C6. Régresser
(et progresser en statistiques)

Comment lire ?

- Âge ?
  - paramètre -0.16
  - ***
  - ‘Écart’-type 0,01
- Corpulence ?
- Profession
  - Agriculteur ?
  - Ouvrier : Réf ?
- Constante

	Hommes		Femmes	
	Paramètre	Ecart-type	Paramètre	Ecart-type
Constante	181.9 ***	0.74	168 ***	0.68
Corpulence	0.15	0.17	-0.57 ***	0.14
Âge de la personne	-0.16 ***	0.01	-0.09 ***	0.01
Région habitée				
Région parisienne	-1.1 *	0.58	-2.44 ***	0.52
Bassin parisien	-1.23 ***	0.48	-1.85 ***	0.43
Méditerranée	-1.58 ***	0.54	-1.78 ***	0.49
Est	-0.52	0.6	-0.89	0.54
Ouest	-2.21 ***	0.51	-2.89 ***	0.45
Sud-Ouest	-1.74 ***	0.55	-2.6 ***	0.49
Centre-Est	-1.65 ***	0.54	-1.93 ***	0.49
Nord	Ref.		Ref.	
Profession de la personne				
Agriculteur	2.26 ***	0.54	0.79	0.53
Artisan, commerçant, entrepreneur	2.16 ***	0.45	1.6 ***	0.53
Cadre, profession libérale, prof. intell. supérieure	2.67 ***	0.4	2.35 ***	0.5
Profession intermédiaire	2.01 ***	0.33	1.38 ***	0.38
Employé	1.72 ***	0.41	1.08 ***	0.31
Ouvrier	Ref.		Ref.	
Profession du père				
Agriculteur	-0.17	0.37	0.55 *	0.32
Artisan, commerçant, entrepreneur	0.49	0.4	0.14	0.36
Cadre, profession libérale, prof. intell. supérieure	0.69	0.49	0.67	0.43
Profession intermédiaire	0.94 **	0.43	0.95 ***	0.36
Employé	1.1 ***	0.4	0.51	0.35
Ouvrier	Ref.		Ref.	
Âge auquel la personne quitte l’école				
13 ans et moins	-1.06 ***	0.39	-0.55 *	0.33
14 ou 15 ans	-1.04 **	0.48	-0.44	0.41
16 ou 17 ans	-0.71	0.49	-0.77 *	0.43
18 ou 19 ans	-0.33	0.47	-0.62	0.4
20, 21 ou 22 ans	-0.35	0.48	-0.81 **	0.39
23 ans et plus	Ref.		Ref.	

Lecture : la taille de l'homme et celle de la femme sont régressées séparément sur le même ensemble de variables.
*** : significatif au seuil de 1 %, ** : significatif au seuil de 5 %, * : significatif au seuil de 10 %,
Réf. : catégorie de référence.
Champ : 30 ans et plus, France métropolitaine. Source : Panel européen, vague 2001, Insee.

Comment prédire la taille de ces personnes

- Femme, IMC 18, 30 ans, région parisienne, cadre, fille d'ouvrier, ayant quitté l'école à 22 ans

- Femme, IMC 25, 50 ans, Sud-Ouest, ouvrière, fille d'agriculteur, ayant quitté l'école à 14 ans

D'après la régression : quelle est la différence de taille entre

- Deux hommes aux caractéristiques identiques, sauf l'âge :
  - 40 ans versus 50 ans ?
  - 30 ans versus 40 ans ?
- Deux hommes aux caractéristiques identiques, sauf la CS ?
  - Cadre versus Ouvrier ?
  - Agriculteur versus employé ?

Introduction historique : de la droite de régression vers la médiocrité à la droite de régression

- Galton (cousin de Darwin). Mesure de la taille des descendants en fonction de la taille des ascendants (petits pois, puis hommes).
- Trace une droite pour représenter les données :

  (Taille des fils-Taille moyenne des fils) =2/3 * (taille des pères-taille moyenne des pères)

- 1886 : « Regression towards mediocrity in heredity stature ».
- Cette première droite appelée : droite de régression vers la médiocrité.

Une démarche inférentielle

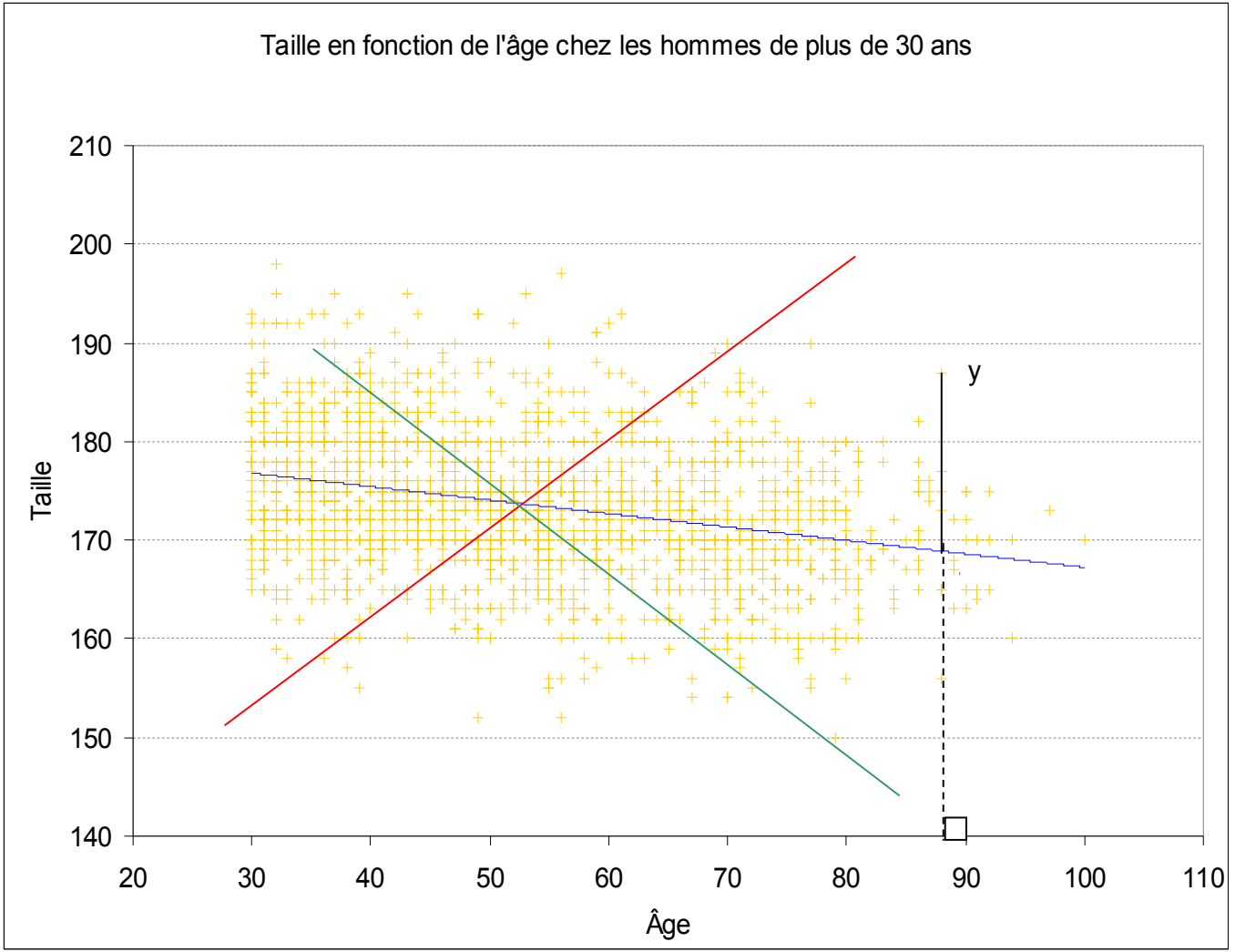
- Variable expliquée : Une variable d'intérêt que l'on cherche à :
  - Expliquer
  - Prévoir
- À partir des valeurs prises par d'autres variables : les variables explicatives

