Lecture 1. Remembering regression

Olivier Godechot

Sciences Po

M2. Sociology Master

History: From regression towards mediocrity to regression

- Galton, Francis. 1886. "Regression towards mediocrity in heredity stature", *The Journal of the Anthropological Institute of Great Britain and Ireland* Vol. 15 (1886), pp. 246-263.

- Galton (Darwin's cousin). Linking the height of descendants to that of ascendants (beans first, than men).

- Plots a regression line in order to represent the link and finds :

(Descendant's height-Mean descendents height) =2/3 * (Ascendant's height-Mean ascendants height)

- This line is called : regression towards mediocrity line.

ANTHROPOLOGICAL MISCELLANEA.

_____

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

An inferential method

- <u>Dependent variable</u> : An interest variable that we try to
  - Explain
  - Predict
  - Also called "explained variable"
- With the values of other variables : the <u>independent variables</u>
  - Also called "explanatory variables"

Tableau B
Les facteurs socioéconomiques de la taille

	Les hommes		Les femmes	
	Paramètre	Écart-type	Paramètre	Écart-type
Constante	181,9***	0,74	168,0***	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	- 1,10*	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	0,60	- 0,89	0,54
Ouest	- 2,21***	0,51	- 2,89***	0,45
Sud-Ouest	- 1,74***	0,55	- 2,60***	0,49
Centre-Est	- 1,65***	0,54	- 1,93***	0,49
Nord	*Réf.*		*Réf.*	
Profession de la personne				
Agriculteur	2,26***	0,54	0,79	0,53
Artisan, commerçant, entrepreneur	2,16***	0,45	1,60***	0,53
Cadre, profession libérale, prof. intell. supérieure	2,67***	0,40	2,35***	0,50
Profession intermédiaire	2,01***	0,33	1,38***	0,38
Employé	1,72***	0,41	1,08***	0,31
Ouvrier	*Réf.*		*Réf.*	
Profession du père				
Agriculteur	- 0,17	0,37	0,55*	0,32
Artisan, commerçant, entrepreneur	0,49	0,40	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	1,10***	0,40	0,51	0,35
Ouvrier	*Réf.*		*Réf.*	
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	0,40
20, 21 ou 22 ans	- 0,35	0,48	- 0,81**	0,39
23 ans et plus	*Réf.*		*Réf.*	

Linear regression with one linear variable

- Example: we try to explain height among adult males with age (Herpin, 2003)
  - The dependent variable, height, is a numeric/continuous/quantitative variable
  - The independent variable, age is also a continuous variable.
- We estimate the following linear relationship:

  height= a + b. Age + error

  Most common presentations:

  $$y_i = a + b.x_i + u_i$$

  or $y = a + b.x + u$

- We try to find the a and b that limit the most errors.

Ordinary least squares principles (OLS)

- In order to find *a* and *b*, we minimize the sum of the squared errors (i.e. "least square"), that is the sum of the squared deviation of predicted height from observed height.
- Analytical solutions when we have only one independent variable:

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

  - *b*, it's the mean variation of *y* (height) when *x* (age) moves by one unit it's the mean slope
  - *a*, it's the constant or intercept. It's the height when *x*=0 (although we can calculate it, it may not correspond to a realistic situation)

- Ex : *a*= 180.89   *b*= -0.136

More on OLS parameters

- Parameters *a* and *b* are « means ».
    - *a* is the mean value of *y* when *x=0*
    - *b* is the mean slope
    - As for any mean, we can compute the mean's standard deviation, which is the standard error.
- Parameter standard error (in the one independent variable regression)

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

With s, residuals standard deviation,

$$s = \sqrt{\frac{\sum_i \hat{u}^2}{n-2}}$$

that is the root square of the mean square residuals

Graphical presentation

Males over 30 height explained by age

```
Call:
lm(formula = height ~ AGE, data = h30)

Residuals:
    Min        1Q    Median       3Q      Max
-22.2028   -4.3392   -0.2937    4.3426   23.7517

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 180.88469    0.56798   318.5   <2e-16 ***
AGE          -0.13636    0.01026   -13.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.72 on 1815 degrees of freedom
  (1965 observations deleted due to missingness)
Multiple R-squared:  0.08876, Adjusted R-squared:  0.08825
F-statistic: 176.8 on 1 and 1815 DF,  p-value: < 2.2e-16
```

An indicator of model's quality: R²

- Basic variance equation with OLS regressions

Variance of y (dependent variable)
$\qquad$ = Variance of $\hat{y}$ (prediction of y based on x) + Variance of u (error)

"Total variance"= "explained variance" + "residual variance"

- R² or « the share of variance explained »
  - R²=explained variance / total variance
  - Example: R²=8,8%

Parameters nullity test

- Is the slope b significant? Can we reasonably believe in the fact that the slope is negative… => Parameter test
- We need one supplementary hypothesis: We suppose that the error *u* follows a normal distribution
- Many possible tests. The most common is two-tailed nullity Student t-test
  - H_0 : *b*=0
  - We try to "nullify" this hypothesis
    - Showing that the probability of a deviation to H0 as large as the empirical one measured is very small under H0 parameter.
  - *(b/ standard error(b))* follows a Student's t-test with *n-1-k* degrees of freedom (where *k* is the number of independent variables).
- Ex : b= -0.13 σ_b=0.01026   T= -0.13 /0.01026=-13.30

