Année : 2008-2009 1er semestre
Niveau : MASTER IS 1ère année
Cours : Logiciel R
Enseignant : A. Illig

TP7 : La régression linéaire avec R

Table des matières

1 Commandes R **3**
- 1.1 Cas des données `cars` 3
- 1.2 Commande `lm` 3
- 1.3 Interprétation des résultats 4
- 1.4 Représentations graphiques 5

2 Exercice d'application **7**

1 Commandes R

La régression linéaire simple ou multiple est obtenue très facilement en R grâce à la commande `lm` (qui permet d'ailleurs d'implémenter d'autres modèles comme l'analyse de la variance). Nous détaillons ici dans le cadre de la régression simple certaines des possibilités offertes par la fonction `lm`.

1.1 Cas des données `cars`

Attachons les données `cars` dans R et affichons le nom des variables ainsi que leur dimension :

```
> attach(cars)
> names(cars)
> dim(cars)
```

Le graphique de la figure Fig 1. obtenu par la commande

```
> par(bg=''lightgrey'')
> plot(cars, pch=22)
```

suggère que la distance de freînage `dist` s'exprime linéairement en fonction de la vitesse `speed`.

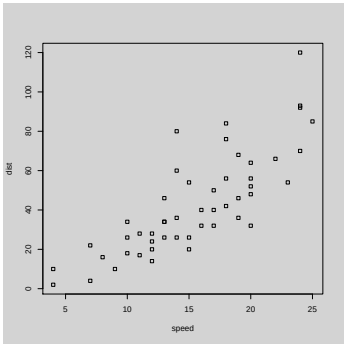

Fig. 1 – Nuage de points des données `cars`

1.2 Commande `lm`

Effectuons la régression linéaire de `dist` en fonction de `speed` et enregistrons les résultats dans une liste nommée `reg` :

```
reg=lm(dist~speed,data=cars)
```

La commande `lm` utilise une formule (`help(formula)` pour plus de détails) décrivant la régression à effectuer :

– `y ~ x` correspond à la régression $y = \beta_0 + \beta_1 x + \epsilon$,
– `y ~ x+0` ou `y ~ x-1` désigne la régression $y = \beta_1 x + \epsilon$.

Il est utile de préciser les données utilisées, ici `data=cars`, en option de la commande `lm` lorsque les noms des variables, ici `dist` et `speed`, sont communs à plusieurs ensembles de données.

La liste `reg` contient maintenant les informations suivantes :

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```

comme l'indique l'appel à la commande `names(reg)`. Par exemple, les coefficients de la régression linéaire, les valeurs ajustées et les résidus s'affichent de la manière suivante :

```
> reg$coefficients
> reg$fitted.values
> reg$residuals
```

1.3 Interprétation des résultats

La commande `summary(reg)` permet d'afficher à l'écran le résumé suivant :

```
Call:
lm(formula = dist ~ speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-Squared: 0.6511,     Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.490e-12
```

Précisons les résultats obtenus dans chaque rubrique :

– `Call` : Rappel de la formule de régression utilisée (identique à `reg$call`).
– `Residuals` : Répartition des résidus (identique à `summary(reg$residuals)`).
– `Coefficients` : La première ligne `(Intercept)` (respectivement la seconde ligne `speed`) rescense les résultats concernant le coefficient β_0 (respectivement β_1) :

  - `Estimate` : valeurs observées des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ des coefficients de la régression linéaire : *L'estimation du coefficient β_0 est -17.5791.*
  - `Std. Error` : valeurs observées des erreurs standardisées $\hat{\sigma}^2_{\beta_0}$ et $\hat{\sigma}^2_{\beta_1}$.
  - `t value` : valeurs observées des statistiques pivotales $\frac{\hat{\beta}_0}{\hat{\sigma}^2_{\beta_0}}$ et $\frac{\hat{\beta}_1}{\hat{\sigma}^2_{\beta_1}}$ permettant d'effectuer le test de significativité pour chacun des coefficients.
  - `Pr(>|t|)` : p-values des tests de significativité : *La p-value du test de significativité du coefficient β_0 est égale à 0.0123. Au niveau* 5%*, on rejette H_0 : $\beta_0 = 0$ car* $0.0123 < 0.05$.
- `Signif.codes` : Codes de significativité attribués aux coefficients en fonction de la valeur de la p-value du test de significativité. *La significativité du coefficient β_0 est symbolisée par $*$ car la p-value est de l'ordre de 0.01.*
- `Residual standard error` : Observation de l'estimateur sans biais $\hat{\sigma}^2$ de la variance du bruit ϵ.
- `Multiple R-Squared` : Pourcentage R^2 de la variance expliquée par le modèle.
- `Adjusted R-squared` : Ajustement de R^2 pour pénaliser les grandes dimension dans le cadre de la régression multiple.
- `F-statistic` : Valeur observée (et p-value) de la F statistique permettant de tester $H_0 : \beta_0 = \beta_1 = 0$ contre $H_1 : \beta_0 \neq 0$ ou $\beta_1 \neq 0$. On rejette H_0 lorsque F est trop grande. *La p-value est égale à* $1.490e^{-12}$. *Au niveau* 1%*, on rejette H_0 car* $1.490e^{-12} < 0.01$.