La régression linéaire à une variable

- Exemple : on cherche à expliquer la taille par l'âge chez les adultes
  - La variable expliquée, la taille, est une variable continue
  - La variable explicative, l'âge, est aussi une variable continue.
- On estime une relation linéaire :
  - Taille= a + b. Âge + erreur
  - Présentation plus fréquente :
  - $Y = a + b.X + u$
- On cherche donc a et b en essayant de limiter l'erreur avec la technique des « Moindres carrés ordinaires ».

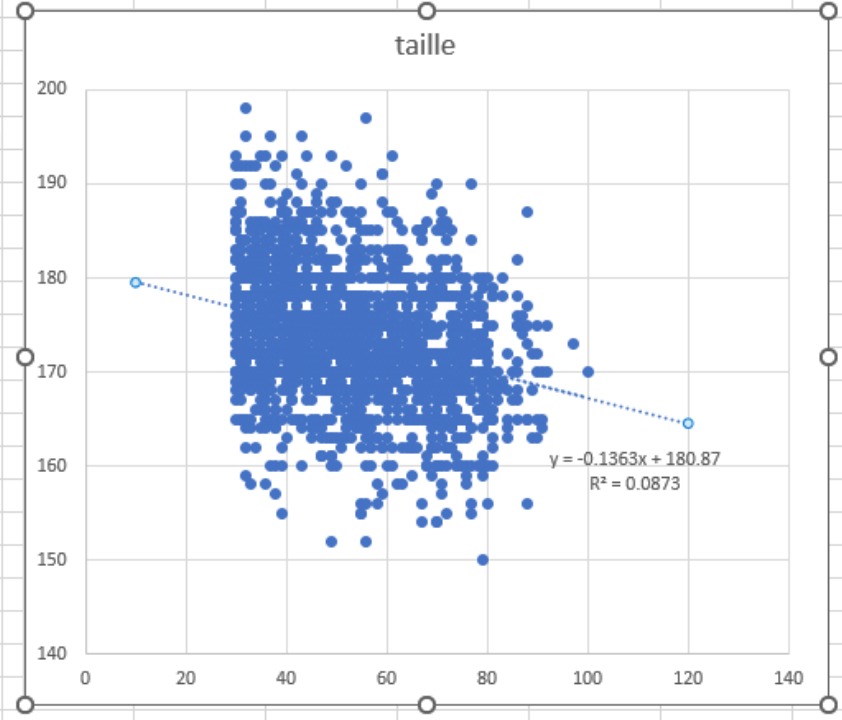
Présentation graphique

- Moindre carré ordinaire :
- Minimisation du « carré des erreurs »
- Erreur : distance entre le point et la prédiction



Avec Excel

	A	B
1	age	taille
2	30	167
3	30	192
4	30	186
5	30	170
6	30	185
7	30	178
8	30	168
9	30	174
10	30	182
11	30	173
12	30	178
13	30	172
14	30	177
15	30	169
16	30	187
17	30	178
18	30	175
19	30	180
20	30	170
21	30	185
22	30	176
23	30	183
24	30	186
25	30	182
26	30	187
27	30	180
28	30	190



Format de courbe de tendance

Options de courbe de tendance

Options de courbe de tendance

- Exponentielle
- Linéaire
- Logarithmique
- Polynomiale — Degré 2
- Puissance
- Moyenne mobile — Période 2

Nom de la courbe de tendance

- Automatique — Linéaire (taille)
- Personnalisé

Prévision

- En avant 20.0 périodes
- En arrière 20.0 périodes
- Définir l'interception 0.0
- Afficher l'équation sur le graphique
- Afficher le coefficient de détermination (R^2) sur le graphique

La technique des moindres carrés ordinaires (MCO)

- Pour trouver a et b, on minimise le carré des erreurs (i.e. moindres carrés), soit le carré de l'écart entre la taille observée et la taille prédite.
- Solutions analytiques dans le cas de la régression à deux variables:

$$b = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} = \frac{\mathrm{cov}(X,Y)}{V(X)} \qquad a = \overline{Y} - b.\overline{X}$$

  - b, c'est la variation (moyenne) de Y (la taille) consécutive à la variation d'une unité de X (l'âge)
  - a, la constante. C'est la taille si X=0 (ce qui dans le cas de l'âge adulte n'a pas de sens concret – des adultes de l'âge de 0 ans)

- Ex : a= 180.87 b= -0.136.

