GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Cours 10

- Données multivariées
- Régression multiple
- Logiciel R / interprétion des sorties R
- Géometrie de régression
- Introduction : 1-way ANOVA (anova à une voie)

Données multivariées

Individus	X_1	X_2	$\ldots$	X_j	$\ldots$	X_p
i_1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
i_2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
$\ldots$						
i_i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\ldots$						
i_n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{np}

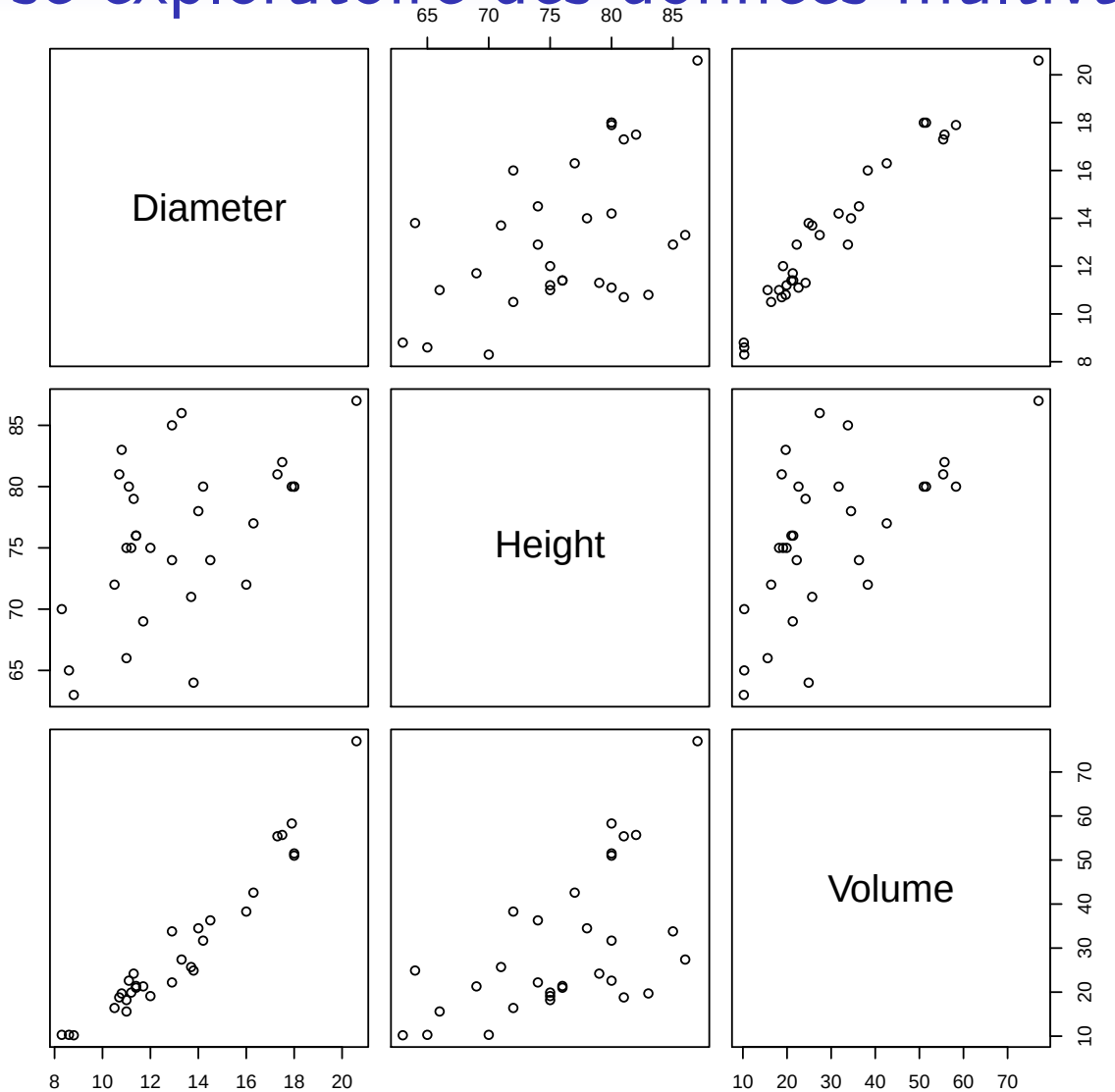
vecteur des moyennes : $(\overline{x}_1, \ldots, \overline{x}_p)$
matrice des variances-covariances (ou *matrice de dispersion*) :

