TP 5 – SY02
Régression linéaire

Les questions/sections marquées par un  sont des questions qui sont prévues pour être traitées en autonomie en dehors de la séance de TP.

En `R`, pour réaliser une régression linéaire, on appelle la fonction `lm` (*linear model*). Le premier argument de `lm` est un nouvel objet `R` qu'on appelle une formule et qui spécifie une « sortie » et des « entrées » séparées par le signe `~`. Les entrées et sortie sont des noms de colonnes d'un `data.frame` qu'il faut spécifier en deuxième argument. Par exemple, si on veut réaliser la régression des données `vary` en fonction des données `varx`, on écrira

```
donnees <- data.frame(varx = c(0, 0.2, 0.3, 0.6),
                      vary = c(1.01, 1.44, 1.55, 2.1))
lm(vary~varx, data = donnees)
```

On fera attention à l'ordre des éléments dans une formule. La variable à régresser se situe à gauche, le ou les régresseurs à droite.

(1) Quelles sont les estimations de l'ordonnée à l'origine (*intercept*) $\widehat{a}$ et de la pente $\widehat{b}$?

(2) À l'aide des fonctions `plot` et `abline`, tracer les points de coordonnées `x` et `y` ainsi que la droite des moindres carrés.

Pour avoir plus d'informations sur la régression effectuée, il faut stocker l'objet renvoyé par la fonction `lm` dans une variable et appeler la fonction `summary` avec cette variable en argument.

Toutes les données affichées par `summary` sont accessibles programmatiquement (voir la table 1 pour quelques exemples)

(3) À l'aide des correspondances indiquées dans la table 1, vérifier que la somme des résidus vaut 0 et que l'image de $\overline{x}$ par la droite des moindres carrés est $\overline{y}$.

1 Qualité de l'ajustement

1.1 Équation d'analyse de la variance

(4) À l'aide des correspondances indiquées dans la table 1, calculer/vérifier successivement

Notations	Code `R`
x_i	`x`
y_i	`y`
$\overline{x}$	`mean(x)`
$\overline{y}$	`mean(y)`
$\widehat{y}_i$	`m$fitted.values`
$y_i - \widehat{y}_i$	`m$residuals`
$\widehat{a}$	`m$coefficients[1]`
$\widehat{b}$	`m$coefficients[2]`

TABLE 1 – Correspondances notations/code `R`, où `m` est l'objet renvoyé par la fonction `lm`

1. la variance totale
$$S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2;$$
2. la variance expliquée par le régression
$$S_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2;$$
3. la variance résiduelle
$$S_{\text{res}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2;$$
4. la variance totale est égale à la somme de la variance expliquée par la régression et de la variance résiduelle,
5. R^2, le **coefficient de détermination** qui est égal à la proportion de la variance expliquée dans la variance totale soit :
$$R^2 = \frac{S_{\text{reg}}}{S_Y^2};$$
vérifier que R^2 est égal au carré du coefficient de corrélation de Pearson entre les observations y_i et les prédictions $\widehat{y}_i$.

1.2 Homoscédasticité, indépendance et normalité des résidus

Le coefficient de détermination est insuffisant pour rendre compte de la qualité de l'ajustement. À titre d'exemple, on utilise le jeu de données d'Anscombe qui consiste en 4 ensembles de 11 points du plan décrits à la figure 1. Pour rendre directement disponibles les colonnes en tapant leur nom, on pourra « attacher » ce jeu de données avec l'instruction

```
attach(anscombe)
```

Dès lors, au lieu de spécifier le jeu de données puis le nom de colonne

```
anscombe$x1
```

on peut se contenter de spécifier `x1`.

De même, pour éviter de définir systématiquement un `data.frame` avant de l'utiliser dans `lm`, on peut utiliser directement des vecteurs dans des formules. On peut alors simplement écrire

```
lm(y1 ~ x1)
```

Attention, cette écriture rend impossible la prédiction en de nouveaux points. On préférera donc utiliser la syntaxe détaillée en début de TP (syntaxe qui précéde la question 1) lorsqu'il sera nécessaire de faire de la prédiction.

