TP Régression linéaire simple

La fonction `lm`

La fonction R utilisée pour le modèle linéaire gaussien est la fonction `lm()`. Elle prend comme argument, le modèle que l'on souhaite étudier.
Dans le cas de la régression linéaire simple, le modèle est le suivant : $Y_i = a + bx_i + \epsilon_i, \quad i = 1, \dots, n$, ce qui en langage R s'écrit : $y \sim x$.

Sortie : res=lm($y \sim x$, data) (data étant le jeu de données dont proviennent x et y) donne une liste. Pour connaître les élements de la liste, on tape names(res).

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

La commande `summary(res)` donne une liste de résultats :

— `summary(res)$coefficients`

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	$\hat{a}$	$\hat{\sigma}_{\hat{a}}$	statistique $\frac{\hat{a}-0}{\hat{\sigma}_{\hat{a}}}$	p value du test de $H_0 : a = 0$
x	$\hat{b}$	$\hat{\sigma}_{\hat{b}}$	statistique $\frac{\hat{b}-0}{\hat{\sigma}_{\hat{b}}}$	p value du test de $H_0 : b = 0$

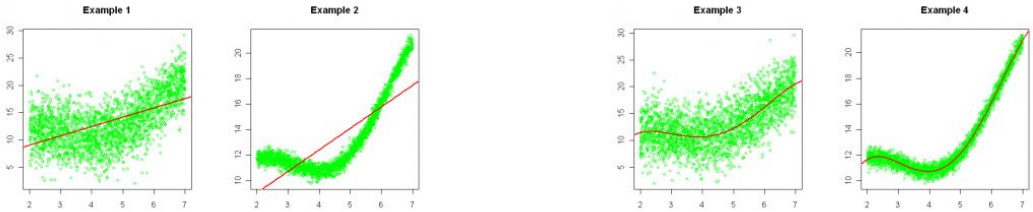
On obtient donc $\hat{\theta} = (\hat{a}, \hat{b})'$ avec `summary(res)$coefficients[,1]`.
Ce tableau permet d'avoir les résultats des tests de Student de $H_0 : a = 0$ contre $H_1 : a \neq 0$, et de $H_0 : b = 0$ contre $H_1 : b \neq 0$.

— l'estimation de la racine carrée de la variance résiduelle appelée "Residual standard error". On l'obtient aussi en tapant dans R `summary(res)$sigma`

— la statistique "F-statistic" du test de Fisher (test de H_0 : (modèle M_0) $Y_i = a + \epsilon_i$ contre H_1 : (modèle M_1) $Y_i = a + bx_i + \epsilon_i$), les ddl("DF") et la p-value associés. On l'obtient aussi en tapant dans R `summary(res)$fstatistic`

— le coefficient de détermination R^2 appelé "Multiple R-squared". Il représente la part de variabilité des données qui est expliquée par le modèle (cad le modèle permet d'expliquer $100 \times R^2\%$ de la variance de Y). On a la décomposition suivante : $SCT = SCR + SCM$ avec $SCT = \text{var}(Y)$ la somme des carrées totale, SCM la somme des carrées expliquée par le modèle, et SCR la somme des carrées résiduelle. Par définition, $R^2 = 1 - \frac{SCR}{SCT}$.

rq : Attention à l'interprétation du R^2. C'est un critère facile à utiliser (on espère R^2 proche de 1 pour dire que le modèle explique bien les données), mais ne pas se fier uniquement à cette valeur ! Voici pour vous en convaincre un exemple.



Exemples 1 et 2 : on obtient en ajustant un modèle de régression linéaire respectivement $R^2 = 0.3372$ et $R^2 = 0.6414$
Exemples 3 et 4 : en ajustant cette fois un modèle polynôme de degré 4, on obtient respectivement $R^2 = 0.4968$ et $R^2 = 0.987$.
Sur l'exemple 3, l'ajustement est "bon" au sens où il serait difficile de faire mieux, c'est simplement qu'il est très bruité. Avec un R^2 de 0.5, l'exemple 3 parait donc proposer un "meilleur" ajustement que l'exemple 2, alors que le R^2 est de 0.65. Dans le premier cas, l'ajustement semble bon, mais très bruité, alors que pour le second, le modèle est plutôt mauvais, mais le bruit est relativement faible.

— R^2 augmente en ajoutant des variables explicatives, ce qui le rend difficilement utilisable pour comparer des modèles. C'est pour cela que l'on donne le R^2 ajusté appelé "Adjusted R-squared".