Prob(stud > |T|)=2,93*10^{-39}

Two variables regression: graphical representation

- $y = b_0 + b_1.x_1 + b_2.x_2 + u$
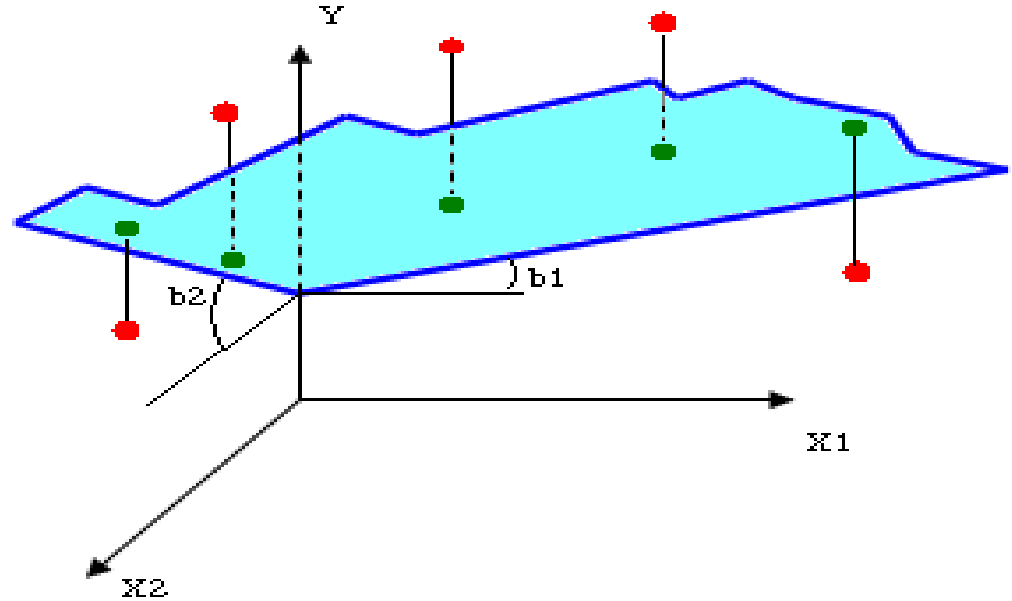
Multiple regression:
$$y = b_0 + b_1.x_1 + b_2.x_2 + \ldots + b_k.x_k + u$$

- We project the y on a plane/or a surface with k dimensions,
  - In a direction parallel to the y axis (or in a direction orthogonal to the x_k plane)
  - Along this direction we try to find the plane that minimize the squared deviation of the observation y from its projection $\hat{y}$.
- Matrix formula : $\widetilde{b} = (X'X)^{-1}.X'.y$
- $V(\widetilde{b}) = \widetilde{\sigma}^2. (X'.X)^{-1}$

Geometrical method or probabilistic method

- OLS: a geometrical approach of regression.
- Alternative method: Maximum likelihood
  - Likelihood is the product of density functions
  $$\mathcal{L}(\theta\,;\,x_1,\ldots,x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$
  - We look for the parameter θ that maximizes this product.
- Very scary and complex. What should we take away ?
  - In the linear case, there's an equivalence between OLS and maximum likelihood. (So let's go for OLS)
  - In some non-linear cases (logit, probit), no geometrical solution => maximum likelihood (similar to OLS)

Interpretation: "everything being equal"

- b_j slope of the best plane in order to predict y.
- If the j independent variable in the regression increases by one unit while other variables are remaining constant, then the predicted variation of y is equal to b_j .
- Linearity: we add the effects of each independent variables
  - Ex : one effect for age, one for corpulence, etc..
  - A simultaneous variation in age by one unit and in corpulence by one unit is equal to the sum of the parameters of age and corpulence (b1+b2)

"Everything being equal" avoids the confounding variable problem

- When we measure the link between two variables, there might be a confounding variable
  - The independent variable captures the effect of another more relevant variable to which it is correlated
  - Example : Age and corpulence are very much correlated
  - When we measure the impact of corpulence alone on height, it is strongly significant not because of corpulence but because of age to which it is correlated
- Regression enables to handle this problem …
  - When we introduce the two variables, we are able to separate the role of the two factors. There's « enough » observations where ages are equal and corpulence different and where corpulence are equal and age different in order to know what is due to each of the two variables.
  - Reading: « when age is constant (or controlled), corpulence adds this ». « Corpulence being constant, age adds this ».

Example: age and corpulence on height

- Age and corpulence are correlated :
  r= 0.15327 (p<.0001)

- Corpulence alone has a negative effect on height:

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.12183    1.18385    148.77  <.0001 ***
BMI          -0.09597    0.04559     -2.11  0.0354 *
```

- But is this an effect of corpulence or an effect of age to which it is tied ?

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.086091   1.194750   151.57   <2e-16 ***
BMI          -0.007964   0.044122    -0.18    0.857
AGE          -0.136030   0.010452   -13.01   <2e-16 ***
```

- Multivariate regression shows that this is due to the effect of age

Categorical independent variables

- How can we introduce categorical independent variables?
- Solution we transform each category of the categorical variable into a dichotomous one. We introduce all those dichotomous variables in the regression EXCEPT ONE.
- The omitted category in regression is the <u>reference category</u>.
  - The marginal effect of the reference category is 0.
  - In a model with only categorical variables, prediction for people in the reference category is the model's constant
- Why do we omit a category?
  - Technical answer, variables would be tied.
    Ex : Male + Female = Constant
  - Intuitive answer : it is like measuring the gaps between the rungs of a ladder : you can measure them only by fixing a reference rung.