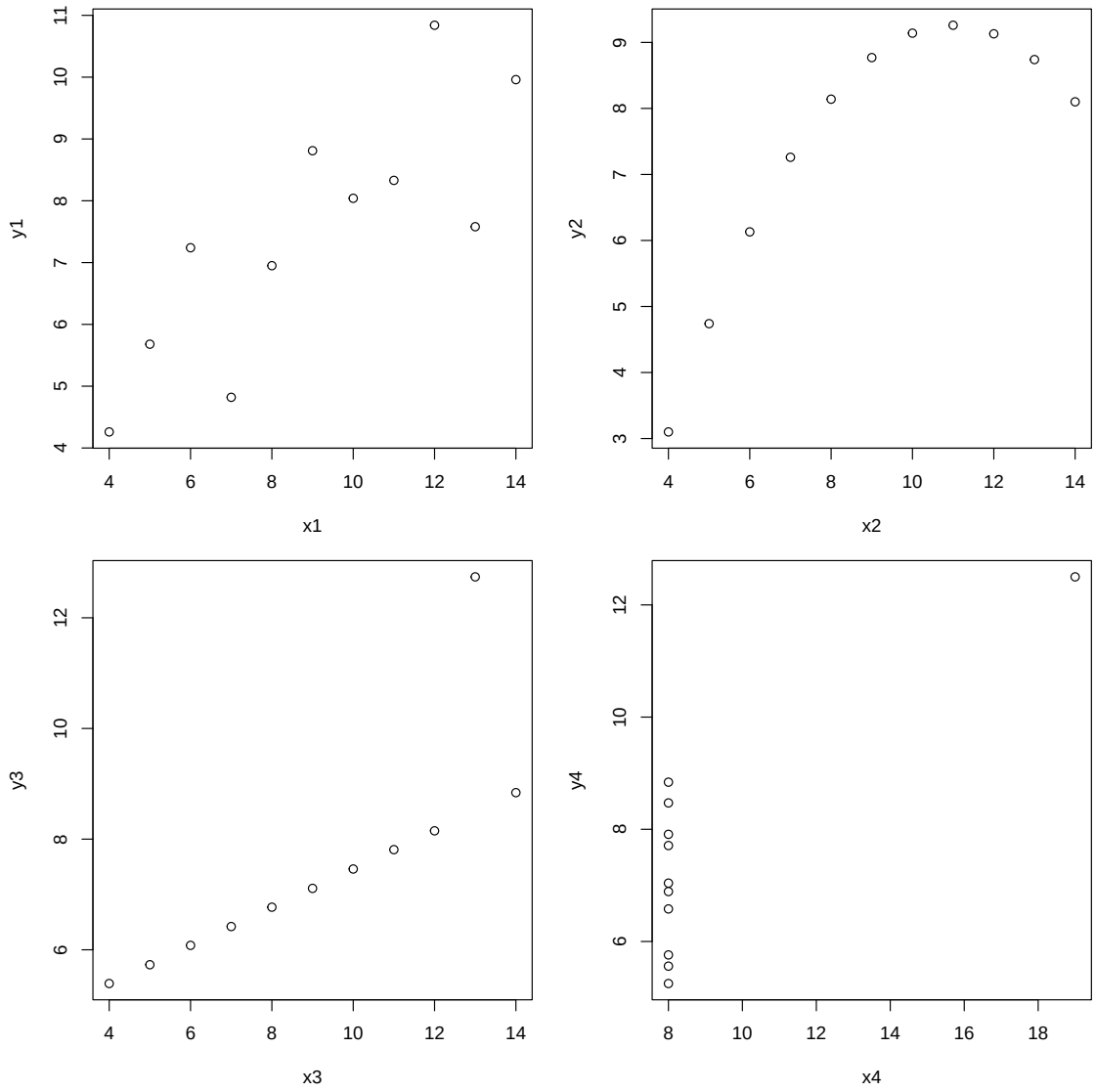

FIGURE 1 – Diagrammes de dispersion des 4 ensembles de 11 points du jeu de données d'Anscombe

(5) Effectuer les régressions linéaires sur les 4 ensembles de points. Que remarquez-vous ?

