GC – Probabilités et Statistique

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

Lecture 10

- Multivariate data
- Multiple regression
- R software / interpretion of R output
- Geometry of regression
- Introduction : 1-way ANOVA (anova à une voie)

Multivariate data

Individuals	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}

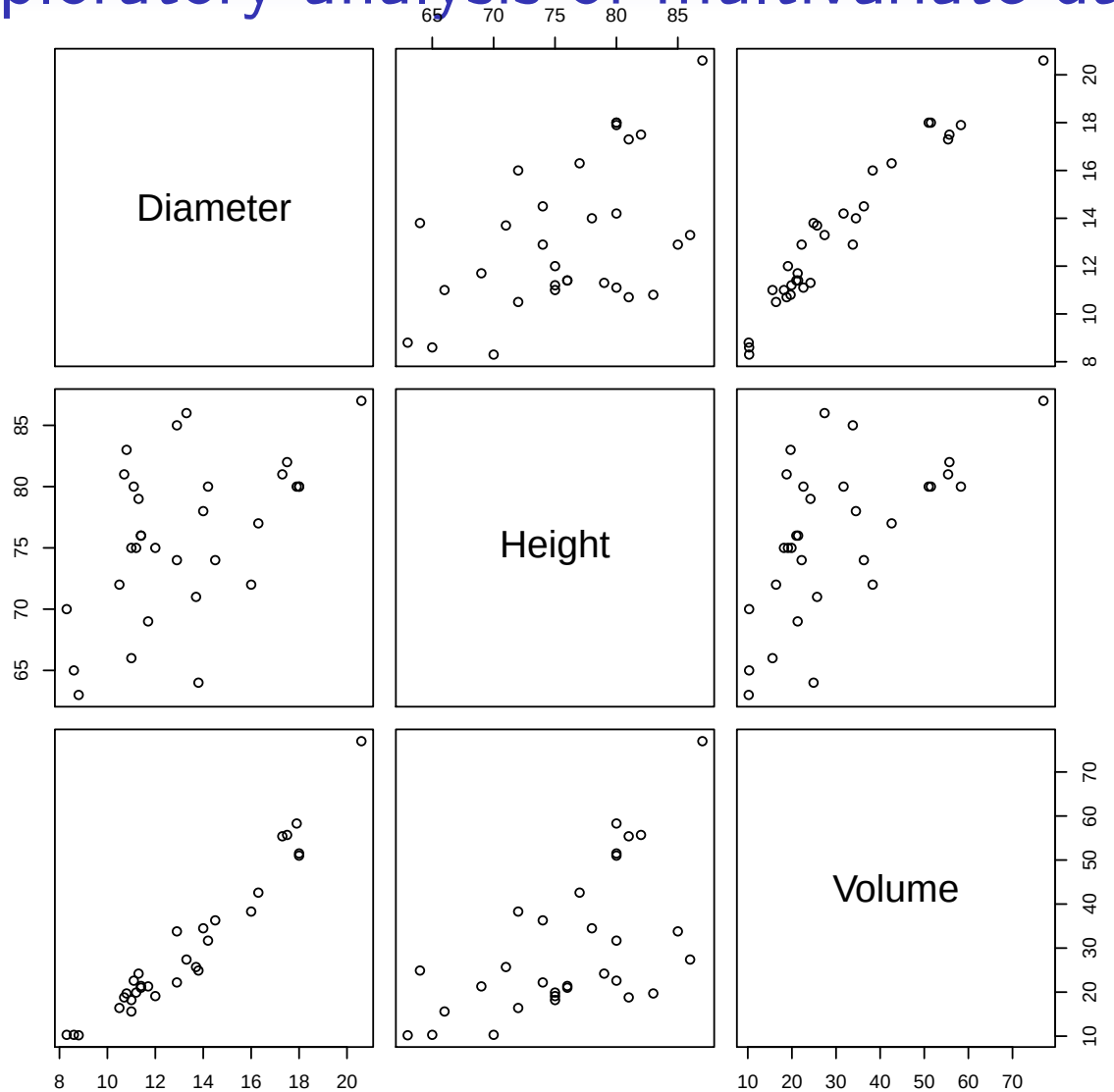
vector of means : $(\overline{x}_1, \ldots, \overline{x}_p)$
matrix of variances-covariances (or *dispersion matrix*) :