La technique des moindres carrés ordinaires (MCO)

- Les paramètres a et b sont des « moyennes ».
  - Comme toute moyenne, on peut leur associer des écarts-types de la moyenne ou erreurs-types.
- Les erreurs types des paramètres (dans le cas de la régression à une seule variable explicative)

$$s(b) = \frac{s}{\sqrt{\sum_i (X_i - X)^2}} = \frac{s}{\sqrt{n.V(X)}} \qquad s(a) = \sqrt{\frac{s^2 \sum_i X_i}{n. \sum_i (X_i - \overline{X})^2}}$$

- Avec $s = \sqrt{\frac{\sum_i \hat{u}^2}{n-2}}$, l'écart –type des résidus,

  soit la racine carrée de la moyenne du carré des résidus

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 180.8747     0.5857  308.80   <2e-16 ***
age          -0.1363     0.0106  -12.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 6.729 on 1728 degrees of freedom
Multiple R-squared:  0.0873,    Adjusted R-squared:  0.08678
F-statistic: 165.3 on 1 and 1728 DF,  p-value: < 2.2e-16
```

Présentation canonique

- Les coefficients en colonne
- Suivi des erreurs types entre parenthèses
- Le R2
- Le nombre d'observations
- Les indicateurs (***) de significativité

```
============================
                  Model 1
----------------------------
(Intercept)     180.875 ***
                  (0.586)
age              -0.136 ***
                  (0.011)
----------------------------
R^2                0.087
Num. obs.         1730
============================
*** p < 0.01; ** p < 0.05; * p < 0.1
```

Avec Excel :

D	E	F	G	H	I	J	K	L	M	N
taille	imc	age								
167	19.3624727	30						Formule		
192	25.7703993	30				-0.13625271	180.874691	=DROITEREG(D2:D1731;F2:F1731;1;1)		
186	25.1474159	30				0.01059787	0.58573953			
170	19.7231834	30				0.08730421	6.72887482			
185	27.1731191	30				165.292406	1728			
178	24.9337205	30				7484.06931	78239.9631			
168	21.2585034	30								
174	21.4691505	30								
182	26.8687357	30								
173	23.3886866	30								
178	25.2493372	30								
172	23.6614386	30								
177	21.7051294	30								
169	20.3074122	30								
187	26.8809517	30								
178	25.5649539	30								
175	20.244898	30								
180	23.4567901	30								
170	22.1453287	30								
185	21.6216216	30								
176	27.1177686	30								
183	22.3954134	30								
186	21.6788068	30								

Les arguments de DROITREG(…) /LINEST(...)

- Contrainte : **aucune cellule vide ou cellule caractère** dans la plage de données → que des variables quantitatives non vides.
- =DROITEREG(D2:D1731;F2:F1731;1;1)
  - D2:D1731 : plage de la variable dépendante (sans les intitulés de variable)
  - F2:F1731 : plages variables explicatives (sans les intitulés de variable)
  - 1 → constante (1. oui ou 0. non)
  - 1 → détail statistiques (1. oui ou 0. non)

Comprendre la sortie Excel

	age	*constante*	Formule
Paramètres beta	-0.136252707	180.874691	=DROITEREG(D2:D1731;F2:F1731;1;1)
Erreurs-types	0.010597872	0.58573953	
R2	0.087304214	6.72887482	*Erreur type résiduelle*
Statistique F de Fisher	165.2924063	1728	*Degré de liberté*
Somme des carrés expliqués	7484.069308	78239.9631	*Somme des carrés des résidus*

- La sortie est à l'envers : dernière variable explicative en premier, première en avant dernier, constante en dernier.

Que lire dans une régression ?

- Paramètre
  - Signe et valeur
- Significativité
  - Et pour ça il faut connaître l'erreur-type et le DL du paramètre
- R2
  - Statistique de qualité du modèle… d'une qualité discutable

Le paramètre

- Signe → Liaison positive / Liaison négative
- Valeur : de combien y change quand x augmente d'une unité
  - Ex. une année de plus d'âge → -0.13 cm en taille
- Dépend de l'échelle de mesure de x
  - En années, mois, jours ou décennies → variation du coefficient
- Possibilité de standardiser les variables explicatives (variables quantitatives surtout)
  - Avant la régression on peut diviser la variable explicative par son écart-type. Lecture : « 1 écart-type de x augment y de tant ».

Le R^2 ou la qualité du modèle

- Somme des carrés de y = Somme des carrés de $\hat{y}$ + Somme des carrés des erreurs
- « Variance totale » = « variance expliquée » + « variance résiduelle »
- R^2, ou « la part de la variance expliquée »
  - R^2=variance expliquée / variance totale
    - Exemple : R^2=8,8%
- Le R^2 n'explique pas grand-chose… Un modèle avec un R^2 faible peut être tout-à-fait valable, un modèle avec un R^2 élevé peut être biaisé.
- Le R^2 permet surtout de comparer la qualité de modèles similaires (même jeu de variables), mais pas de modèles de nature différente

Le degré de liberté

- Correspond à :

  DL=Nombre d'observations-Nombre de variables explicatives-1

  DL = N – K - 1

- Nécessaire pour faire les tests de Student
- Calculé dans la sortie Excel

L'erreur-type

- Parfois appelée abusivement écart-type
- Permet de mesurer l'intervalle de confiance du paramètre *b*
- *b* se situe dans l'IC à 95 % (pour DL > 120)
  - [b-1.96*erreur-type; b+1.96*erreur-type]
  - Ex. [-0.136-1.96*0.011; -0.136+1.96*0.011]=[-0.158 ; -0.114]
- Permet de calculer des tests de Student

Le test de nullité des paramètres

- La pente b est-elle vraiment significative ? Peut-on dire raisonnablement qu'on ne peut écarter l'idée d'une relation négative… → Test de Student de nullité du paramètre.
  - H0 : b=0. (On suppose que dans notre population théorique notre variable n'a pas d'impact ; → Hypothèse contrariante qu'on veut « nullifier »).
  - On calcule la proportion que des échantillons de même taille tirés dans cette population théorique génère une distance aussi importante que la distance à l'hypothèse nulle mesurée dans notre échantillon empirique

Exemple test de student

- Paramètre b= -0.13
- Erreur type σ_b=0.011
- Statistique T= -0.13 /0.011=-12.86
- Degré de liberté : N-K-1 =1728
- P-valeur : Prob($X_{Student} > |T|$)=$3{,}40*10^{-36}$

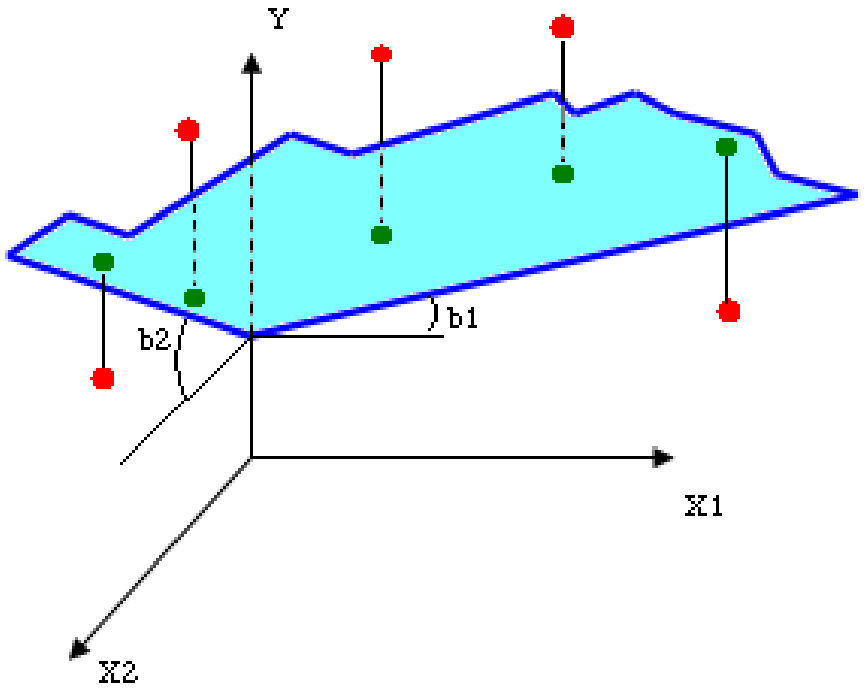
Avec Excel

- Il faut calculer soi-même le test
- Similaire à un test de Student de comparaison de moyennes

	age	*constante*	Formule
Paramètres beta	-0.136252707	180.874691	=DROITEREG(D2:D1731;F2:F1731;1;1)
Erreurs-types	0.010597872	0.58573953	
R2	0.087304214	6.72887482	*Erreur type résiduelle*
Statistique F de Fisher	165.2924063	1728	*Degré de liberté*
Somme des carrés expliqués	7484.069308	78239.9631	*Somme des carrés des résidus*
T de Student	-12.85660944	308.79714	=I4/I5
P-valeur	3.40177E-36	0	=LOI.STUDENT(ABS(I10);J7;2)

La régression à deux variables : présentation graphique

- $y=b_0+b_1.x_1+b_2.x_2+u$



Exemple avec Excel et R