$$\begin{pmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_2^2 & \cdots & s_{2,p} \\ \cdots & s_i^2 & s_{i,j} & \cdots \\ s_{p,1} & s_{p,2} & \cdots & s_p^2 \end{pmatrix}$$

Exemple

- Un échantillon de cerisiers a été coupé et les mesures prises pour
  - `Diameter` (inches)
  - `Height` (feet)
  - `Volume` (cubic feet)
- Le but de la collecte de ces données était de fournir un moyen de prédire le volume de bois dans les arbres, sachant la hauteur et le diamètre
- Utilise un modèle de régression

Analyse exploratoire des données multivariées



Régression multiple

- On peut avoir *plusieurs* variables explicatives x :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- Même suppositions dans le cas régression simple :
  $\epsilon \sim$ iid $N(0, \sigma^2)$
- Résumé suppositions :
  - **L**inear model (in the parameters)
    [modèle **L**inéaire (dans les paramètres)]
  - **I**ndependent errors / observations
    [**I**ndépendance des erreurs / observations]
  - **N**ormal errors / observations
    [erreurs / observations **N**ormales]
  - **E**qual error variances
    [**E**galité des variances des erreurs]

Algèbre matricielle pout la régression (simple)

- Le modèle :

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Moindres carrés (ordinaires) pour la régression multiple

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Trouver une solution $\hat{\boldsymbol{\beta}}$ qui minimise la somme des carrés des résidus (solution de *moindres carrés ordinaires (MCO)*) :

$$\min \sum_{i=1}^{n} e_i^2 \implies \frac{\partial \left(\sum_{i=1}^{n} e_i^2 \right)}{\partial \hat{\beta}_j} = 0, \quad j = 0, \ldots, p$$

$$\implies \sum_{i=1}^{n} x_{ij} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \quad j = 0, \ldots, p$$

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0 \implies \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

$$\implies \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

où $\boldsymbol{X}$ est la *matrice d'expérience* (*design matrix*) et $\boldsymbol{X}'$ est la transposée de $\boldsymbol{X}$

Logiciel : *R*

Pourquoi *R* ?

- Puissant, flexible, extensible langue et environnement pour le calcul statistique
- Large gamme de fonctions statistiques intégrées et 'packages' disponibles
- De haute qualité, des capacités graphiques excellentes
- Disponible pour les systèmes Unix / Linux, Windows, Mac
- Tout cela et ... *R* est gratuit !
- `http://cran.r-project.org/`

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

équation → y, x_1, x_2

$\hat{\beta}_0$ → (Intercept) −57.9877

$\hat{\beta}_1$ → Diameter 4.7082

$\hat{\beta}_2$ → Height 0.3393

Volume = -57.99 + 4.71 x Diameter + 0.34 x Height

*** Interprétation des coefficients ***

- Les coefficients de régression correspondent aux changements anticipés dans la réponse lorsqu'un changement d'une unité survient dans une variable explicative/prédictrice
- Pour la régression simple :
  - la pente est le changement espéré de la variable réponse si la variable explicative (x) est augmentée de 1 unité
  - l'ordonnée à l'origine est la valeur prédite de la réponse (y) lorsque $x = 0$
- Une distinction très importante – lorsque l'équation comporte *plusieurs* variables prédictrices :
  - chaque coefficient $\beta_1, \ldots, \beta_p$ correspond à la contribution d'une variable lorsque **toutes les autres variables présentes dans l'équation sont contrôlées ou tenues constantes**
  - le coefficient β_0 est la valeur prédite de la réponse (y) lorsque **toutes les variables** $x_1, \ldots, x_p = 0$