Graph of a simple regression with just one categorical variable

Height of males depending on the fact they are manager

$y = 4.1566x + 166.86$
$R^2 = 0.0219$

+ Height
— Linéaire (Height)

Height

Belonging to the manager group
1 =yes, 0 = no
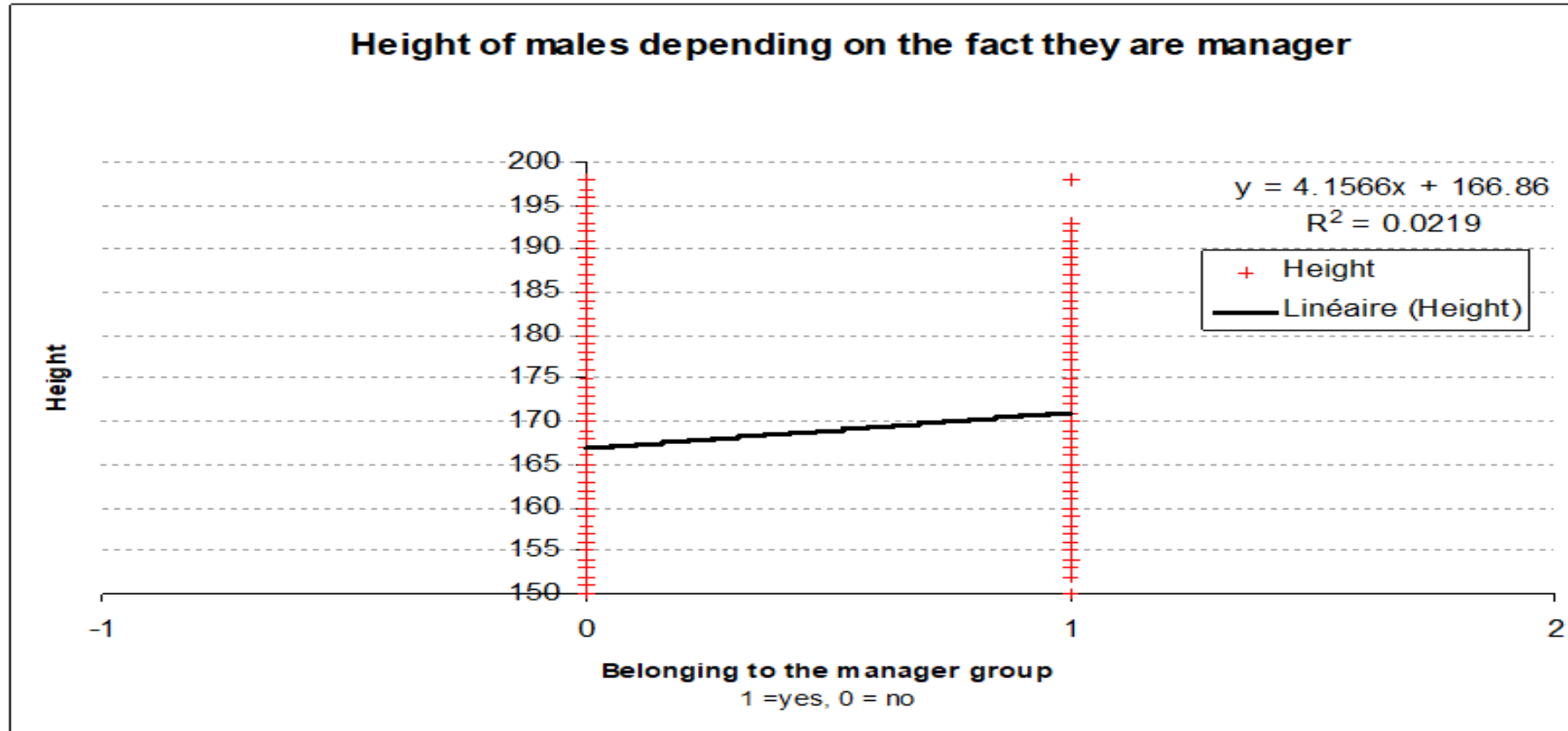
```
Call:lm(formula = tailleb ~ corpulence + AGE + reg + cs6 + agediplo, data = h30)
Residuals:
     Min       1Q   Median       3Q      Max
-21.7206  -4.3031  -0.2273   4.2869  23.7295
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        180.95376    1.58089 114.463  < 2e-16 ***
BMI                  0.03759    0.04447   0.845 0.398155
AGE                 -0.12285    0.01181 -10.405  < 2e-16 ***
reg1_Region Par.    -1.66781    1.01121  -1.649 0.099264 .
reg2_Bassin Par     -0.19911    0.91467  -0.218 0.827701
reg4_Est            -0.30088    0.87387  -0.344 0.730656
reg5_Ouest          -2.17721    0.85792  -2.538 0.011244 *
reg7_Sud-Ouest      -0.82433    0.85286  -0.967 0.333910
reg8_Centre-Est     -1.02188    0.85475  -1.196 0.232042
reg9_Mediterrannee  -0.29088    0.85576  -0.340 0.733965
cs61Agriculteurs     1.19071    0.67426   1.766 0.077581 .
cs62Arti-Comm        1.69546    0.59667   2.842 0.004543 **
cs63Cadres           2.15593    0.63808   3.379 0.000744 ***
cs64Prof. Int        0.40244    0.48029   0.838 0.402193
cs65Employes         0.29687    0.55452   0.535 0.592467
agediplo(12,13]     -2.95758    0.76759  -3.853 0.000121 ***
agediplo(13,15]     -1.21978    0.74028  -1.648 0.099591 .
agediplo(15,17]     -1.77661    0.89606  -1.983 0.047562 *
agediplo(17,19]     -0.61104    0.78522  -0.778 0.436568
agediplo(19,22]     -0.62371    0.76311  -0.817 0.413861
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.543 on 1710 degrees of freedom
  (2052 observations deleted due to missingness)
Multiple R-squared:  0.146, Adjusted R-squared:  0.1365
F-statistic: 15.38 on 19 and 1710 DF,  p-value: < 2.2e-16
```

How to choose the reference category?