```
==========================
                  Model 1
--------------------------
(Intercept)     181.050 ***
                  (1.222)
imc               -0.007
                  (0.045)
age               -0.136 ***
                  (0.011)
--------------------------
R^2                0.087
Adj. R^2           0.086
Num. obs.       1730
==========================
*** p < 0.01; ** p < 0.05; * p < 0.1
```

	Formule	*age*	*imc*	*constante*
Paramètres beta	=DROITEREG(D2:D1731;E2:F1731;1;1)	-0.136	-0.007	181.050
Erreurs-types		0.011	0.045	1.222
R2	*Erreur type résiduelle*	0.087	6.731	#N/A
Statistique F de Fisher	*Degré de liberté*	82.613	1727.000	#N/A
Somme des carrés expliqués	*Somme des carrés des résidus*	7485.278	78238.754	#N/A
T de Student	=I4/I5	-12.674	-0.163	148.146
P-valeur	=LOI.STUDENT(ABS(I10);J7;2)	0.000	0.870	0.000

La régression multiple :

$y=b_0+b_1.x_1+b_2.x_2+\ldots+b_k.x_k+u$

- On projette les y sur un plan/une surface en k dimensions,
  - parallèlement à l'axe des y (orthogonalement à la surface des xi) de manière à ce que le plan minimise le carré des erreurs entre y et sa projection.
- Formules matricielles [que l'on peut oublier] :

$\tilde{b}=(X'X)^{-1}.X'.y$

$V(\tilde{b})=\tilde{\sigma}^2.(X'.X)^{-1}$

L'interprétation : « toutes choses égales par ailleurs »

- b_i pente du meilleur plan pour approcher y.
- Si un facteur *i* de la régression varie d'une unité alors que les autres restent constants, alors la variation de y est égale à b_i .
- Linéarité : les variables s'additionnent
  - Ex : un effet pour l'âge, un effet pour la corpulence, etc..
  - Une variation simultanée de l'âge et de la corpulence est égale à la somme des paramètres d'âge et de corpulence

L'interprétation : « toutes choses égales par ailleurs »

- Éviter qu' « un effet en cache un autre »
  - L'âge et la corpulence sont deux facteurs très corrélés.
  - Quand on mesure l'effet « seul » de la corpulence sur la taille, on peut se dire ça joue non à cause de la taille elle-même mais parce que ça reflète aussi l'âge.
  - Quand on introduit les deux variables, on arrive à faire la part des deux facteurs. Il y a « suffisamment de cas » où les âges sont égaux et les corpulences différentes et de cas où les corpulences sont égales et les âges différents pour voir ce qui est dû à l'un ou l'autre des facteurs.
  - La lecture : « à âge fixé (contrôlé), la corpulence rajoute tant ». « À corpulence fixée, l'âge rajoute tant ».

Exemple

- AGE et corpulence sont corrélés :
  cor = 0.154 (p<.0001)
- La corpulence toute seule a un effet négatif sur la taille.
- Mais est-ce un effet propre de la corpulence ou un effet de l'âge auquel elle est liée ?
- La régression multivariée montre que c'est plutôt l'effet de la corrélation à l'âge qui impactait la taille

	Model 1	Model 2
(Intercept)	176.094 ***	181.050 ***
	(1.210)	(1.222)
imc	-0.095 **	-0.007
	(0.047)	(0.045)
age		-0.136 ***
		(0.011)
R^2	0.002	0.087
Num. obs.	1730	1730

*** p < 0.01; ** p < 0.05; * p < 0.1

Portée et limite de l'interprétation : toutes choses égales par ailleurs

- « Les choses égales par ailleurs » se limitent aux variables de contrôle !
- On mesure des effets nets « moyens ».
- On considère que ces effets « moyens » s'additionnent.
- L'effet d'ajouter une année d'âge est le même que l'on ait un IMC de 20 ou de 25.
- Sauf spécification contraire, on n'estime pas les effets croisés :
