Initiation à la régression linéaire simple avec R

mercredi 12 février 2014 par Mehdi Khaneboubi

Pour citer cet article :

Khaneboubi Mehdi (2014). Initiation à la régression linéaire simple avec R. *Adjectif.net* Mis en ligne mercredi 12 février 2014 [En ligne] http://www.adjectif.net/spip/spip.php?article275

Résumé :

Pour tester le lien entre deux variables qualitatives, il est fréquent de réaliser un test de khi d'indépendance sur un tableau de contingence. Comment créer un tableau de contingence a été présenté ici et on pourra consulter le processus du test là. Si le croisement de deux variables qualitatives produit un tableau croisé, comment analyser le lien entre deux variables quantitatives ? Autrement dit, que peut-on dire d'un nuage de points à deux dimensions ? Comment calculer l'équation d'une droite qui résumera ce nuage de points ? Ce premier billet cherche à présenter la première étape pour analyser le lien entre deux variables quantitatives par la méthode des moindres carrés avec le logiciel R.

Mots clés :

Logiciel R

R est un logiciel libre et gratuit et un langage de traitement statistique qui a notamment été présenté ici. Le langage R (R Development Core Team, 2013) est dit orienté objet comme Python ou Ruby. Un des avantages de R réside dans la possibilité de communiquer des scripts par écrit, car le plus souvent on l'utilise en mode console. Cela évite de se soucier des problèmes techniques produits par la variété de versions des systèmes d'exploitation et d'avoir recours à des copies d'écrans, mais permet surtout de communiquer des calculs et des analyses statistiques en quelques lignes de texte.

On trouvera un guide sur son installation ici. Le lecteur pourra aussi suivre différents MOOC très complets et notamment celui-ci. Un tutoriel d'initiation particulièrement ludique est disponible sur cette page.

Résumer un nuage de points

Le principe élémentaire des statistiques descriptives consiste à résumer des données que l'on ne peut pas appréhender une à une. On cherche à synthétiser ces données par le biais d'indicateurs. Ainsi un indicateur qui est très utilisé pour résumer une variable quantitative est la moyenne arithmétique. Pour résumer un nuage de points dont les coordonnées sont constituées par deux variables quantitatives on va chercher l'équation d'une droite. À titre d'exemple, considérons une première variable constituée des valeurs 1 2 3 4 5 6 7 8 9 10 11 et 12 et une seconde dont les valeurs sont : 40 42 44 45 48 50 52 55 58 63 68 et 70.

Saisie des valeurs dans R
Dans R, assignons la première distribution à un objet que l'on nomme xi :

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

Lorsque l'on tape le nom de l'objet puis la touche entrer, la console renvoie les valeurs :

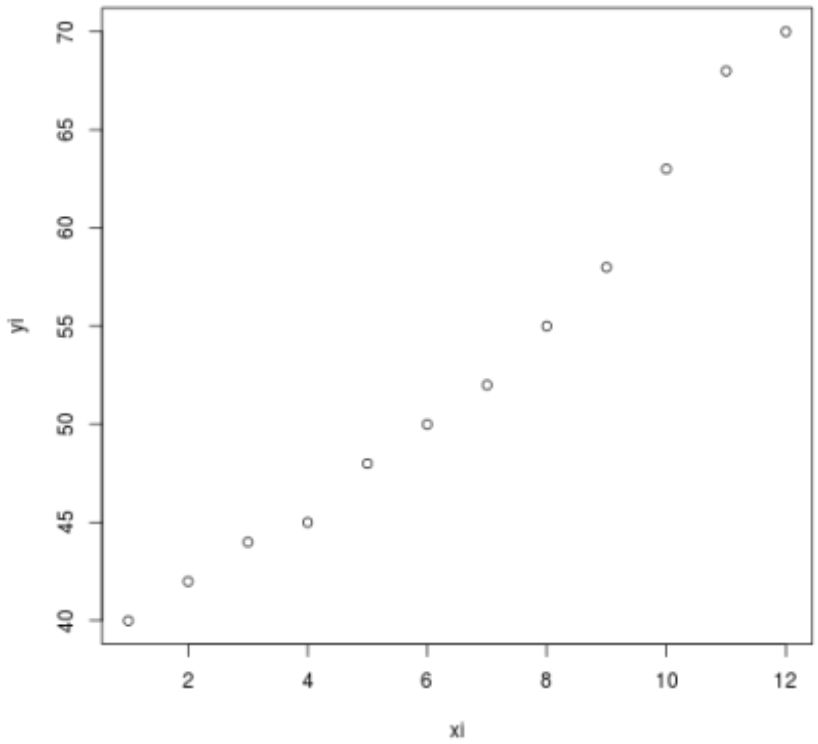
```
xi
## [1] 1 2 3 4 5 6 7 8 9 10 11
12
```

Entrons la seconde variable dans un objet que l'on nomme yi ainsi :

```
yi <- c(40, 42, 44, 45, 48, 50, 52, 55, 58, 63, 68,
70)
yi
## [1] 40 42 44 45 48 50 52 55 58 63 68 70
```

Représentation graphique
Il est maintenant possible de voir comment se répartissent ces deux distributions l'une par rapport à l'autre en faisant un graphique grâce à la fonction `plot()` :

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



On remarque une régularité dans le nuage, presque un alignement. Cet élément suggère qu'une droite serait un bon résumé de ces points. L'équation d'une droite est de la forme y = ax + b . La méthode de la régression linéaire par les moindres carrés va nous permettre de connaître les valeurs *a* et *b* de façon à minimiser l'écart entre la droite et l'ensemble des points. Pour plus de détail sur la méthode pas à pas avec un tableur le lecteur pourra consulter (Monino, Kosianski, & Cornu, 2007).