La commande `anova(res)` permet en plus la sortie de la table d'analyse de variance, càd les sommes des carrées ayant servis à calculer la statistique de Fisher :

Source de variation	ddl	Somme carrés	Moyenne des carrés	statistique F	$Pr(>F)$
Modele	1	SCM= $\|\|\hat{Y}_{M_1} - \hat{Y}_{M_0}\|\|^2$	SCM/1	(SCM/1)/SCR/(n-2)	p value
Residuelle	n-2	SCR= $\|\|Y - \hat{Y}_{M_1}\|\|^2$	SCR/(n-2)		

Exercice 1

On reprend les données *ozone* (TP statistique descriptive). La variable à expliquer y est la concentration en ozone, notée "max03", et la variable explicative x est la température à midi, notée "T12".

1. Représenter ces données par un nuage de points. Un modèle de régression simple vous semble t-il adapté aux données ?
2. Donner l'équation de la droite des moindres carrés. La rajouter sur le graphique.
3. Donner un intervalle de confiance pour a et un intervalle de confiance pour b de niveau 95%.
4. Tester si la température à midi influe de manière significative sur la concentration en ozone. Vous avez 3 méthodes possibles :
   (a) avec un test de Student (soit en calculant le quantile, soit avec la p-value).
   (b) en utilisant l'intervalle de confiance.
   (c) avec un test de Fisher (soit en calculant le quantile, soit avec la p-value).
   Vérifier que vous obtenez à chaque fois le même résultat.
5. Test unilatéral : la température à midi influe positivement ou négativement sur la concentration en ozone ? Pour y répondre, chosir ce qu'il faut mettre dans les hypothèses H_0 et H_1 et utiliser un test de Student.
6. Tester si la droite de régression passe par l'origine. Vous avez deux méthodes possibles : avec un test de Student ou en utilisant l'intervalle de confiance.
7. Un des buts de la régression est de faire de la prévision. Soit donc une nouvelle valeur de température (par exemple 19 degrés), pour laquelle nous voulons prédire la concentration en ozone.
   (a) D'après le modèle, que vaut la concentration en ozone y_{19} pour une température de 19 degrés ? Que vaut $\mathbb{E}(y_{19})$?
   (b) Donner la concentration $\widehat{y}_{19}$ prédite par le modèle pour une température de 19 degrés.
   (c) On peut obtenir des intervalle de confiance pour $\mathbb{E}(y_{19})$ et pour y_{19} à partir de l'estimation $\widehat{y}_{19}$. A votre avis quel intervalle de confiance va être le plus large ?
   (d) Un intervalle de confiance de niveau $1-\alpha$ pour $\mathbb{E}(y_{19}) = a + 19b$ est donné par :

$$\left[\widehat{y}_{19} - t_{1-\alpha/2,n-2}\widehat{\sigma}\sqrt{(\frac{1}{n} + \frac{(19-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2})}, \quad \widehat{y}_{19} + t_{1-\alpha/2,n-2}\widehat{\sigma}\sqrt{(\frac{1}{n} + \frac{(19-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2})}\right]$$

   Un intervalle de confiance de niveau $1-\alpha$ pour y_{19} est donné par :

$$\left[\widehat{y}_{19} - t_{1-\alpha/2,n-2}\widehat{\sigma}\sqrt{(1+\frac{1}{n} + \frac{(19-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2})}, \quad \widehat{y}_{19} + t_{1-\alpha/2,n-2}\widehat{\sigma}\sqrt{(1+\frac{1}{n} + \frac{(19-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2})}\right]$$

   Cet intervalle est appelé intervalle de prédiction.
   (e) Sous R la fonction predict permet de calculer ces intervalles de confiance.

```
predict19 <- predict(res, newdata=data.frame(T12=19), interval="confidence")
predict19 <- predict(res, newdata=data.frame(T12=19), interval="prediction")
```

   (f) En faisant varier $T12$ dans la fonction *predict*, donner un intervalle de confiance de la droite de régression et un intervalle de prédiction de la droite de regression. Sur le graphe, tracer la droite de regression, et les 2 bornes de l'intervalle de confiance, et les 2 bornes de l'intervalle de prédiction (ce sont les 2 branches d'une hyperbole).
   (g) Que pouvez-vous dire de la taille de l'intervalle de prédiction lorsque la nouvelle valeur de $T12$ est loin de la moyenne $\bar{x}$?