Propriétés de l'estimateur MCO

Dans le cas

1. $E(\epsilon_i) = 0,\ i = 1, \ldots, n$;
2. $Var(\epsilon_i) = \sigma^2$ (constante) ;
3. $Cov(\epsilon_i, \epsilon_j) = Cor(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

on a :

- Espérance : $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- Variance : $Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$
  ($(\boldsymbol{X}'\boldsymbol{X})$ symétrique)
- 0ptimalité :
  - Le théorème **Gauss-Markov** nous dit que *parmi toute estimation linéaire non biaisée*, l'estimateur MCO possède la *variance minimale*
  - On peut le résumé en disant : **l'estimateur MCO est le « BLUE »** (Best Linear Unbiased Estimator)

Test/intervalle de confiance pour les coefficients

- En supposant en plus $\epsilon_1, \ldots, \epsilon_n \sim$ iid $N(0, \sigma^2)$, on a

$$\hat{\boldsymbol{\beta}} \sim MVN\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

- Donc, $Var(\hat{\beta}_i) = \sigma^2 \left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}$
- L'IC avec indice de confiance $1 - \alpha$ pour β_i prend la forme

$$\hat{\beta}_i \pm \hat{\sigma} \sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}}\; t_{n-p-1,1-\alpha/2}$$

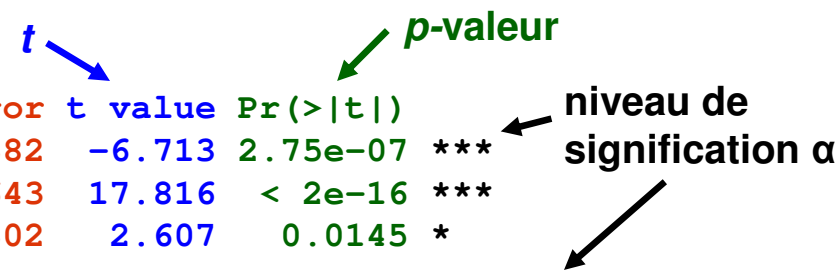
- Pour tester $H : \beta_i = 0$ contre $A : \beta_i \neq 0$

$$t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma} \sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}}}$$

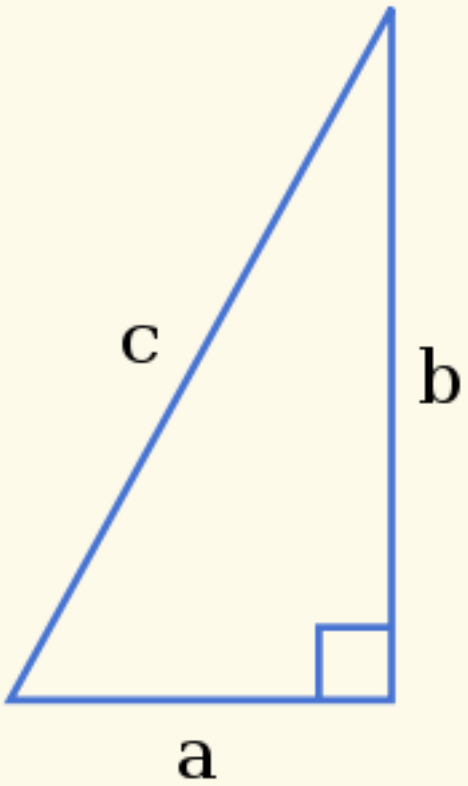
- On REJETTE H si : $|t_{obs}| > t_{n-p-1,1-\alpha/2}$
  (également si l'IC ne contient pas la valeur 0)