- The choice of the reference category of one categorical variable does not modify the estimates for other categorical variables
- But changing the reference category of one categorical variable modifies the estimates for other categories of the same variable
  - Parameter of manager is different if you use intermediates as reference category or unskilled blue collar
- Interpretation may be influenced by the choice of the categorical variable.
  - A category in the middle diminish apparent significance
  - An extreme category increases apparent significance
- In a treated versus control framework: set the control group as the reference category
- Otherwise: using the most common situation might be a good policy= avoid to construct monster as the reference category (ex : unqualified worker with PhDs)

Model with R

a<-**lm(**dependent**~**indep1**+**indep2**, data=**mydb**)**

summary(a**)**

Or eventually

a<-lm(mydb$dependent~mydb$indep1+mydb$indep2)

summary(a)

lm() stands for linear model

Object a contains many things not displayed in the summary, predictions, residuals. Look inside

str(a)

Independent can be a quantitative numerical variable or categorical one (factor) : no need to compute the dichotomous variables.

Improve layout

library(texreg)

screenreg(list(m1,m2,m3))

htmlreg(list(m1,m2,m3),
        file="out.html")

	iie10_bb	iie10_nuts1	iiu10_nuts1	iie10_wnuts1
year	0.232^{***}	0.220^{**}	-0.009	0.229^{***}
	(0.059)	(0.070)	(0.009)	(0.061)
Num. obs.	2807	2807	2807	2807
R^2 (full model)	0.202	0.959	0.990	0.919
R^2 (proj model)	0.037	0.372	0.011	0.459
Adj. R^2 (full model)	0.200	0.957	0.989	0.915
Adj. R^2 (proj model)	0.035	0.339	-0.041	0.431
Num. groups: country	8			
Num. groups: paste(nuts1, country)		139	139	139

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Statistical models

Choosing reference category with R

If one introduces a "factor" type categorical variable in the regression with R, R chooses by default the first category (following the variable's categorical order) as the reference category.

- A simple way of changing it : relevel

dat$myvar <- relevel(dat$myvar,ref="my category")

"Everything being equal" isn't everything… Unobserved heterogeneity limit

- "Everything being equal" holds true only for variables used in the model! It is more an "All others variables being controlled"

- What if the suspected confounding variable is missing?

  - True model is $y=a+ b.x+ c.z$ but z is missing.

  - If x is correlated positively with z and y and z is missing and z positively correlated with x.

  - We estimate $y=a'+ b'.x$

  - What's the relation between b and b'?

"Everything being equal" isn't everything…
Unobserved heterogeneity limit

$$b' > b$$

- b' overestimates the true b
- b' captures both the effect of x and that of z

- If $c<0$ than b' underestimates the true b
- Worse, there can be a sign reversal.

Control variables. Virtues and limits

- Risk of unobserved heterogeneity ==> multiply control variables

- If x effect remains robust to introduction of many control variables ==> believe in causal effects.
  - Sequential models strategy

- But risks "Control variables" could be
  - Alternative proxy of x
  - Antecedent
  - Mediator (Mechanism)

- ==> Attenuate the effect of x on y
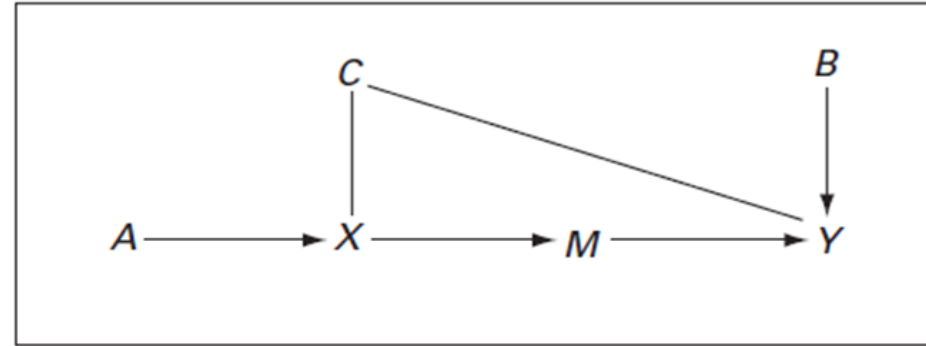
Of theoretical interest
X = Causal factor
Y = Outcome
M = Mechanism

Background factors
A = Antecedent
B = Covariate
C = Confounder

General features
⟶ = Causal relationship
—— = Covariation (possibly causal)

An elaborated causal graph

"Everything being equal" isn't everything… Linearity limit

- We only measure average effects.
- We consider that those average effects can just be added.
- For instance, in previous model, one supplementary year of age is the same for someone whose body mass index is of 20 or of 25.
- Except in alternative specification (see further: next class), we don't look for interactions:
  - If height increases with age for this category but decreases with age for other category, we won't see it. We will only estimate the average age effect.
  - Ex. If height decreases only with age for blue collars, but not for managers, we won't see it in the regression. We will see only average effect on age and on blue collar variable.

What to read in a regression ?

- Quality of the model
  - The R-square
    - Capacity for the model to reproduce reality
    - Increases with the number of variables
      - To avoid this problem: adjusted R-Square
    - R-square might be very high with trivial correlations (ex.: vote and political positioning)
  - F test
    - Tests the global validity of the model (almost always significant)

What to read in a regression ?

- The models parameters b
  - b (or β): the value of your link
    - Sign tells you about the direction of the effect
    - Value tells about the magnitude of an effect
    - Value depends on the scale. If the variable is age is calculated in months instead of years b is divided by 12.
    - For categorical variable, depends on the reference category

What to read in a regression ?

- The models parameters s
  - ($s(b)$ or s or sometimes σ): the precision of your link
    - Indicator of the dispersion of your b parameter
    - Enables to compute Student's t (=b/s), p-value and ***
    - Enables to compute confidence interval of your b parameter
      - The 95% confidence interval of your b in the full population
        [b-1.96*s, b+1.96*s] (when n>30)

What to read in a regression?

