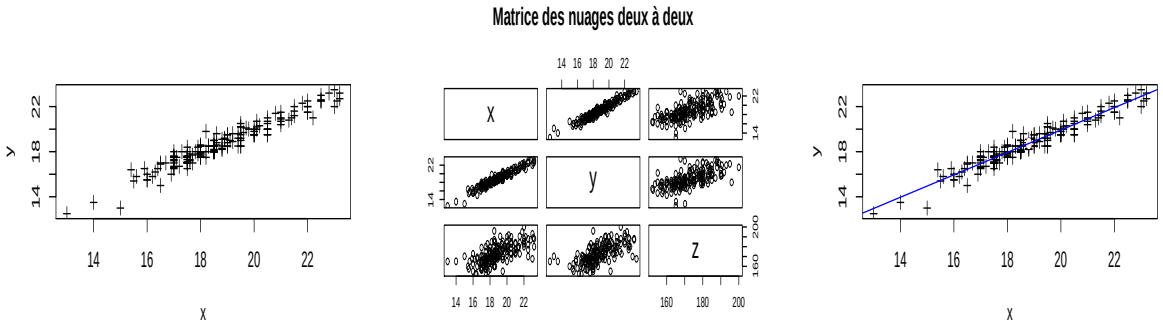
Cours 06
Régression linéaire

1 Nuage et droite

Dans ce chapitre nous abordons la question de la comparaison de deux (ou plusieurs caractères) `x` et `y`, pour un même individu en considérant les propriétés du *nuage des points* $M_i = (x_i, y_i)$. La commande `plot(x,y,pch=3)` permet de représenter ce nuage ; si on a plus de deux caractères pour un même individu on peut représenter les nuages de caractères deux-à-deux au moyen de la commande

`pairs(~x+y+z,data=MonTableau, main="Matrice des nuages deux à deux")`

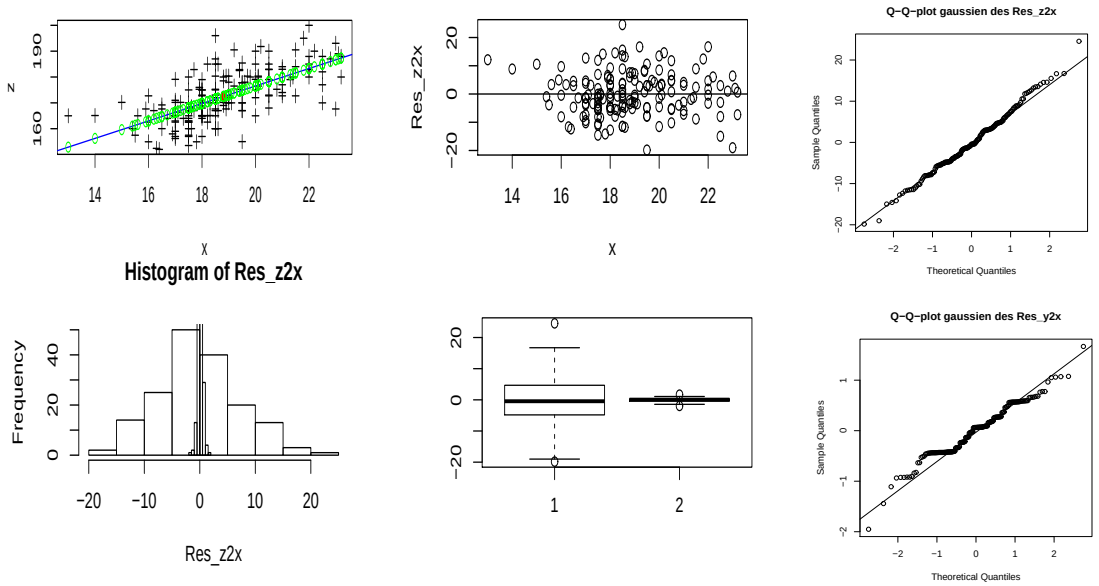
en définissant au préalable `MonTableau` par `MonTableau=data.frame(x,y,z)`.



Le *centre de gravité* du nuage est la moyenne de ses points. Pour le nuages des `x` et `y` c'est donc le point $(\bar{x}, \bar{y})$ =`(mean(x),mean(y))`. Nous abordons ici la question de choisir un *modèle linéaire* dans le cas où ce nuage semble s'étirer le long d'une droite, comme c'est le cas dans les exemples ci-dessus.

2 Résidus

Choisir un modèle linéaire de `y` sur `x` revient à choisir a et b tels que pour tout (x_i, y_i), $y_i = ax_i + b + \epsilon_i$. On dit qu'on régresse les y_i sur les x_i, et les ϵ_i sont les *résidus* pour cette régression $\hat{y} = ax + b$, ou encore `Res`$= \epsilon$, et `Res=y-(a*x+b)`. Le choix des a at b est tel que `Res` est nécessairement de moyenne `mean(Res)` égale à 0. On a choisi de représenter les résidus de la régression des `z` sur les `x` car ils sont plus grands que ceux de la régression des `y` sur les `x` : cette différence se voit aussi sur leurs histogramme et leurs boite à moustaches.



Sous `R` la commande `lm` permet de produire cette régression (choix de a et b), comme dans l'exemple suivant : `DteReg_y_sur_x=lm(y~x,data=MonTableau)`, qui permets ensuite d'obtenir la représentation de la droite de régression au moyen d'une commande telle que `abline(coef(DteReg_y_sur_x),col="blue")`.

