de comprendre ce que renvoient les commandes toutes faites de *R*. Nous allons donc calculer un certain nombre de valeurs intermédiaires. Pour une méthode dans Excel, le lecteur pourra consulter Monino *et al*. (2007).

Les valeurs intermédiaires nécessaires sont les suivantes :

- les moyennes, variances et écarts types des deux distributions,
- la covariance du nuage de points,
- l'équation de la droite d'ajustement : c'est-à-dire la pente et l'ordonnée à l'origine,
- enfin, le coefficient de corrélation linéaire.

Les moyennes arithmétiques

Pour calculer une moyenne on peut faire appel à la fonction mean() ainsi :

```
mean(xi)
## [1]
6.5

mean(yi)
## [1]
52.92
```

Ou faire le calcul « à la main » grâce à la fonction sum() qui calcule la somme des termes d'un objet et la fonction length() qui compte le nombre de termes d'un objet :

```
ximoy <-
sum(xi)/length(xi)
```

```
yimoy <-
sum(yi)/length(yi)
```

Variances

La variance d'une distribution correspond au calcul que l'on résume ainsi « *La moyenne des carrés moins le carré de la moyenne ».* Soit le calcul suivant :

```
xivar <- mean(xi^2) -
mean(xi)^2
```

```
yivar <- mean(yi^2) -
mean(yi)^2
```

On peut aussi utiliser les objets que nous avons créés auparavant ce qui donne pour le calcul de la variance :

```
mean(xi^2) -
ximoy^2
## [1] 11.92
```

```
mean(yi^2) -
yimoy^2
## [1] 92.74
```

La commande qui permet de faire le calcul automatique est var()

```
var(xi)
## [1]
13
```

```
var(yi)
## [1]
101.2
```

Vous remarquerez que le résultat n'est pas le même que le précédent, car *R* calcule la variance « sans biais ». La variance sans biais, comme l'écart type sans biais, correspond au même calcul, mais pour une observation de moins, c'est-à-dire multiplié par le nombre de termes sur le nombre de termes moins un. Autrement dit, pour 12 termes nous allons multiplier la formule de la variance par 12/11 :

```
(mean(xi^2) - mean(xi)^2) *
12/11
## [1] 13
```

```
var(xi)
## [1]
13
```

Cela n'a pas d'importance lorsque l'on a un grand nombre de termes, mais sur un petit échantillon comme dans notre exemple, la différence entre les deux résultats est notable, on préférera pour ce billet la variance ordinaire.

Écarts types

L'écart type est la racine carrée de la variance. La commande sqrt() pour *square root* calcul la racine carrée. L'écart type est donc :

```
sqrt(xivar)
## [1]
3.452
```

```
sqrt(yivar)
## [1]
9.63
```

La commande qui calcul l'écart type est sd() pour *standard deviation* mais, comme pour la variance, *R* calcul l'écart type sans-biais.

Calcul de la covariance

La covariance est un indicateur de la variation simultanée de deux variables. La formule correspond à « *la moyenne des produits moins le produit des moyennes* » ; la covariance de xi et yi se calcule donc ainsi :

```
xiyicovar <- mean(xi * yi) - mean(xi) *
mean(yi)
```

Dans *R* la commande cov() fait le calcul, mais là encore c'est un calcul sans biais.

Récapitulatif

Nous disposons maintenant d'un ensemble d'indicateurs :

- les moyennes des deux distributions qui sont dans les objets ximoy et yimoy,
- les variances dans les objets : xivar et yivar,
- les écarts types que l'on déduit des variances en faisant la racine carrée : sqrt(xivar) et sqrt(yivar),
- la covariance du nuage de point : xiyicovar. À partir de ces valeurs, nous allons (re)calculer l'équation de la droite d'ajustement et le coefficient de corrélation linéaire.

Calcul de l'équation de la droite d'ajustement

Pour mémoire une régression linéaire simple consiste à trouver l'équation d'une droite résumant au mieux un nuage de points. On peut écrire l'équation de cette droite ainsi : y = ax + b et nous chercherons à trouver les valeurs de a (la pente) et de b (l'ordonnée à l'origine).

Pente de la droite d'ajustement

La pente de la droite d'ajustement se calcule simplement en divisant la covariance du nuage de point par la variance des xi :

```
pente <-
xiyicovar/xivar
```

On a donc a = 2,744. La formule éclatée de la pente, c'est-à-dire sans utiliser d'objets intermédiaires, est la suivante :

```
sum((xi - mean(xi)) * (yi - mean(yi))/sum((xi -
mean(xi))^2))
## [1] 2.745
```

Ordonnée à l'origine

Maintenant que nous disposons de la valeur de la pente de notre droite, il est simple de calculer l'ordonnée à l'origine puisque si y = ax + b alors b = y - ax. On effectue le calcul ainsi avec les valeurs des moyennes :

```
ordorig <- mean(yi) - pente *
mean(xi)

print(ordorig)
## [1] 35.08
```

On peut vérifier facilement avec les commandes coef() et lm()

```
coef(lm(yi ~
xi))
## (Intercept)
xi
## 35.076 2.745
```

Calcul du coefficient de corrélation linéaire

La valeur du coefficient de corrélation linéaire correspond à la covariance que divise l'écart type des xi, multiplié par l'écart type des yi :

```
xiyicovar/(sqrt(xivar) *
sqrt(yivar))
## [1] 0.9839
```

On peut vérifier que l'on ne s'est pas trompé grâce à la commande cor() :

```
cor(xi, yi, method =
'pearson')
## [1] 0.9839
```

Perspectives

D'autres éléments d'évaluations du modèle seront présentés dans un prochain billet, nous verrons en détail comment interpréter les éléments donner par *R* avec la combinaison des commandes summary() et lm() :

Références

Monino, J.-L., Kosianski, J.-M., & Cornu, F. L. (2007). *Statistique descriptive*. Éditions Dunod.