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```



Théorème de Pythagore



$$a^2 + b^2 = c^2$$

Géometrie de moindres carrés

- On considère $\boldsymbol{y}$ comme un vecteur dans l'espace n-dim
- Les vecteurs des colonnes de $\boldsymbol{X}$ forment un sous-espace (de l'estimation ou du modèle) p-dim
  - Variation des valeurs estimées des coefficients de régression localise des points différents du sous-espace
- Les valeurs prédites $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$ représentent le point du sous-espace le plus proche des observations : MCO est la *projection orthogonale* de $\boldsymbol{y}$ sur le sous-espace de $\boldsymbol{X}$
- Le résidu $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ est *orthogonal* aux vecteurs du sous-espace
- $SCE = \sum e_i^2 = \boldsymbol{e}'\boldsymbol{e}$ est le carré de la distance du vecteur des obs. au point le plus proche dans le sous-espace
- Partition de $\boldsymbol{y}$ en *deux composantes orthogonales* :
  - $\hat{\boldsymbol{y}}$ (sous-space du modèle, p dims)
  - $\hat{\boldsymbol{y}} - \boldsymbol{y}$ (sous-espace de l'erreur, $n - p$ dims)
- (degrés de liberté correspondent aux dims des sous-espaces)

Géométrie de MC

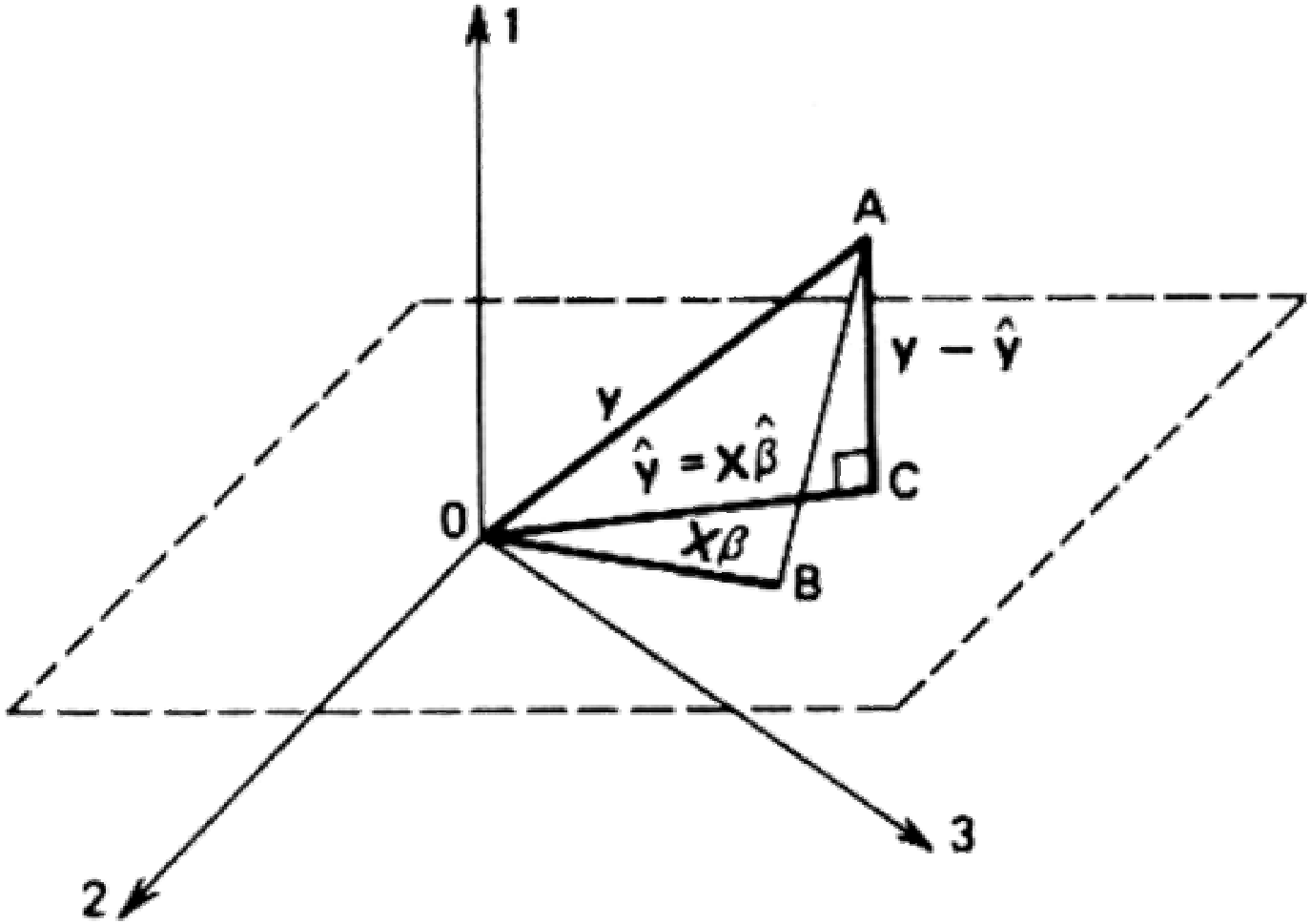


Figure 4.2 A geometrical interpretation of least squares.

Tableau de l'analyse de variance (ANOVA)

- Il s'agit d'une *partition de la somme des carrés totaux (SCT)*
- Théorème de Pythagore :

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- également :

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Cette égalité présentée dans un tableau :

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)	F	p-valeur
régression	p	$SCM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	SCM/p	CMM/CME	$P(F_{obs} > F_{p,n-p-1})$
erreur	$n-p-1$	$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$SCE/(n-p-1) (= \hat{\sigma}^2)$		
total (corr.)	$n-1$	$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$			

PAUSE

F-test - régression

- La statistique $F_{obs} = CM(\text{source})/CME$ teste l'hypothèse $H_0 : \beta_1 = \ldots = \beta_p = 0$ vs. A : au moins 1 $\beta_i \neq 0$