- Significance: T-test, p-value and ***
  - R default * for p-values: *** p<0.001, ** p<0.01, * p<0.05, ·p<0.1
  - Most common thresholds:
    - *** p<0.001, ** p<0.01, * p<0.05
    - *** p<0.01, ** p<0.05, * p<0.1
  - Measure how much you can trust an effect
  - Often researchers mainly interested just by ± (parameters' sign) and ***

What to read in a regression?

- Limitation of ****
    - * make a huge qualitative difference between p=0.099 and p=0.101
    - Sensitive to the number of observations. p-values decrease (and significance increase) proportionally to $\sqrt{n}$
    - Sensitive to strong auto-correlation of the independent variables. Might kill significance.
    - Sensitive to the choice of the reference category
    - Depends on the number of variables for describing the phenomenon

What to read in a regression?

- Other possible statistics and figures
- Number of observations (extremely important)
  - Not displayed, but in the R object
  - *DF=n-k-1* where k number of variables (-1 holds for the constant)
- Sum of squares
- Residuals
- Predicted values

Comparing the effect of two independent variables

- Comparing significance
  - Might inform when one variable is significant and the other is not
  - But significance is about how much you can trust there's an impact, not about intensity
- Comparing directly the parameters
  - For instance categorical variable
    - Belonging to this group influences more y than belonging to that one.
  - Continuous variable if scales are comparable
  - But if scales are not comparable

Comparing the effect of two independent variables (2)

- Standardization might be the solution
  - Standardization consists in dividing a variable by its standard deviation
  - Regression can be y or x standardized or both.
  - The new unit is the sd. One sd of x_1 increases y by b_1 sd of y
- Sometimes debatable
  - Magnitude more difficult to grasp
    - Earnings increase by 100 \$ or by 10% or by 0.1 standard deviation
  - Especially for categorical variables
    - "A standard deviation of male" impacts by b * a standard deviation of earning

Example

```
> reg<-lm(tailleb~corpulence+AGE,data=h30)
> summary(reg)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 181.086091   1.194750   151.57   <2e-16 ***
corpulence   -0.007964   0.044122    -0.18    0.857
AGE          -0.136030   0.010452   -13.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> sd(h30$corpulence,na.rm=TRUE)
[1] 3.632575
> sd(h30$AGE,na.rm=TRUE)
[1] 14.69266

> h30$corpulence_std<-h30$corpulence/sd(h30$corpulence,na.rm=TRUE)
> h30$AGE_std<-h30$AGE/sd(h30$AGE,na.rm=TRUE)

> reg<-lm(tailleb~corpulence_std+AGE_std,data=h30)
> summary(reg)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    181.08609   1.19475   151.57   <2e-16 ***
corpulence_std  -0.02893   0.16028    -0.18    0.857
AGE_std         -1.99864   0.15357   -13.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example (Godechot, 2016)

Table 2: Impact of the Finance Share of the GDP on Income Inequality

	A. Classical panel regression models (Equation 1)							
	Gini Index	D5/D1	D9/D1	D9/D5	Top 10% share	Top 1% share	Top 0.1% share	Top 0.01% share
GDP per capita (t-1)	−0.51∗	0.62∗	0.34∗	0.13	−0.21	0.04	−0.02	0.02
Union rate (t-1)	−0.27∗	−0.16∗	−0.23∗	−0.25∗	−0.36∗	−0.23∗	−0.1∗	−0.14∗
Import rate (t-1)	−0.15∗	0.41∗	0.17	−0.03	−0.11∗	−0.13∗	−0.15∗	0.17
Finance & insurance/ GDP (t-1)	−0.04	−0.04	0.16∗	0.18∗	0.12∗	0.23∗	0.28∗	0.41∗
Adj. within R2	0.150	0.081	0.086	0.152	0.174	0.147	0.127	0.229
Nb. obs./ countries/ years	673/18/42	391/18/42	391/18/42	391/18/42	604/18/42	623/18/42	538/17/42	368/14/42

Regression hypotheses, limits and solutions

OLS 6 hypotheses

- 1. Linearity
- 2. Full rank Matrix and absence of auto-correlation between independent variables
- 3. Homeoskedasticity
- 4. Absence of auto-correlation of residuals
- 5. Gaussian normal residuals
- 6. Absence of correlation between independent variables and the residual *in the theoretical model.*

1. Linearity

- Is the relation between the variables really the following form ?

  $y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \ldots + \beta_k.x_k + u$

- But linear models are very flexible
  - Transforming dependent variable
    - Log, Asinh, etc.
  - Transforming independent variable
    - Log, Square, Cubic, etc.
- Prediction of y* can go beyond y variations
  - y is a proportion
  - y is a duration
  - y is a grade
  - y is a count variable
- Linear model is not necessarily the most adapted

Handling non linearity

- Introduce a quadratic function
  - Ex: $height = b_0 + b_1.age + b_2.age^2 + \ldots + u$
  - Optimum of age: $-b_1/(2.b_2)$
- Introduce interaction
  - Ex : $height = b_0 + b_1.age + b_2.manager + b_3.manager*age + \ldots + u$
- Use a log dependent variable transforms a linear model into a multiplicative one
  - Ex: $log(height) = b_0 + b_1.age + b_2.age^2 + \ldots + u$

    $\Leftrightarrow height = exp(b_0 + b_1.age + b_2.age^2 + \ldots + u)$

    $\Leftrightarrow height = exp(b_0)*exp(b_1.age)*exp(b_2.age^2)*\ldots* exp(u)$

Reading interactions

- Marginal effects reading
  - Main effect manager : + 1.4* cm
  - Main effect master : + 2.1* cm
  - Interacted effect : + 0.5 cm
    - Shows whether there's a significant interaction effect
- Reconstitute all the cases
  - Non manager w/o master 173.1 cm
  - Non manager with master 173.1+2.1=175.2 cm
  - Manager w/o master: 173.1+1.4=174.5
  - Manager with master: 173.1+1.4+2.1+0.5=177.1 cm