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

Example

- A sample of cherry trees was cut and measures were taken of :
  - `Diameter` (inches)
  - `Height` (feet)
  - `Volume` (cubic feet)
- The goal of the collection of these data was to furnish a means of predicting wood volume of the trees, knowing height and diameter 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
- Use a regression model

Exploratory analysis of multivariate data



Multiple regression

- We can have *several* explanatory variables x :

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

- Same assumptions as for simple regression : $\epsilon \sim$ iid $N(0, \sigma^2)$
- Assumptions summarization :
  - **L**inear model (in the parameters)
  - **I**ndependent errors / observations
  - **N**ormal errors / observations
  - **E**qual error variances

Matrix algebra for (simple) regression

- The model :

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

(Ordinary) least squares for multiple regression

- $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \Rightarrow \quad E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$
- Find a solution $\hat{\boldsymbol{\beta}}$ that minimizes the sum of the squared residuals (*ordinary least squares solution (OLS)*) :

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

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

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

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

where $\boldsymbol{X}$ is the *design matrix* and $\boldsymbol{X}'$ is the transpose of $\boldsymbol{X}$

Software : *R*

Why *R* ?

- Powerful, flexible, extensible language and environment for statistical calculation/computation
- Large number of integrated statistical functions statistiques and 'packages'
- High quality, excellent graphical capacities
- Available for Unix / Linux, Windows, Mac
- All this and ... *R* is free !
- http://cran.r-project.org/

Regression estimation output

```
> 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
```

Regression estimation output

```
> 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), $\hat{\beta}_1$ → Diameter, $\hat{\beta}_2$ → Height

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

*** Interpretation of the parameters ***

- The regression coefficients correspond to expected changes in the response for a change of 1 unit in an explanatory / predictor variable
- For simple regression :
  - the slope is the expected change in the response variable (y) if the explantory variable (x) increases by 1 unit
  - the intercept is the predicted value of the response (y) when $x = 0$
- A very important distinction – when there are *several* predictor variables in the equation :
  - each coefficient $\beta_1, \ldots, \beta_p$ corresponds to the contribution of a variable when **all the other variables in the equation are held constant**
  - the coefficient β_0 is the predicted value of the response (y) when **all variables** $x_1, \ldots, x_p = 0$

Properties of the least squares estimator (OLS)

In the case

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

we have :

- Expected value : $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- Variance : $Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$
  $((\boldsymbol{X}'\boldsymbol{X})$ symmetric)
- 0ptimality :
  - The **Gauss-Markov** theorem tells us that *among all linear unbiased estimators*, the least squares estimator (OLS) has *minimum variance*
  - We can summarize that by saying that **the OLS estimator is the « BLUE »** (Best Linear Unbiaised Estimator)

Test/confidence interval for the coefficients

- If we also suppose that $\epsilon_1, \ldots, \epsilon_n \sim$ iid $N(0, \sigma^2)$, we have

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

- Thus, $Var(\hat{\beta}_i) = \sigma^2 \left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{i+1,\, i+1}$
- The CI with level $1-\alpha$ for β_i takes the form :

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

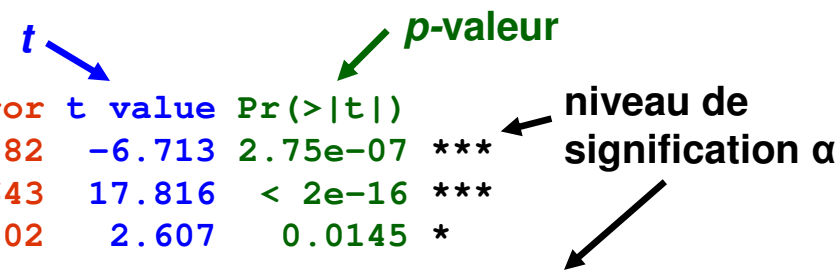
- To test $H : \beta_i = 0$ vs. $A : \beta_i \neq 0$

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

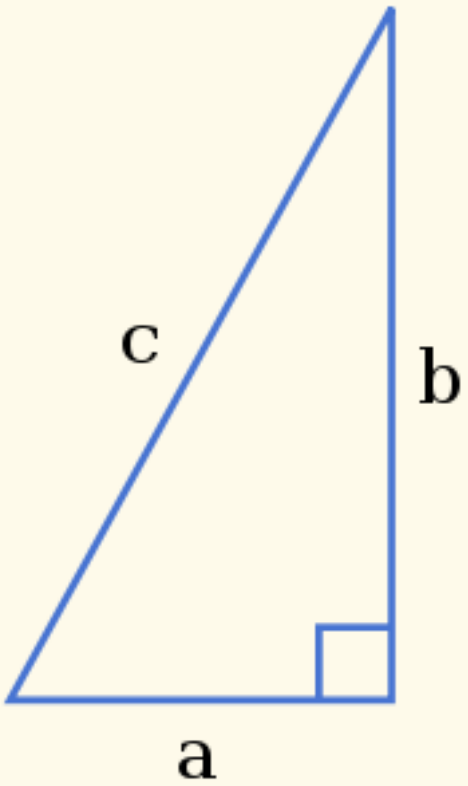
- We REJECT H if : $|t_{obs}| > t_{n-p-1,1-\alpha/2}$
  (equally, if the confidence interval does not contain 0)

Regression estimation output