- La distribution de F_{obs} si H est vraie est *la distribution* $F_{p,n-p-1}$ de Fisher
- Au numérateur de la statistique F_{obs} se trouve *la variance expliquée par le modèle de régression*
- Au dénominateur se trouve *la variance résiduelle*
- On REJETTE l'hypothèse nulle H pour *grandes valeurs de F*
- Lorsqu'on teste une seule pente ($H : \beta_i = 0$), $F_{1,n} = t_n^2$

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

$F_{p,n-p-1}$

p-valeur

Coefficient de détermination

- La valeur y_i d'une observation peut être décomposée en deux parties : une partie *expliquée par le modèle* et une partie *résiduelle*
- La dispersion de l'ensemble des observations se décompose donc en :
  1. variance expliquée par la régression, et
  2. variance résiduelle, inexpliquée
- Le *coefficient de détermination* (ou *corrélation multiple*) R^2 se définit alors comme la part de variance expliquée par rapport à la variance totale
- Également, $R^2 = 1 - SCE/SCT$
- Dans le cadre d'une régression linéaire *simple*, c'est *le carré du coefficient de corrélation*

Coefficient de détermination ajusté

- Le *coefficient de détermination ajusté* R^2_{aj} tient compte du *nombre de variables*
- En effet, le défaut principal du R^2 est de *croître avec le nombre de variables explicatives*
- Un excès de variables produit des modèles *peu robustes*
- Donc on s'intéresse davantage à cet indicateur (R^2_{aj}) qu'au R^2
- Ce n'est pas vraiment un 'carré' – il peut même être négatif

$$R^2_{aj} = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

R^2

R^2-ajusté

R^2 ou R^2-ajusté ?

UTILISEZ LE R^2 AJUSTÉ !

MARRE DU R^2 ? Comme monsieur Statos, optez pour une qualité de régression plus sûre !!!

« Avant, j'utilisais un R^2 normal, j'étais fatigué et ça se voyait sur mon visage ; depuis que j'ai découvert le R^2 ajusté, ma vie a complètement changé ! »

Dépêchez-vous !

SATISFAIT ou REMBOURSÉ (*)

VU SUR INTERNET !!!

Dernière minute :

Pour vous souhaiter la bienvenue, la somme des carrés des résidus vous est offerte !