  - Si la taille est croissante en fonction de l'âge pour telle catégorie et décroissante en fonction de l'âge pour telle autre, la régression ne permettra pas de le voir (sauf à introduire une variable croisant l'âge et la catégorie en question). On aura que l'effet moyen de l'âge.
  - Ex. Si la taille décroît très fortement en fonction de l'âge chez les OS, mais stagne en fonction de l'âge chez les cadres, on ne verra pas cet effet. On verra uniquement les effets moyens.

Les deux risques de la régression : 1. La variable manquante

- Autre nom : variable cachée, confondante, hétérogénéité inobservée
- Une (ou plusieurs) variable importante est manquante
- Cette variable est corrélée à la fois et à la variable dépendante et à une des variables explicatives
- La variable explicative va « capturer » le pouvoir explicatif de la variable manquante
- Risque d'erreur d'interprétation

Les deux risques de la régression : 2. La causalité inverse

- La variable explicative dépend de la variable dépendante
- Phénomène : poule et œuf. Ex : Aspirations et réussite, quantité et prix
- Risque d'erreur d'interprétation si on interprète la régression comme l'expression d'une relation causale

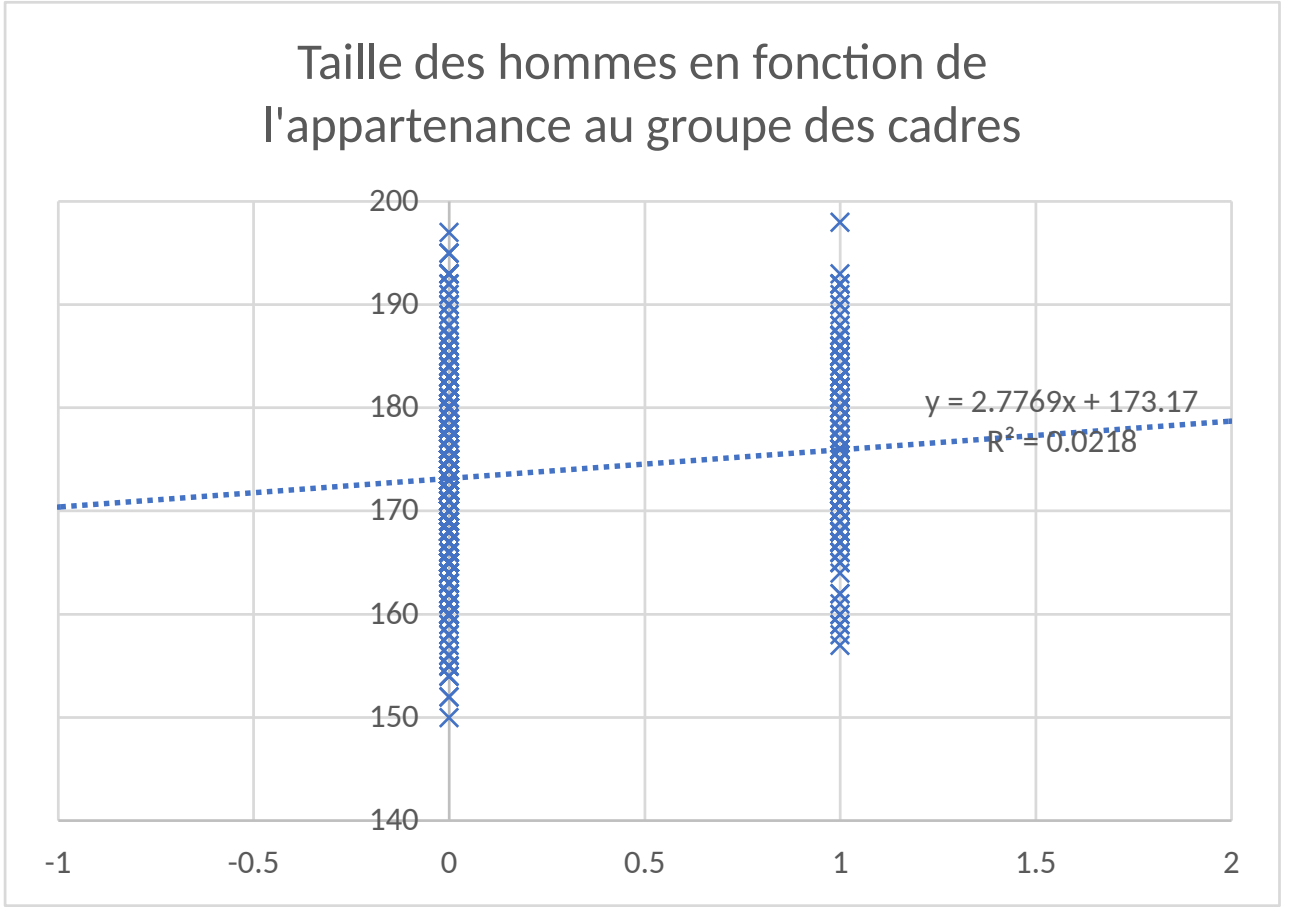
Les solutions face à ces deux risques

- 1. Expérience randomisée
  - La ventilation aléatoire dans le groupe traité / de contrôle assure qu'il n'y a pas de variables cachées qui soient corrélées au traitement
- 2. Expérience « naturelle »
  - Phénomène de la vie sociale qui conduit à une ventilation en groupe similaire à une expérience randomisée
- 3. La prudence dans l'interprétation + réflexivité + de bonnes variables de contrôle
  - Vérifier ce qui se passe avec ou sans.

Les variables qualitatives explicatives

- Comment introduire une variable comme la CS ou la région ?
- Solution on transforme chaque modalité en variable dichotomique. On introduit toutes les modalités SAUF UNE dans la régression.
- La modalité non introduite dans la régression est la modalité de référence.
- Pourquoi ?
  - Réponse technique, les variables seraient liées.
    Ex : Sexe1 + Sexe2= Constante
  - Réponse pratique : Calculer les paramètres *b* associés aux modalités c'est comme placer les barreaux sur une échelle et mesurer leur écartement : il faut le mesurer par rapport à un barreau de référence.

Schéma d'une régression simple avec une variable qualitative explicative



Exemple région et cs

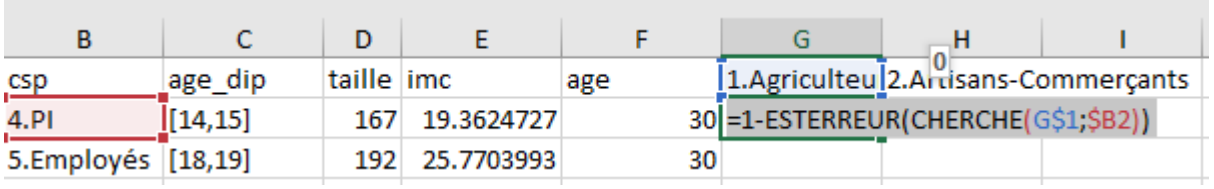
```
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                180.31319    1.44092 125.138  < 2e-16 ***
imc                          0.03071    0.04473   0.687 0.492396
age                         -0.14624    0.01077 -13.578  < 2e-16 ***
reg_1.Region_parisienne     -1.70261    1.01355  -1.680 0.093168 .
reg_2.Bassin_Parisien       -0.23562    0.91875  -0.256 0.797624
reg_5.Est                   -0.40202    0.87803  -0.458 0.647108
reg_6.Ouest                 -2.31350    0.86206  -2.684 0.007352 **
reg_7.Centre-Est            -1.19699    0.85764  -1.396 0.162993
reg_8.Sud-Ouest             -0.89257    0.85562  -1.043 0.297011
reg_9.Mediterrannee         -0.42540    0.86032  -0.494 0.621037
cs_1.Agriculteurs            1.31777    0.67082   1.964 0.049643 *
cs_2.Artisans-Commerçants    2.26543    0.58829   3.851 0.000122 ***
cs_3.Cadres                  3.75679    0.48164   7.800 1.07e-14 ***
cs_4.PI                      1.22319    0.43837   2.790 0.005324 **
cs_5.Employés                0.66542    0.55177   1.206 0.227992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Variables explicatives qualitatives avec Excel