```
> 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
```



Pythagorean theorem



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

Geometry of least squares

- Consider $\boldsymbol{y}$ as a vector in n-dimensional space
- The column vectors of $\boldsymbol{X}$ form a p-dim subspace
- The predicted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$
  represent the point in the subspace that is closest to the observations : OLS is the *orthogonal projection* of $\boldsymbol{y}$ on the subspace of $\boldsymbol{X}$
- The residual $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is *orthogonal* to vectors in the subspace
- $SSE = \sum e_i^2 = \boldsymbol{e}'\boldsymbol{e}$ is the square of the distance from the vector of obs. to the closest point in the subspace
- Partition $\boldsymbol{y}$ in *two orthogonal components* :
  - $\hat{\boldsymbol{y}}$ ((model subspace, p dims)
  - $\boldsymbol{y} - \hat{\boldsymbol{y}}$ ((error subspace, $n - p$ dims)
- (degrees of freedom correspond to the subspace dims)

Geometry of LS

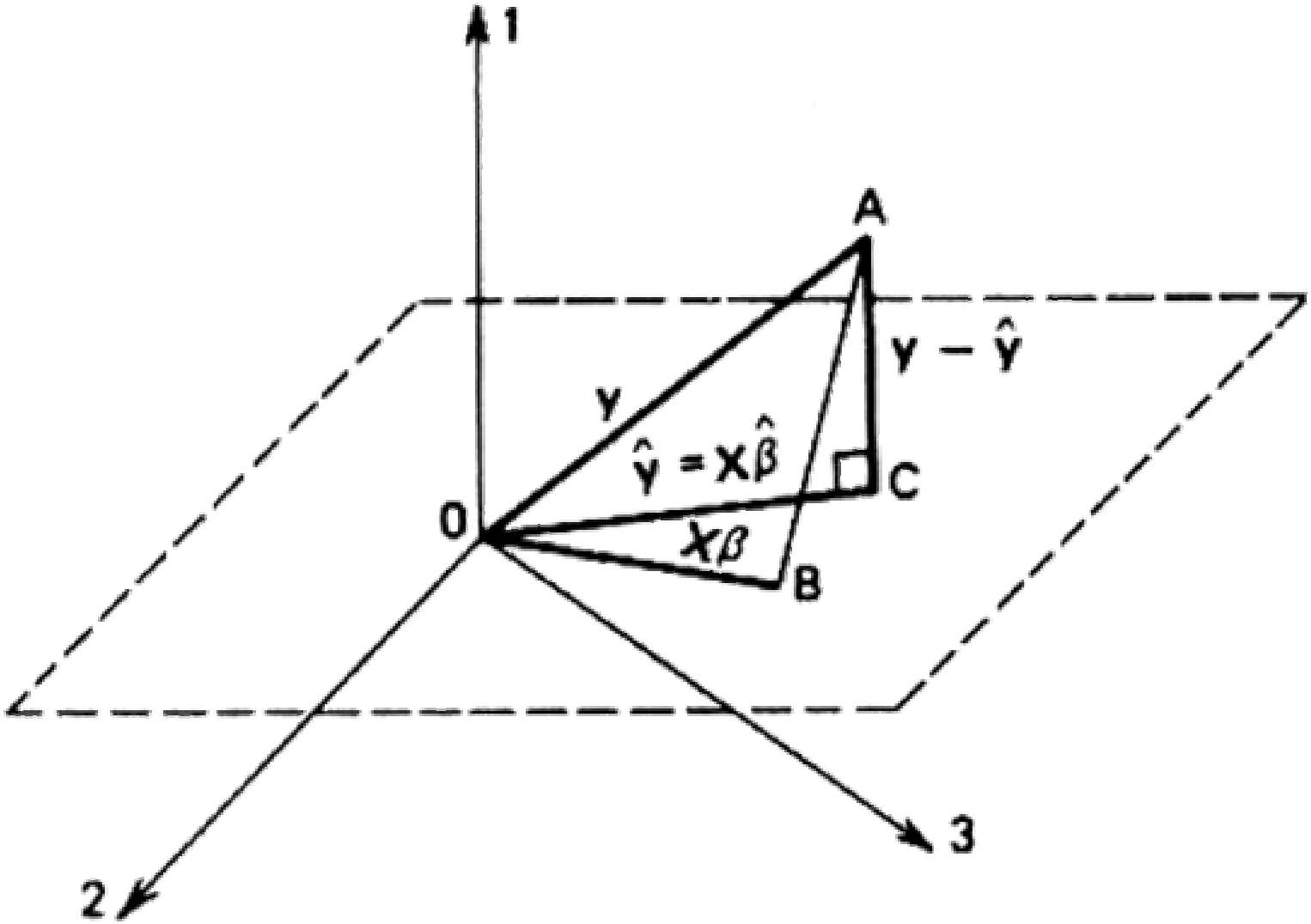


Figure 4.2 A geometrical interpretation of least squares.

Analysis of variance table for regression

- Uses the Pythagorean theorem to *partition the total sum of squares (SST)*
- Pythagorean theoreme :

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

- equally :

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

- We can present this equality in the form of a table :

ANOVA table

source	df	SS	MS ($=SS/\text{df}$)	F	p-value
regression	p	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	SSR/p	MSR/MSE	$P(F_{obs} > F_{p,n-p-1})$
error	$n-p-1$	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$SSE/(n-p-1) (= \hat{\sigma}^2)$		
total (corr.)	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$			

BREAK

F-test - regression

- The statistic $F_{obs} = MS(\text{source})/MSE$ tests the hypothesis $H_0 : \beta_1 = \ldots = \beta_p = 0$ vs. A : at least 1 $\beta_i \neq 0$
- The distribution of F_{obs} if H is true is *the Fisher distribution* $F_{p,n-p-1}$
- The numerator of F_{obs} is *the variability explained by the regression model*
- The denominator contains *the residual variance*
- Under the null, the expected value of F is 1 and under the alternative the expected value is bigger than 1
- REJECT the null hypothesis H for *large values of F*
- When testing a single coeffcient ($H : \beta_i = 0$), $F_{1,n} = t_n^2$

Regression estimation output