Équation de la droite

Valeurs calculées par la commande `lm()`

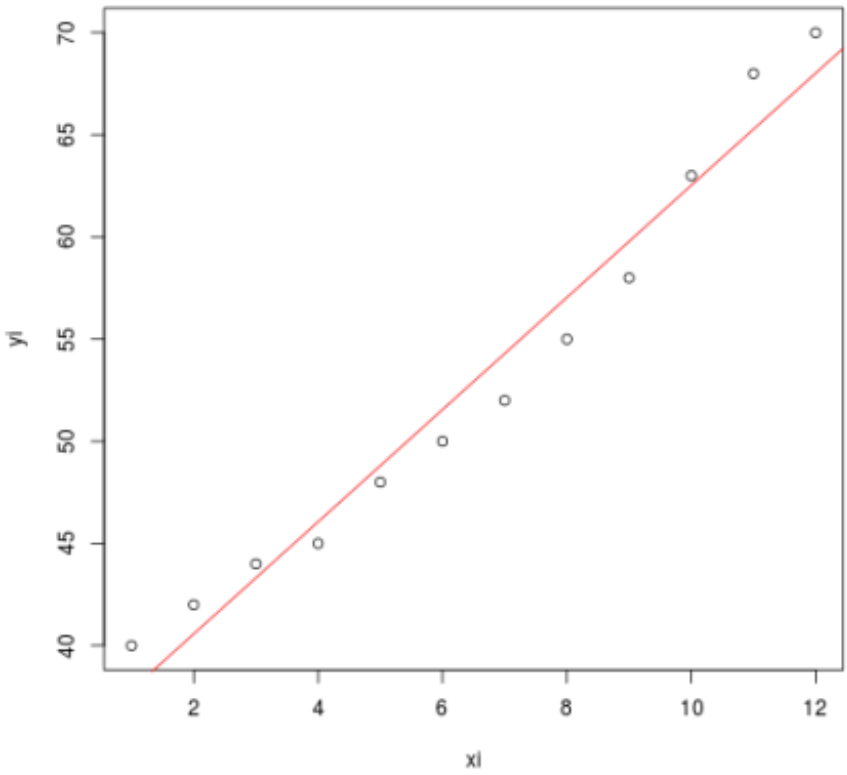
Dans *R* on fera appel à la commande `lm()`, qui signifie *linear model* ou modèle linéaire en français. Pour obtenir l'équation de la droite d'ajustement simplement entrer la commande :

```
lm(yi ~ xi)
##
## Call:
## lm(formula = yi ~
xi)
##
## Coefficients:
## (Intercept)      xi
## 35.08
 2.74
```

En notation anglo-saxonne, « Intercept » correspond ici à l'ordonnée à l'origine le « b » de notre droite et le « x » est la pente de la droite ce qui correspond au « b » dans notre notation. L'équation, de notre droite est donc y = 2,74x + 35,08

Représentation de la droite
Pour dessiner les points et la droite, on fait :

```
plot(xi, yi)
abline(35.076, 2.745, col =
'red')
```



Dans le cas de ce nuage de points, la relation entre les deux variables (le xi et le yi) est assez nette.

Interprétation

On peut interpréter l'équation de la droite y = 2,745x + 35,076 en disant :

« lorsque les xi augmentent d'une unité les yi augmentent de 2,745 unités. »

Prédictions

Maintenant que nous disposons de l'équation, on peut réaliser des prédictions. Ainsi pour une valeur de x qui serait de 13 on calcule une approximation de y en calculant 2,745 × 13 + 35,076, dans *R* on le fera ainsi :

```
2.745 * 13 +
35.076
## [1] 70.76
```

y(13) sera donc aux alentours de 70.8.

Bien sûr *R* pourra le faire pour nous pour une série de valeur le faire automatiquement avec la commande predict(). Mais nous verrons cela dans un autre billet.

Quelle est la fiabilité de ce modèle ?

Nous avons donc obtenu un modèle statistique élémentaire, mais nous ne nous sommes pas interrogé sur sa validité et sa robustesse. Pour ce faire la première étape est toujours de revenir aux données et de les examiner lorsque c'est possible ou d'examiner des indicateurs de tendances centrales et des critères de dispersions notamment en utilisant la commande summary() :

```
data.frame(xi,
yi)
## xi yi
## 1 40
## 2 42
## 3 44
## 4 45
## 5 48
## 6 50
## 7 52
## 8 55
## 9 58
## 10 63
## 11 68
## 12 70

summary(xi)
## Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
## 1.00   3.75     6.50   6.50   9.25   12.00

summary(yi)
## Min.  1st Qu.  Median  Mean  3rd Qu.  Max.
## 40.0   44.8    51.0    52.9   59.2    70.0
```

Cette relation forte entre les deux variables que l'on trouve dans notre exemple, n'est pas toujours présente surtout lorsqu'il y a un grand nombre de valeurs. C'est pour cette raison qu'il est important de regarder et d'interpréter d'autres éléments permettant de savoir si la régression linéaire est adaptée aux données et les résume convenablement.

Dans un prochain billet, nous allons voir comment calculer et interpréter le coefficient de corrélation linéaire qui indique dans quelle mesure l'équation d'une droite de régression s'ajuste à un nuage de points.

Le lecteur impatient est invité à consulter des explications détaillées en consultant notamment Cornillon & al. (2010), Rodriguez (2013) ou l'article Wikipedia traitant de la régression linéaire sujet .

Références