- Il faut créer soit même des variables dichotomiques pour chaque modalité de la variable qualitative en question
  - Si la variable est en colonne B et en mettant les noms des modalités en 1ère ligne

    =1-ESTERREUR(CHERCHE(G$1;$B2))

- On exclut de la plage la modalité dont on veut qu'elle soit la situation de référence



B	C	D	E	F	G	H	I
csp	age_dip	taille	imc	age	1.Agriculteu	2.Artisans-Commerçants	
4.PI	[14,15]	167	19.3624727	30	=1-ESTERREUR(CHERCHE(G$1;$B2))		
5.Employés	[18,19]	192	25.7703993	30			

Variables qualitatives avec Excel

Modèle 1, avec juste IMC, âge et la région

	Formule	reg_9.Me	reg_8.Sud-	reg_7.Cer	reg_6.Ou	reg_5.Est	reg_2.Bas	reg_1.Reg	age	imc	Constante
Paramètres beta	=DROITEREG(D2:D1731;E2:F1731;1;1)	-0.250	-0.758	-1.158	-2.260	-0.309	0.196	-1.036	-0.136	-0.008	181.907
Erreurs-types		0.874	0.868	0.870	0.874	0.893	0.932	1.026	0.011	0.045	1.446
R2	*Erreur type résiduelle*	0.099	6.702	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Statistique F de Fisher	*Degré de liberté*	20.974	1720.000	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Somme des carrés expliqués	*Somme des carrés des résidus*	8477.438	77246.594	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
T de Student	=AB4/AB5	-0.286	-0.873	-1.331	-2.586	-0.346	0.211	-1.010	-12.658	-0.184	125.764
P-valeur	=LOI.STUDENT(ABS(AB10);AC7;2)	0.775	0.383	0.183	0.010	0.730	0.833	0.313	0.000	0.854	0.000

Modèle 2, avec toutes les variables : IMC, âge, région, CS, et âge de fin d'études