```
> 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 of determination

- The value y_i can be decomposed in two parts : one part *explained by the model* and one part *residual*
- The dispersion for the data can therefore be decomposed as :
  1. variance explained by the regression, and
  2. residual (unexplained) variance
- The *coefficient of determination* (or *multiple correlation*) R^2 is defined as the ratio between the explained and total variance SSR/SST
- Equally, $R^2 = 1 - SSE/SST$
- In *simple regression*, this is just *square of the correlation coeffcient*

Adjusted coefficient of determination

- The *adjusted coefficient of determination* R^2_{aj} takes account of the *number of variables*
- In fact, the principal failing of R^2 is that it *increases with the number of explanatory variables*
- An excessive number of variables produces *non-robust* models
- Thus, this measure (R^2_{aj}) is more useful than R^2
- R^2_{aj} is not a true 'square' – it can even be negative

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

Regression estimation output

```
> 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 or adjusted-R^2 ?

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

Testing submodels

- Full model (Ω) : $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- Submodel (ω) : $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$, $q < p$
- $H : \beta_{q+1} = \cdots = \beta_p = 0$ vs. A : at least $1 \beta_i \neq 0$, $q + 1 \leq i \leq p$

ANOVA Table

source	df	SS	MS ($=SS/\text{df}$)
ω	q	$SSM(\omega)$	SSM/q
suppl. terms	$p-q$	$SSE(\omega) - SSE(\Omega)$	$(SSE(\omega) - SSE(\Omega))/(p-q)$
error	$n-p-1$	$SSE(\Omega)$	$SSE(\Omega)/(n-p-1)$
total (corr.)	$n-1$	SCT	

- The F-statistic for testing the significance of the extra terms in Ω is :

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

- Thus we REJECT H when $F_{obs} > F_{p-q,n-p-1}(1-\alpha)$

Example 10.1

For a random sample of 66 communes, we have the following data :

- Y = percentage of adults who vote
- X_1 = percentage of adults who own property
- X_2 = percentage of adults who are persons of color
- X_3 = median family income (thousands of CHF)
- X_4 = median age
- X_5 = percentage of adults resident at least 10 years

Example 10.1, cont.

(a) Complete the output :

	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	----	----

Example 10.1, cont.

(b) Write the prediction equation and interpret the coefficient for '% adults who own property'

(c) Does it seem necessary to include all of these explanatory variable in the model ? Explain.

(d) The F-value is used for which test ? Interpret the result of this test.

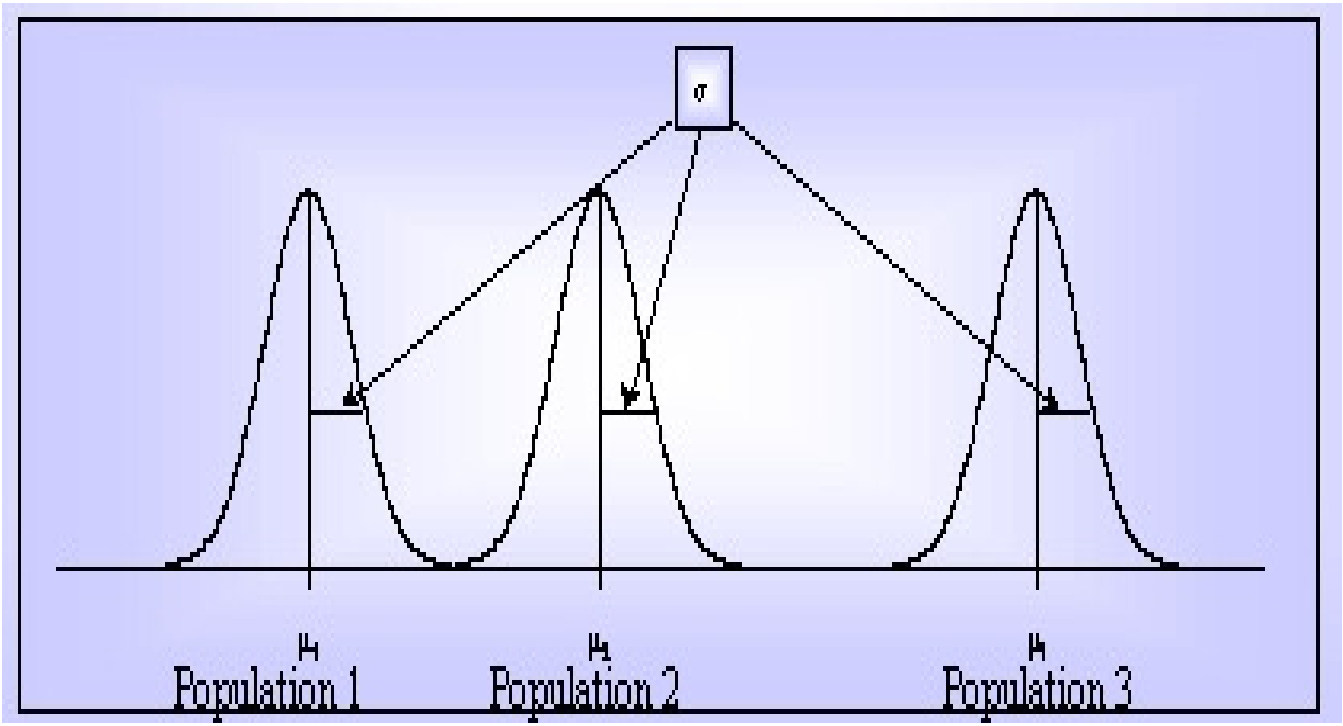
(e) The t-value for the variable X_1 is used for which test ? Interpret the result of this test.

(f) Give a 95% CI for the average change in Y when the percentage of property owners increases by 1, controlling for the effects of the other variables ; interpret.