(*) voir conditions au verso

Tester un sous-modèle

- Modèle complet (Ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_p$
- Sous-modèle (ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_q$, $q < p$
- $H : \beta_{q+1} = \cdots = \beta_p = 0$ vs. A : au moins 1 $\beta_i \neq 0$, $q + 1 \leq i \leq p$

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)
ω	q	$SCM(\omega)$	SCM/q
termes suppl.	$p-q$	$SCE(\omega) - SCE(\Omega)$	$(SCE(\omega) - SCE(\Omega))/(p-q)$
erreur	$n-p-1$	$SCE(\Omega)$	$SCE(\Omega)/(n-p-1)$
total (corr.)	$n-1$	SCT	

- La statistique F pour tester la signification des termes supplémentaires dans Ω est :

$$F_{obs} = \frac{(SCE(\omega) - SCE(\Omega))/(p-q)}{SCE(\Omega)/(n-p-1)} \sim F_{p-q,n-p-1} \text{ sous } H$$

- Donc on REJETTE H lorsque $F_{obs} > F_{p-q,n-p-1}(1-\alpha)$

Exemple 10.1

Pour un échantillon aléatoire de communes, on a les données suivants :

- Y = pourcentage des adultes qui votent
- X_1 = pourcentage des adultes proprietaires
- X_2 = pourcentage des adultes personnes de couleur
- X_3 = revenu médiane de la famille (milliers CHF)
- X_4 = âge médiane
- X_5 = pourcentage des adultes résident au moins 10 années

Exemple 10.1, cont.

(a) Remplir les sorties :

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	----	---	----	----	----	
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0500	----	----

Exemple 10.1, cont.

(b) Écrire l'équation de prédiction et interpréter le coefficient pour « % résidents adultes propriétaires de maisons ».

(c) Semble-t-il nécessaire d'inclure toute ces variables explicatrices dans le modèle ? Expliquer.

(d) La valeur F est utilisée pour quel test ? Interpreter le résultat de ce test.

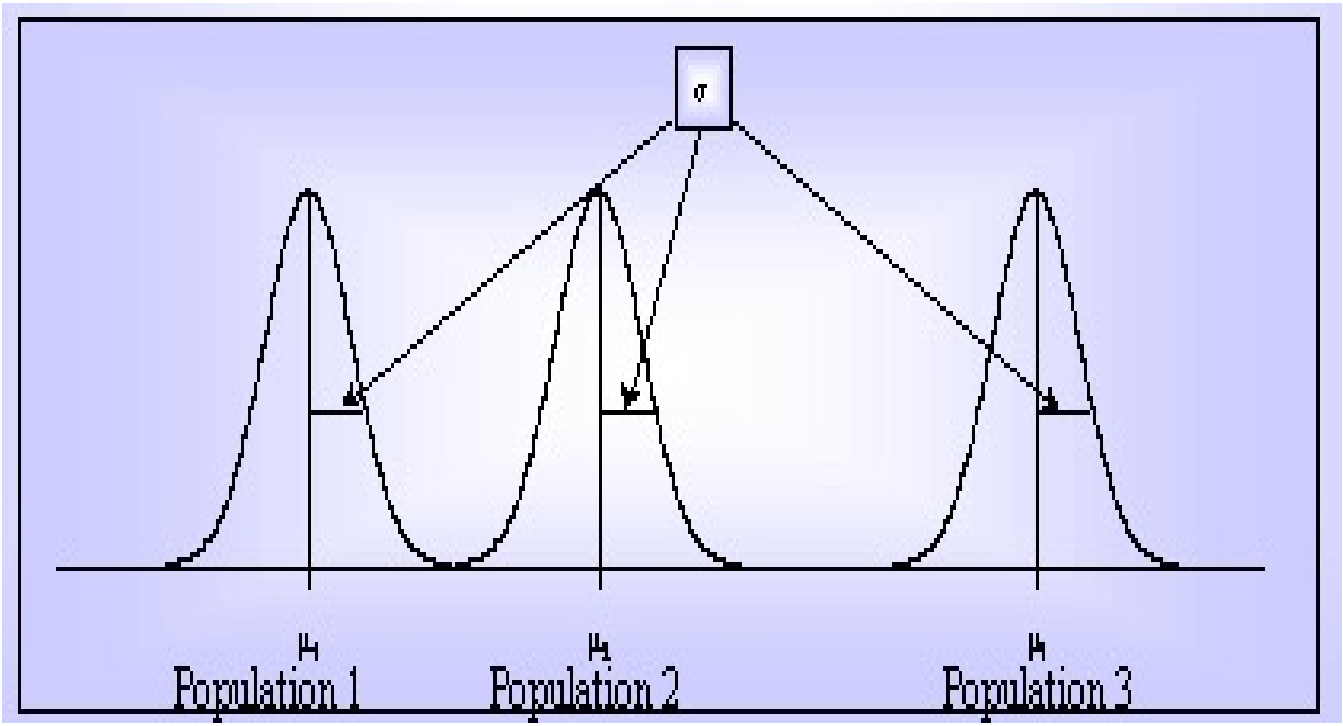
(e) La valeur t de la variable X_1 est utilisée pour quel test ? Interpreter le résultat de ce test.