Le modèle de régression linéaire fait les hypothèses suivantes :

1. Lin : la relation entre y_i et x_i est linéaire
2. Norm : normalité des résidus (induite par la normalité des termes d'erreurs)
3. Ind : indépendance des résidus (induite par l'indépendance des termes d'erreurs)
4. Hom : homoscédasticité : $\forall i,\ \mathrm{Var}(\epsilon_i) = \sigma^2$

Pour estimer la qualité de l'ajustement, on utilise quelques diagnostics graphiques qui permettent de vérifier empiriquement la validité de ces hypothèses (cf section 7.5 du poly de cours).

(6) Faire une analyse des résidus et discuter de la validité des hypothèses Norm, Hom et éventuellement des hypothèses Lin et Ind pour l'ensemble des régressions linéaires proposées par le jeu de données Anscombe.

En particulier, on tracera

— Pour Norm, le diagramme quantile–quantile des résidus avec les fonctions `qqnorm` et `qqline` (on pourra également tracer l'histogramme des résidus corrigés et y superposer la densité d'une normale d'espérance leur moyenne empirique et d'écart-type, leur écart-type empirique) ;

— Pour Hom et éventuellement Ind et Lin), les résidus standardisés (`rstandard`) en fonction des prédictions (`fitted.values`) ou bien en fonction des valeurs de la variable explicative ;

2 Prédiction

Le fichier hooker-data.data contient un jeu de données recueillies par le botaniste anglais Joseph Dalton Hooker. Il s'agit de températures d'ébullition de l'eau relevées pour différentes altitudes. Dans cette section, il est indispensable d'effectuer la régression linéaire en utilisant la syntaxe détaillée en début de TP (celle située avant la question 1).

(7) Faire une étude de régression linéaire qui explique la pression atmosphérique.

(8) À l'aide de la function `confint`, donner un intervalle de confiance sur les coefficients de la droite des moindres carrés au niveau de confiance $1 - \alpha = 0.99$.

(9) À l'aide de la fonction `predict`, calculer un intervalle de confiance sur la pression pour une température d'ébullition mesurée de 97 °C.

Pour plus d'informations sur les arguments à fournir à la fonction `predict`, on pourra utiliser l'instruction suivante

```
?predict.lm
```

3 Étude de cas

Loi de Moore

La loi de Moore est une loi empirique qui dit que le nombre de transistors croit de manière exponentielle avec le temps. Autrement dit, on suppose que le nombre de transistors N_t au temps t est égal à

$$N_t = \alpha \exp(\beta t).$$

(10) À l'aide du fichier moore-data.data et en utilisant une régression linéaire, estimer les paramètres α et β et donner un intervalle de confiance et de prédiction sur N_{2018}. Retrouver le fait que le nombre de transistors double tous les 2 ans.

Hauteur et diamètre de cèdres

Le fichier cedar-data.data contient le diamètre et la hauteur de 139 cèdres. On cherche à prédire la hauteur d'un cèdre en fonction de leur diamètre.

(11) Faites l'étude de la régression linéaire expliquant la hauteur en fonction du diamètre. Analyser le diagramme des résidus. Que remarquez-vous ?

(12) On souhaite appliquer une transformation sur la variable explicative. On choisit la transformation de Box–Cox définie comme suit

$$f_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(x) & \text{si } \lambda = 0, \end{cases}$$

de sorte que le problème de régression est maintenant

$$\mathbb{E}[Y \mid X] = \alpha + \beta f_\lambda(X).$$

Créer une fonction qui réalise la transformation de Box–Cox.

(13) Parmi les valeurs -1, $-1/2$, 0, $1/3$, $1/2$, 1, trouver la valeur de λ qui semble le mieux expliquer la hauteur des cèdres.

(14) Refaites la même étude en appliquant une transformation logarithmique à la variable explicative.