1.4 Représentations graphiques

Les informations de la liste `reg` permettent de tracer le nuage de points des données `cars` et la droite de régression :

```
> par(bg=''lightgrey'')
> plot(cars, pch=22)
> abline(reg$coefficients, col=3)
```

Par ailleurs, la commande `plot(reg)` permet d'afficher successivement 4 graphiques relatifs à la régression effectuée. Pour afficher les 4 graphiques simultanément (c.f. figure Fig 2), faire appel à la commande `layout` :

```
> layout(matrix(1:4,ncol=2))
> plot(reg)
```

Le graphique 3 de la figure Fig 2 permet de vérifier la normalité des résidus : *en dehors des points 23, 35 et 49, les points sont à peu près alignés sur la droite de Henry.*

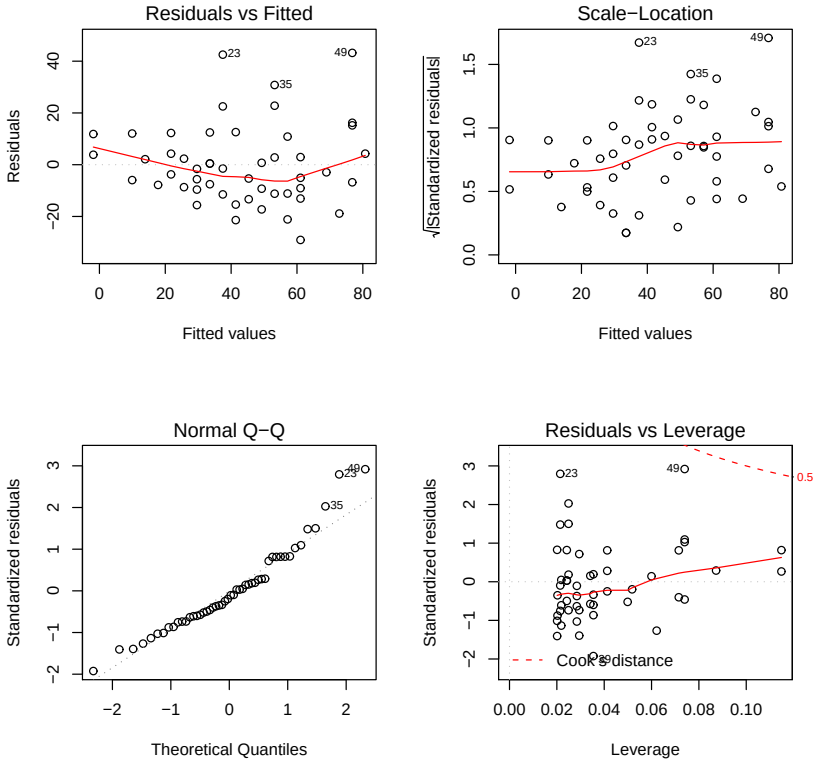

FIG. 2 – Graphiques de la régression linéaire

2 Exercice d'application

1. Attacher les données `attitude`. Afficher les noms des différentes variables, leur taille et rechercher leur signification en tapant `help(attitude)`.
2. Effectuer la régression multiple de l'indice attribué à un employé `rating` en fonction de tous les facteurs. Afficher le résumé de la régression et déterminer les coefficients significatifs.
3. Représenter sur un graphique l'indice attribué à un employé `rating` en fonction de la quantité de réclamations émises par cet employé `complaints`.
4. Effectuer la régression linéaire simple de `rating` en fonction de `complaints`. Analyser le résumé de la régression. Produire les graphiques de la régression. Afficher sur un même graphique le nuage de points et la droite de régression.