(g) Give a 95% CI for the average change in Y when the percentage of property owners increases by 50, controlling for the effects of the other variables ; interpret.

ANOVA

- Abbreviation for *AN*alysis *O*f *VA*riance (analyse de variance)
- But it's a test for a difference in *means*
- The idea :



Test principle

- 1-factor analysis of variance tests the effect of one *factor* A having k modalities on the means of a *quantitative variable* X
- The tested hypotheses are :

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

- Test if the ratio of 2 variance estimators is close to 1
- The associated variance estimators are :
  - *Total variance* : $SS_{total}/(n-1)$
  - *Variance due to factor A* (MS_{trts}) : $SS_{trts}/(k-1)$
    $\Longrightarrow$ estimator of σ^2 is H is true
  - *Residual variance* (MS_{error}) : $SS_{error}/(n-k)$
    $\Longrightarrow$ estimator of σ^2 whichever model

The models

- $\epsilon_{ij} \sim$ iid $N(0, \sigma^2)$
- Under H, the model is :

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

- Under A, the model is :

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

where α_i ia *the effect of modality/level i* of factor A on the variable X

- For each model, we can derive an estimator for the residual variance

Pairs of tests : why not ?

Why not start off by carrying out tests (z or t) for each pair of samples ?

- For m comparisons (independent), the probability of rejecting at least one H can be expressed as $\alpha_m = 1 - (1 - \alpha)^m$; now, for $\alpha = 0.05$:
- 3 tests $\Longrightarrow$ Type I error $= 0.14$
- 5 tests $\Longrightarrow$ Type I error $= 0.23$
- 10 tests $\Longrightarrow$ Type I error $= 0.4$
- 21 tests $\Longrightarrow$ Type I error $= 0.66$

$\Longrightarrow$ Type I error no longer controlled at level $\alpha = 0.05$ (anti-conservative/liberal test)

Test statistic

- Under H, $SS_{trts}/(k-1)$ and $SS_{error}/(n-k)$ $\Rightarrow$ *estimators of the same parameter* σ^2
- Thus (under H), the ratio $\dfrac{SS_{trts}/(k-1)}{SS_{error}/(n-k)} \approx 1$
- Under A, at least 1 $\alpha_i \neq 0$ and $SS_{error}/(n-k)$ is a unique estimator of σ^2 ; $SS_{trts}/(k-1) >> SC_{error}/(n-k)$
- Thus (under A), the ratio $\dfrac{SS_{trts}/(k-1)}{SS_{error}/(n-k)}$ *much larger than 1*
- $\Rightarrow$ F-Test *unilateral* in every case
- $F_{obs} = \dfrac{SS_{trts}/(k-1)}{SS_{error}/(n-k)} = MS_{trts}/MS_{error}$
- Test statistic is distributed according to a Fisher F distribution, with $k-1$ (numerator) and $n-k$ (denominator) degrees of freedom (df)

ANOVA table

ANOVA table

source	df	SS	MS $(=SS/\text{df})$	F	p-value
treatments	$k-1$	SS_{trts}	$SS_{trts}/(k-1)$	MS_{trts}/MS_{error}	$P(F_{obs} >$
error	$n-k$	SS_{error}	$SS_{error}/(n-k)(=\hat{\sigma}^2)$		$F_{k-1,n-k})$
total (corr.)	$n-1$	SS_{total}			

- 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
```