Les *valeurs prédites* par le modèle linéaire : `y2x[i]`$= \hat{y}_i = ax_i + b$. Elles sont données par la commande `y2x=fitted(DteReg_y_sur_x)` dans le premier exemple du cours.
Les *résidus* : `Res[i]`$= \epsilon_i = y_i - \hat{y}_i$
Les a et b sont choisis de manière à minimiser la somme des carrés des résidus $\sum_{i=1}^n \epsilon_i^2$. Soit $\varphi(a,b)$ cette somme :

$$\varphi(a,b) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - ax_i - b)^2 = \sum_{i=1}^n y_i^2 + a^2 \sum_{i=1}^n x_i^2 + nb^2 - 2a \sum_{i=1}^n x_i y_i - 2b \sum_{i=1}^n y_i + 2ab \sum_{i=1}^n x_i.$$

Pour que cette fonction soit minimale, il faut qu'elle le soit tant du point de vue de a que de b, et donc les deux dérivées (partielles) doivent être nulles $\frac{\partial \varphi}{\partial a} = 0 = \frac{\partial \varphi}{\partial b}$; ce qui nous donne les deux équations $0 = \frac{\partial \varphi}{\partial a} = 2a\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i y_i + 2b\sum_{i=1}^n x_i$ et $0 = \frac{\partial \varphi}{\partial b} = 2nb - 2\sum_{i=1}^n y_i + a\sum_{i=1}^n x_i$. Notons $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ les moyennes des x et des y. La deuxième équation donne alors $\boxed{b = \bar{y} - a\bar{x}}$, et la première équation $0 = a\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i + (\bar{y} - a\bar{x})n\bar{x} = a(\sum_{i=1}^n x_i^2 - n\bar{x}^2) - (\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})$, d'où $\boxed{a = \dfrac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}}$.

Au prochain cours nous définirons (enfin) les notions de variance $\mathrm{Var}\,(x)$ d'un échantillon `x` et covariance $\mathrm{cov}\,(x,y)$ de deux échantillons `x` et `y` et cette dernière formule pourra encore s'écrire $a = \mathrm{cov}\,(x,y)/\mathrm{Var}\,(x)$.

Normalité des résidus : Il y a de bonne raison pour espérer que les résidus d'une régression forment un échantillon gaussien (on dit aussi qu'ils résultent d'un "bruit" tel que des erreurs d'arrondis). Un peut examiner cette hypothèse en comparant les quantiles des résidus avec ceux d'une loi normale au moyen de `qqnorm`. C'est ce qui a été reproduit sur la figure, où on a noté `Res_y2x` et `Res_z2x` les résidus des regressions de `y` et de `z` sur `x`. Pour la régression de `y` sur `x` nous retrouvons l'artefact déjà observé que les décimales 0 et 5 dans la mesure en centimètres ont été favorisées, peut-être parce que certaines personnes chargées de procéder aux mesures des empans ont renoncé à la précision du millimètre et se sont limitées à la précision du demi-centimètre.

3 Résumé

Le code utilisé comporte à la fois des commandes `R` déjà connues et des commandes nouvelles, qui elles-même se divisent en commandes de type gestion graphique, et des commandes spécifiques à la notion de statistique qui est l'objet de ce cours : la régression linéaire ; il y a aussi quelques commandes de manipulation de données qui seront approfondies plus tard : nous pensons que vous pouvez commencer à les utiliser en mimant les exemples. Il y a enfin des noms de variables, choisis de façon à être facilement mémorisés, mais dont le choix est totalement arbitraire. C'est un bon exercice de distinguer ces trois derniers types, en soulignant/stabilotant de façon distincte ces trois derniers types (réservez le stabilo à la régression linéaire).

Regression linéaire

`lm` Exemple : `DteReg_y_sur_x=lm(y~ x,data=MonTableau)`. Effectue la régression des `y` sur les `x`.
`coef` Exemple : `coef(DteReg_y_sur_x)`. Retournes les coefficients b =`Intercept` et a.
`abline` Exemple : `abline(coef(DteReg_y_sur_x),col="blue")`. Ajoute la droite de régression.
`fitted` Exemple : `y2x=fitted(DteReg_y_sur_x)`. Retourne le vecteurs des valeurs $\hat{y}_i$ que prédit le modèle pour chaque x_i =`x[i]` de l'échantillon `x` ; les différences avec les valeurs mesurées `y` s'appellent les résidus du modèle.

Commandes graphiques

`plot` Exemple : `plot(x,z,type="p",pch=8)` : dessine le nuage de points.
`points` Exemple : `points(mean(x),mean(y),col="white",pch=19)` : ajoute un ou des points, ici le centre de gravité du nuage.
`windows` Exemple : `windows()` : ouvre et active une nouvelle fenêtre.
`dev.cur` Exemple : `dev.cur()` : récupère le numéro de la fenêtre active.
`dev.set` Exemple : `dev.set(2)` : active la fenêtre 2.

Commandes qui seront revues plus tard

`data.frame` Exemple : *MonTableau*=`data.frame(x,y,z)`
`pairs` Exemple : `pairs(~x+y+z,data=`*MonTableau*`, main="`*Matrice des nuages deux à deux*`")`. Donne le scatter-plot des échantillons *x*, *y* et *z*

Les noms de variables utilisées

Ce sont *x*, *y*, *z* qui sont les noms des échantillons utilisés, *MonTableau* qui est le tableau (data.frame) constitué de ces trois échantillons, *DteReg_y_sur_x* qui est le résultat de la régression linéaire des `y` sur les `x`, *DteReg_z_sur_x*, idem pour les `z`, *y2x* et `z2x` sont, pour l'échantillon `x`, les valeurs qui seraient prédites par le modèle linéaire $y = a_1 x + b_1$ et $z = a_2 x + b_2$. Enfin *Res_z2x*, et *Res_y2x* sont les résidus des deux régressions.