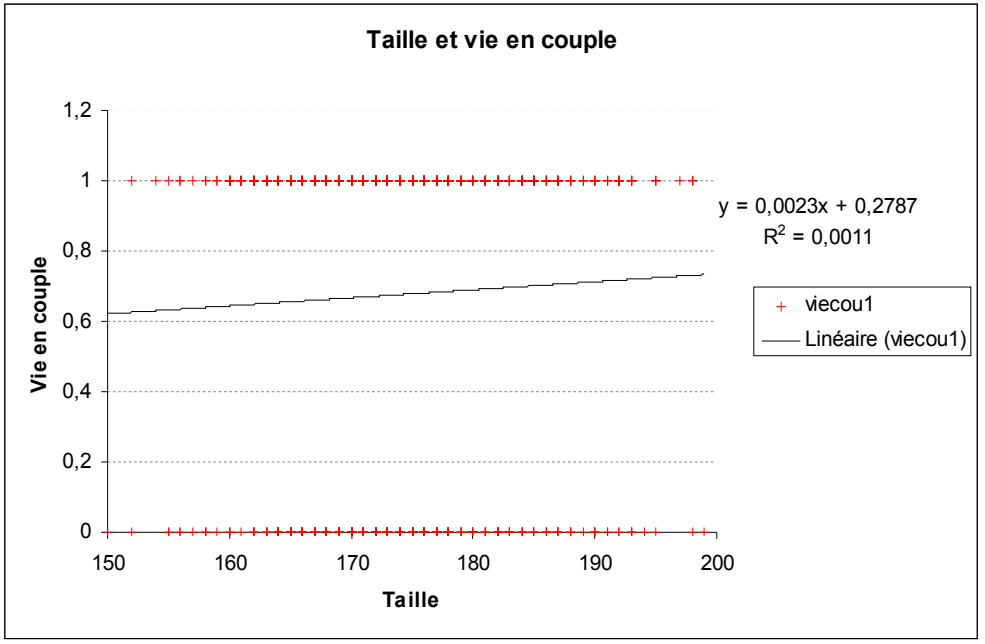
	Formule	age_dip_	age_dip_	age_dip_	age_dip_	age_dip_	cs_5.Emp	cs_4.PI	cs_3.Cadr	cs_2.Artis	cs_1.Agri	reg_9.Me	reg_8.Suc	reg_7.Cer	reg_6.Ou	reg_5.Est	reg_2.Bas	reg_1.Reg	age	imc	Constan
Paramètres l	=DROITEREG	0.148	0.085	-0.044	-4.355	-0.070	0.640	1.211	3.746	2.270	1.308	-0.446	-0.904	-1.204	-2.316	-0.419	-0.267	-1.695	-0.146	0.032	184.60
Erreurs-types		0.318	0.488	0.405	3.836	0.347	0.553	0.441	0.483	0.589	0.672	0.861	0.857	0.859	0.863	0.879	0.920	1.015	0.011	0.045	4.06
R2	*Erreur type r*	0.133	6.592	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Statistique F	*Degré de libe*	13.842	1710.000	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Somme des c	*Somme des c*	########	########	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
T de Student	=I4/I5	0.467	0.173	-0.109	-1.135	-0.202	1.157	2.749	7.760	3.852	1.947	-0.517	-1.055	-1.402	-2.682	-0.477	-0.290	-1.670	-13.348	0.710	45.37
P-valeur	=LOI.STUDEN	0.641	0.862	0.913	0.256	0.840	0.248	0.006	0.000	0.000	0.052	0.605	0.292	0.161	0.007	0.633	0.772	0.095	0.000	0.478	0.00

Comment choisir la situation de référence ?

- Le choix de la situation référence d'une variable ne modifie pas l'estimation des autres variables
- Les estimations sont toutes équivalentes mathématiquement quelle que soit la référence
- Mais l'interprétation peut être influencée par le choix de la situation de référence.
- Prendre la situation la plus courante peut être une bonne solution = permet d'éviter de construite un monstre en situation de référence (ex : diplômé du supérieur, ouvrier non qualifié)

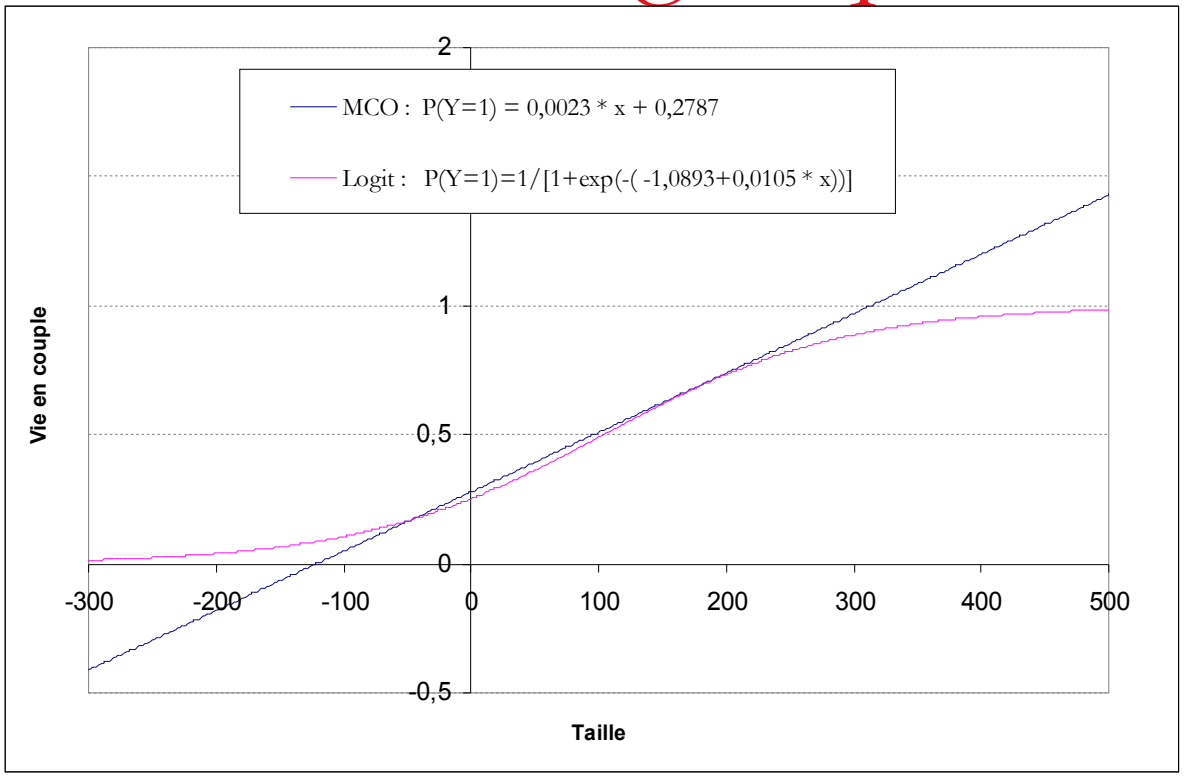
MCO et régression logistique

- On peut utiliser les MCO sur des variables qualitatives
- Exemple : le rôle de la taille sur le fait de vivre en couple
- Mais on risque de prédire des probabilités <0 ou >1
- Par exemple dans l'équation, un homme de 3m20 a une probabilité de 1,01 d'être en couple…
- Même si un homme de 3m20 n'existe pas (encore), calculer une probabilité > 1 est pour certains un grand scandale intellectuel.