```
============================
                   Model 1
----------------------------
(Intercept)      173.05 ***
                   (0.18)
manager            1.40 *
                   (0.62)
master             2.09 *
                   (0.90)
manager:master     0.51
                   (1.20)
----------------------------
R^2                0.03
Adj. R^2           0.03
Num. obs.          1813
============================
*** p < 0.001; ** p < 0.01; * p <
0.05
```

2. Full rank matrix and absence of autocorrelation variables

- Full rank matrice:
  - An independent variable x_k cannot be a linear combination of other independent variables (including intercept)
  - $x_3 = 3 * x_2 + 2 * x_1 \Rightarrow$ We can't estimate x_3.
  - Reason for the exclusion of the reference category for qualitative variables
    Female = 1 – Male = Intercept – Male
  - The software will not estimate the variable and will tell it's a linear combination
- Variables autocorrelation
  - We are getting close to linear combination without being there fait
  - Parameter instability

Handling multicolinearity

- Multicolinearity of independent variables (strong : $|r|>0,9$)
  - VIF : Variance inflation factor>10
  - Parameter instability (especially standard-errors).

- Solution :
  - Introduce only one of the two correlated variables
  - Use an index
  - Conduct a PCA and use the first axis.

3. Homeoskedasticity and 4. absence residual autocorrelation

- Homeoskeadastiticite.
  - All errors u_i have the same variance σ^2.
  - $\mathrm{Var}[u_i \,|\, x_1,\ldots, x_k] = \sigma^2$
- Absence of residuals auto-correlation
  - Residuals u_i are not correlated with residuals u_k
  - $\mathrm{Cov}[u_i, u_k \,|\, x_1,\ldots, x_k] = 0$
- Combination of the two criteria
  - $\mathrm{Var}[u] = \sigma^2.I$

Handling heteroskedasticity and residual autocorrelation

- Heteroskedasticity (unstability of residual of variance)
  - Unstability of parameters
- Solution:
  - Change the definition of the dependent variable: turn into log…
  - Other methods of estimation: Generalized least squares, Weighted least squares
  - Robust standard errors
- Autocorrelation of residuals
  - => Bad estimation of parameters standard errors
- Solution
  - Cluster robust standard errors

5. Gaussian normal residuals

- Residual u_i follows a normal probability law with a null average and σ^2 variance
- This hypothesis is not compulsory but enables to calculate confidence intervals and to test parameters.

  $y = b_0 + b_1.x_1 + u$

  $y^* = b_0 + b_1.x1 = (y - u)$

  $b_1 = (y - b_0 - u)/x_1$

  If u is normal, the prediction y* and the parameter b_1 also follows a normal law.