(f) Donner un IC à 95% pour le changement de la moyenne d'Y quand le pourcentage de proprietaires augmente par 1, en contrôlant pour les effets des autres variables ; l'interpreter.

(g) Donner un IC à 95% pour le changement de la moyenne d'Y quand le pourcentage de proprietaires augmente par 50, en contrôlant pour les effets des autres variables ; l'interpreter.

ANOVA

- Abréviation de *AN*alysis *O*f *VA*riance (analyse de variance)
- Mais c'est un test de différences des *moyennes*
- L'idée :



Principe du test

- L'analyse de variance à un facteur teste l'effet d'un *facteur* A ayant k modalités sur les moyennes d'une *variable quantitative* X
- Les hypothèses testées sont les suivantes :

$$H : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \text{ contre } A : \exists\, \mu_i \neq \mu_j$$

- Tester si le rapport de ces 2 estimateurs de variance est proche de 1
- Les estimations des variances associées [*carré moyen*] sont :
  - *Variance totale* : $SCE_{totale}/(n-1)$
  - *Variance due au facteur A* (CM_{trts}) : $SCE_{trts}/(k-1)$
    $\Longrightarrow$ estimateur de σ^2 si H est vraie
  - *Variance résiduelle* (CM_{erreur}) : $SCE_{erreur}/(n-k)$
    $\Longrightarrow$ estimateur de σ^2 quelque soit le modèle

Les modèles

- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- Sous H, le modèle est :

$$x_{ij} = \mu + \epsilon_{ij}$$

- Sous A, le modèle est :

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

où α_i est *l'effet de la modalité i du facteur A sur la variable X*

- Pour chaque modèle, on peut produire un estimateur de la variance résiduelle

Paires de tests : pourquoi pas ?

Pourquoi ne pas commencer en faisant des tests (z ou t) pour chaque paire d'échantillons ?

- Pour m comparaisons (indépendantes), la probabilité de rejeter au moins un H peut s'écrire : $\alpha_m = 1 - (1 - \alpha)^m$; pour $\alpha = 0.05$:
- 3 tests $\Longrightarrow$ l'erreur de type I $= 0.14$
- 5 tests $\Longrightarrow$ l'erreur de type I $= 0.23$
- 10 tests $\Longrightarrow$ l'erreur de type I $= 0.4$
- 21 tests $\Longrightarrow$ l'erreur de type I $= 0.66$

Statistique de test

- Sous H, $SCE_{trts}/(k-1)$ et $SCE_{erreur}/(n-k)$ $\Rightarrow$ *estimateurs du même paramètre* σ^2
- Donc (sous H), le rapport $\dfrac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)} \approx 1$
- Sous A, au moins 1 $\alpha_i \neq 0$ et $SCE_{erreur}/(n-k)$ est un unique estimateur de σ^2 ; $SCE_{trts}/(k-1) >> SCE_{erreur}/(n-k)$
- Donc (sous A), le rapport $\dfrac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)}$ *très supérieur à 1*
- $\Rightarrow$ Test *unilatéral* dans tous les cas
- $F_{obs} = \dfrac{SCE_{trts}/(k-1)}{SCE_{erreur}/(n-k)} = CM_{trts}/CM_{erreur}$
- Statistique de test distribuée selon une loi de Fisher à $k-1$ (numérateur) et $n-k$ (dénominateur) degrés de liberté (df = degrees of freedom)