*** Assumptions ***

- *Independence* : The k groups (samples) are independent, as well as the individuals within groups ; the ensemble of the n individuals are placed *at random* (*randomization*) between the k modalities for the controlled factor A, with n_i individuals receiving treatment i.
- *Homoscedasticity* : The k populations have the same variance ; the factor A acts only on the *mean* of the variable X and does not change its variance
- *Normality* : The variable studied follows a Normal distribution in the k populations compared (or the CLT applied to the means if the n_i are 'sufficiently large')
- (*watch video for model evaluation*, which **WILL NOT BE EXAMINED**)

Example 10.2

- Mortar mixes are usually classified on the basis of compressive strength and their bonding propoerties and flexibility.
- In a building project, engineers wanted to compare specifically the population mean strengths of four trypes of mortars :
  1. Ordinary cement mortar (OCM)
  2. Polymer impregnated mortar (PIM)
  3. Resin mortar (RM)
  4. Polymer cement mortar (PCM)
- Random samples of specimens of each mortar type were taken. Each specimen was subjected to a compression test to measure strength (MPa).

Example 10.2, cont.

- An initial question that engineers may have is the following : 'Are the population mean mortar strengths equal among the four types of mortars ? Or, are the population means different ?'
- We take a sample of size $n = 36$, distributed as follows : 8 samples from group OCM ; 10 samples from group PIM ; 10 samples from group RM ; 8 samples from group PCM.

Tableau d'ANOVA

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

- What are your conclusions ?

What does it mean when we reject H ?

- The null hypothesis H is a **joint** (global) one : that *all* the population means are equal
- When we reject the null hypothesis, that does *not* mean that all the means are different ! !
- It means that *at least one* is different
- To know which is different, we can carry out '*post hoc*'/*a posteriori* tests (pairs of tests, for example)

ANOVA : after the test

- Once all the conditions for an ANOVA have been verified and the analysis carried out, two conclusions are possible :
  - we reject H
  - we do not have enough evidence to reject H
- If H is not rejected, we conclude that there are not significant differences between group means
- If we DO reject H, typically we are interested in *identifying the modalities/factor levels* that are responsible for the significant result