  Hence, we can calculate parameters for y* and the parameter b_1 thanks to the properties of Gaussian law.

- But the normal law might not be the best fit…

6. Exogeneity of independent variables

- Independent variables are not tied to the residuals (in the theoretical model).
- $E(u_i \mid x_1, \ldots, x_k) = 0$ ou $Cov(x_k, u_i) = 0$
- If that's not the case, we speak of endogeneity
- => Lecture on instrumental variables

OLS and logistic regressions

OLS and logistic regressions

- We can use OLS for qualitative variables

- Example : the role of size on the fact of living in couple

- Bur risk of predicting probabilities <0 or >1

- For example in the equation, a man of 3m20 has a probability of 1,01 of being in a couple…

- While man of 3m20 don't exist (yet), calculating a probability > 1 is for some a great outrage.
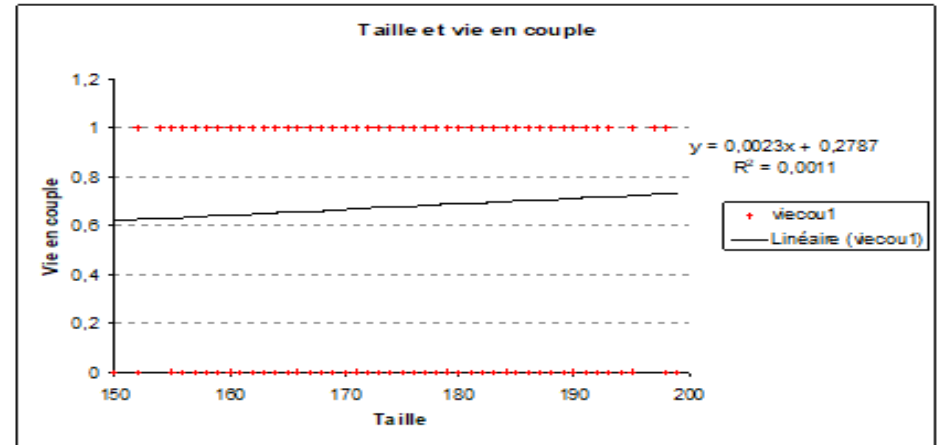
Solution: use the logistic function

- We don't estimate a straight line
- … but a curve bound between 0 and 1 : it's the logistic function
- $P_{(y_i=1)}=1/[1+\exp[-(b_0+b_1X_1+\ldots +b_nX_n + u_i)]]$
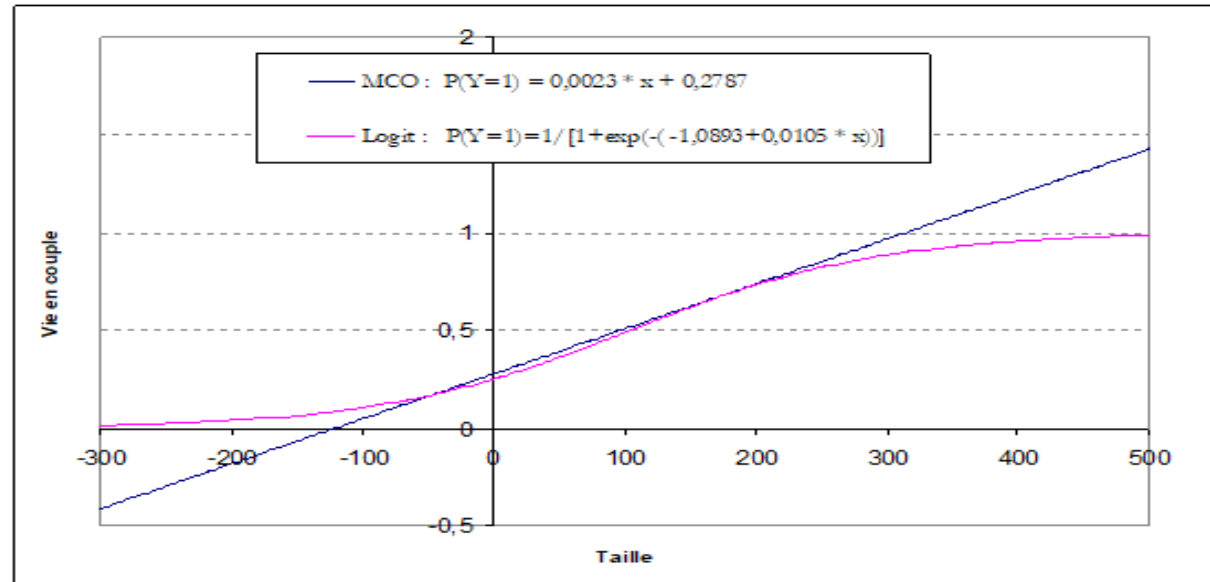
La vie en couple : taille des hommes et autres facteurs sociodémographiques

	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

Living with a partner: men's height and other socio-demographic factors

	Estimated parameter	Standard deviation
Constant	2,11***	0,35
Size		
Great	0,09	0,15
Average	*Ref.*	
Small	- 0,55***	0,17
Normal **weight**	- 0,30**	0,13
Overweight	*Ref.*	
Age of the person		
20 to 29 years old	- 0,86***	0,18
30 to 39 years old	0,16	0,16
40 to 49 years old	*Ref.*	
50 to 59 years old	0,17	0,18
60 to 69 years old	- 0,24	0,23
Inhabited area		
Paris area	- 0,28	0,22
Parisian basin	*Ref.*	
North	0,56*	0,31
East	0,01	0,24
West	- 0,10	0,22
Southwest	- 0,07	0,24
Central East	- 0,23	0,23
Mediterranean	- 0,46*	0,25
Municipality of residence		
Urban unit of 100,000 inhabitants or more	- 0,39***	0,14
School level		
No diploma	0,09	0,27
Primary/secondary or technical	0,01	0,23
Primary/secondary and technical	0,15	0,23
Undergraduate	*Ref.*	
Graduate and post-graduate	- 0,21	0,31
Great schools	- 0,45	0,36
Profession of the person		
Farmer, craftsman, shopkeeper	0,08	0,25
Company manager, liberal profession	0,95*	0,58
Public service executive, professor	- 0,17	0,65
Private sector executives and information and entertainment professionals	- 0,63	0,62
Engineer	- 0,76	0,64
Intermediate profession	*Ref.*	
Employee	- 0,72***	0,21
Worker	- 0,24	0,18
Nationality		
French born in France	- 0,61***	0,21
Activity status		
Unemployed	- 0,81***	0,23

With R