La solution: une fonction logistique

- On estime non pas une droite linéaire



- … mais une courbe bornée entre 0 et 1 : c'est la fonction logistique

$$P(y_i=1)=1/[1+\exp[-(b_0+b_1X_1+\ldots+b_nX_n+u_i)]]$$

	Paramètr	Écart-type
Constante	2,11***	0,35
Taille		
Grande	0,09	0,15
Moyenne	Réf.	
Petite	-0,55***	0,17

La vie en couple : taille des hommes et autres facteurs sociodémographiques

	Paramètre estimé	Écart-type
Constante	2,11***	0,35
Taille		
Grande	0,09	0,15
Moyenne	*Réf.*	
Petite	- 0,55***	0,17
Corpulence		
Normale	- 0,30**	0,13
Surpoids	*Réf.*	
Âge de la personne		
20 à 29 ans	- 0,86***	0,18
30 à 39 ans	0,16	0,16
40 à 49 ans	*Réf.*	
50 à 59 ans	0,17	0,18
60 à 69 ans	- 0,24	0,23
Région habitée		
Région parisienne	- 0,28	0,22
Bassin parisien	*Réf.*	
Nord	0,56*	0,31
Est	0,01	0,24
Ouest	- 0,10	0,22
Sud-Ouest	- 0,07	0,24
Centre-Est	- 0,23	0,23
Méditerranée	- 0,46*	0,25
Commune de résidence		
Unité urbaine de 100 000 habitants et plus	- 0,39***	0,14
Niveau scolaire		
Sans diplôme	0,09	0,27
Primaire/secondaire ou technique	0,01	0,23
Primaire/secondaire et technique	0,15	0,23
Premier cycle universitaire	*Réf.*	
2e et 3e cycles universitaires	- 0,21	0,31
Grandes écoles	- 0,45	0,36
Profession de la personne		
Agriculteur, artisan, commerçant	0,08	0,25
Chef d'entreprise, profession libérale	0,95*	0,58
Cadre de la fonction publique, professeur	- 0,17	0,65
Cadre du privé et profession information, spectacle	- 0,63	0,62
Ingénieur	- 0,76	0,64
Profession intermédiaire	*Réf.*	
Employé	- 0,72***	0,21
Ouvrier	- 0,24	0,18
Nationalité		
Français né en France	- 0,61***	0,21
Situation d'activité		
Au chômage	- 0,81***	0,23

Lire les coefficients d'une régression logistique

- Passer à une notion de rapport de chance avec l'odds ratio
  - Calculer exp(b).
  - Ceci nous explique de combien on multiplie/divise la probabilité de base de y pour une variation d'une unité de x
  - Exemple grande taille : b=0.09 ; exp(0.09)=1.09. Être grand multiplie la chance d'être en couple de 1.09 par rapport à une taille normale
  - Exemple petite taille : b=-0.55 ; exp(-0.55)=0.58. Être petit multiplie la chance d'être en couple de 0.58 par rapport à une taille normale.

Retrouver la différence de probabilité à partir du score

- Méthode approximative :
  - On divise par 4 le coefficient : -0.55/4 = -14 %

- Méthode classique : variation d'une unité autour de la situation de référence.
  - Probabilité d'être en couple pour un individu de la situation de référence :

    b_0=Constante = 2,11

    $P(y_i=1 \mid réf)=1/[1+exp[-(b_0)]]=1/[1+exp[-(2,11)]]=89\%$
  - Probabilité d'être en couple quand on est petit (par rapport à la situation de référence :

    b_1=-0,55

    $P(y_i=1 \mid X_1)=1/[1+exp[-(b_0+b_1X_1)]]$

    $=1/[1+exp[-(2,11-0,55)]]=82\%$
  - Effet marginal :

    $\Delta P=P(y_i=1 \mid X_1) - P(y_i=1 \mid réf)$

    82%-89%=-7%
  - Le fait d'être petit diminue toutes choses égales par ailleurs de 7% la probabilité d'être en couple

Régression logistique / MCO Qu'est-ce que ça change ?

- L'estimation : un calcul plus compliqué
  - On n'utilise pas des méthodes matricielles et géométriques de projection mais la méthode du maximum de vraisemblance
  - Pas de solution analytique simple → recours à un algorithme.
- La lecture : des coefficients plus « abstraits »
  - On peut les transformer en pourcentage marginal
  - … Mais on ne peut pas additionner directement les pourcentages marginaux
- Les indicateurs de qualité du modèle moins consensuels
  - On ne peut pas calculer un R2
  - Plusieurs notions alternatives, comme le pseudo-R2 ou D2 de Sommer, etc. Ne font pas l'unanimité !
- Le test est un test de chi2 et non de student, mais la p-value se lit exactement pareil
- Lecture similaire des coefficients et des erreurs-types

Rappel : Que lire dans une régression ?

- Le R2
  - Capacité du modèle à reproduire la réalité
  - Croît avec le nombre de variables
  - Pas d'équivalent pour les régressions logistiques
- Les paramètres *b*
  - Dépendent de l'échelle de mesure
  - Possibilité de « standardiser », mesurer en écart-type (valable surtout pour les phénomènes normaux. Discutable dans les autres cas (var. qual.))
  - Dépendent (variables qualitatives) de la situation de référence

Que lire dans une régression ?

- Les significativités
  - Mesurent les degrés de crédibilité d'un effet et non leur intensité
  - Sensibles à l'autocorrélation
  - Sensibles au choix de la situation de référence
  - Sensibles à l'effectif
  - Dépendent du nombre de variables pour décrire le phénomène

Appendices R

Sous R

```
a<-lm(expliquee~explicat1+explicat2, data=baz)
summary(a)
```

Ou éventuellement

```
a<-lm(baz$expliquee~baz$explicat1+baz$explicat2)
summary(a)
```