Tableau d'ANOVA

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)	F	p-valeur
traitements	$k-1$	SCE_{trts}	$SCE_{trts}/(k-1)$	CM_{trts}/CM_{erreur}	$P(F_{obs} > F_{k-1,n-k})$
erreur	$n-k$	SCE_{erreur}	$SCE_{erreur}/(n-k) (= \hat{\sigma}^2)$		
total (corr.)	$n-1$	SCE_{totale}			

- Sortie d'ordinateur – ANOVA

```
> redcell.aov<-aov(Folate~Group)
> summary(redcell.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
Group        2  15516    7758  3.7113 0.04359 *
Residuals   19  39716    2090
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

*** Suppositions ***

- *Indépendance* : Les k échantillons comparés sont indépendants ; l'ensemble des n individus est réparti *au hasard* (*randomisation*) entre les k modalités du facteur contrôlé A, n_i individus recevant le traitement i.
- *Homoscédasticité* : Les k populations comparées ont la même variance ; le facteur A agit seulement sur la *moyenne* de la variable X et ne change pas sa variance
- *Normalité* : La variable quantitative étudiée suit une loi normale dans les k populations comparées (ou TCL s'applique pour les n_i 'suffisament grands')
- (*voir diapos / vidéo pour l'évaluation du modèle*, qui **NE SERA PAS EXAMINÉE**)

Exemple 10.2

- Les mélanges de mortier sont généralement classés en fonction de leur résistance à la compression, de leurs propriétés d'adhérence et de leur flexibilité.
- Dans le cadre d'un projet de construction, des ingénieurs ont souhaité comparer spécifiquement les résistances moyennes de quatre types de mortiers :
  1. Mortier de ciment ordinaire (MCO)
  2. Mortier imprégné de polymères (MIP)
  3. Mortier de résine (MR)
  4. Mortier de ciment polymère (MCP)
- Des échantillons aléatoires de chaque type de mortier ont été prélevés. Chaque échantillon a été soumis à un essai de compression pour mesurer sa résistance (MPa).

Exemple 10.2, cont.

- Une première question que les ingénieurs peuvent se poser est la suivante : « Les résistances moyennes des mortiers (dans les 'populations' des mortiers) sont-elles égales pour les quatre types de mortiers ? Ou sont-elles différentes ? »
- On prend un échantillon de taille $n = 36$, réparti comme la suite : 8 échantillons du groupe MCO ; 10 échantillons du groupe MIP ; 10 échantillons du groupe MR ; 8 échantillons du groupe MCP.

Tableau d'ANOVA

source	df	*SC*	*CM*	*F*	*p*-valeur
			506.96		9.576e-07
erreur					
total (corr.)		2483.74			

- Quelles sont vos conclusions ?

Qu'est-ce que cela veut dire quand on rejette H ?

- L'hypothèse nulle H est **conjointe** : que *toutes* les moyennes des populations sont égales
- Lorsqu'on rejette l'hypothèse nulle, cela *ne signifie pas* que les moyennes sont toutes différentes ! !
- Cela signifie qu'*au moins une* est différente
- Pour en savoir qui est différente, on peut faire des tests 'post-hoc'/*a posteriori* (paires de t-tests, par exemple)

ANOVA : après le test

- Une fois que toutes les conditions d'une ANOVA ont été vérifiée et que l'analyse a été effectuée, deux conclusions sont possibles :
  - on rejette H
  - on n'a pas assez de preuves pour rejeter H
- Si on ne rejette pas H, on conclut qu'il n'y a pas de différences significatives entre les groupes
- Si on rejette H, on veut *identifier les modalités/niveaux du facteur* qui sont responsables du résultat significatif (la semaine prochaine)