```
mod <-  glm(dependent~explicat1+explicat2,
            data=baz,
            family = binomial)
summary(mod)


screenreg(mod)
```

Finding the probability from the score

- Classical method: one unit variation from the reference situation.
  - probability of being in couple for an individual in reference group:
    - b_0=Intercept = 2,11
    - $P(y_i=1 \mid ref)=1/[1+\exp[-(b_0)]]=1/[1+\exp[-(2,11)]]=89\%$
  - probability of being in couple when one is short (in contrast to the reference group):
    - b_1=-0,55
    - $P(y_i=1 \mid X_1)=1/[1+\exp[-(b_0+b_1X_1)]]$
    - $=1/[1+\exp[-(2,11-0,55)]]=82\%$
  - Marginal effect:
    - $\Delta P=P(y_i=1 \mid X_1) - P(y_i=1 \mid ref)$
    - 82%-89%=-7%
  - The fact of being short decreases by 7 percentage point the probability of being in couple.

Logistic regression vs OLS
What's the difference?

- Estimation: a little more complex
  - Rather than geometrical matrix projection, using maximum likelihood methods
  - No simple analytical solution → use of an algorithm.
- Reading: coefficients more « abstract »
  - One can compute marginal percentage
  - … But you cannot add up different marginal percentages
- Regression quality indicators are less consensual
  - Can't compute R2
  - Several alternative notions, like pseudo-R2 or Sommer's D2 , etc. They are not unanimous !
- Test is a chi2 test rather than a student one, but you can read p-value exactly the same
- Similar reading of coefficients and standard errors.

Regression logistic vs OLS : the debate

- Two limits with logistic regressions:
  - The parameters are not comparable from one regression to another
    - Mood, 2010, "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It", *Eur. Soc. Rev.*
    - Solution: transforming parameters. Two available techniques
      - *Y standardization*
      - *Average marginal effect*
  - Marginal effects cannot be calculated directly when there's some interactions
    - Ai and al., 2003, « Interaction terms in logit and probit models », *Ec. Let.*

- The return of linear probability models(OLS)
  - Need for correcting for heteroskedasticity with robust standard-errors.

Y standardization

- Residual default variance in logistic regression is constant and equal to $(\pi^2)/3$ (=3.29)
- Formula
  - Divide parameters (coefficient) with a standard deviation parameter which is the sum of 1) standard-deviation of logit predictions and 2) assumed standard deviation of logistic regression
  - $sd(Y^*) = sd(\Sigma_i b_0 + b_1 x_1 + \ldots + b_k x_k) + \sqrt{\pi^2/3}$
- Sous R

      logit<-glm(expliquee~explicat1+explicat2, data=baz,family = binomial)
      sdYstar<-sd(logit$linear.predictors)+((pi**2)/3)**0.5
      logit$coefficients/ sdYstar

- Simulations does not seem very robust solutions. Precise meaning of parameters less clear.

Average marginal effect

- We calculate instead the average marginal effect, i.e. the mean (all along x distribution) of an epsilon variation.
- Formula
  - $b_{AME} = \Sigma_i \, bf(b.x_i) \, / \, n$
    - Where f is the density function of the logistic law and F its cumulative distributive function
    - $$f(b.x_i) = exp(b.x_i) / [1 + exp(b.x_i)]^2$$
    $$= [1/[1 + exp(-b.x_i)]].[1 - [1/[1 + exp(-b.x_i)]]]$$
    $$= F(b.x_i)*[1 - F(b.x_i)]$$
- Sous R

      logit<-glm(expliquee~explicat1+explicat2, data=baz,family = binomial)
      AME = logit$coefficient*((1/length(logit$fitted.values))*sum(logit$fitted.values*(1-logit$fitted.values)))
      AME

- Result is robust and very close from that obtained with OLS. You can interpret directly as difference in probability

Other functional forms

Non-linear models

- Dependent variable
  - Binary => Logistic / probit regressions
  - Ordered variable => Logistic / probit ordered regressions
  - Categorical => Multinomial regressions
  - Duration => Duration models: Cox, Weibull, etc.
  - Count variable => Poisson models
  - Non Gaussian linear form (like wage, wealth) => log-normal regressions

Ordered logistic regression

- *« Given the work you furnish, would you say that you are : 1. Very well payed, 2. Rather well payed, 3. Normally payed, 4. Rather badly payed, 5. Very badly payed? » (Exemple : Godechot, Gurgand, 2000)*

```
Ordered Value   _PAYE      Count
        1   1         30
        2   2        240
        3   3       1166
        4   4        844
        5   5        196
```

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Variable Estimate
INTERCP1	1	-1.6842	0.6215	7.3425	0.0067	.	
INTERCP2	1	-0.5761	0.6184	0.8679	0.3515	.	
INTERCP3	1	1.0095	0.6184	2.6649	0.1026	.	
INTERCP4	1	2.3776	0.6197	14.7203	0.0001	.	
SEXE1	1	0.2385	0.8832	0.0729	0.7872	0.119217	

Intercept 1: Very well rather than well, normally, badly payed and very badly payed
Intercept 2: Very well, well, rather than normally, badly payed and very badly payed
Intercept 3: Very well, well, normally, rather than badly payed and very badly payed
Intercept 4: Very well, well, normally, badly payed rather than very badly payed

Ordered Regression with R

- Function polr in MASS library
  - library(MASS)
  - reg <- polr(Sat ~ Sex + Type + Cont, method="logistic", data=aa)

Multinomial regression / Coulangeon (2003)

TABLEAU IVb. — *Estimation des paramètres du modèle logit multinomial – probabilité d'appartenir aux classes I, II, III et V (modèle avec effets d'interaction)*

Modalité de référence **Classe IV**	Modalité active (Modalité active)	Classe V Coeffi-cient	p	Effet marginal	Classe III Coeffi-cient	p	Effet marginal	Classe II Coeffi-cient	p	Effet marginal	Classe I Coeffi-cient	p	Effet marginal
Modalité de référence	Modalité active	Coeffi-cient	p	Effet marginal	Coeffi-cient	p	Effet marginal	Coeffi-cient	p	Effet marginal	Coeffi-cient	p	Effet marginal
Constante		−1,459		18,9%	−2,614		6,8%	−1,423		19,4%	−1,443		19,1%
Sexe *femme*	*homme*	0,183	n.s.		0,485	<,001	+3,8%	−0,066	n.s.		−0,224	<,05	-3,2%
Âge [(1)]		0,100	<,001	+1,6%	−0,025	n.s.		0,065	<,001	+1,0%	0,072	<,001	+1,1%
Diplôme *< bac*	*bac ou plus*	0,167	n.s.		0,353	n.s.		0,116	n.s.		0,883	<,001	+17,3%
Âge×diplôme *< bac*	*Bac ou plus*	0,005	n.s.		0,023	n.s.		0,000	n.s.		0,027	<,01	+0,4%
Pcs *employé*	*agriculteur*	0,466	n.s.		−1,676	n.s.		0,279	n.s.		−0,452	n.s.	
	patron de l'industrie et du commerce	0,333	n.s.		0,473	n.s.		−0,126	n.s.		−0,013	n.s.	
	cadre supérieur	−0,552	n.s.		0,729	n.s.		0,400	n.s.		1,031	<,001	+20,7%
	profession intermédiaire	−0,880	<,05	−10,1%	0,318	n.s.		0,260	n.s.		0,194	n.s.	
	ouvrier	0,027	n.s.		−0,729	n.s.		0,053	n.s.		−0,378	n.s.	
	étudiant	−3,578	n.s.		0,479	n.s.		2,584	n.s.		0,784	n.s.	
	retraité	−0,174	n.s.		0,008	n.s.		0,825	<,05	+16,1%	0,729	<,05	+13,8%
	autre inactif	0,060	n.s.		−0,360	n.s.		−0,028	n.s.		−0,205	n.s.	
Âge×Pcs *employé*	*agriculteur*	0,041	n.s.		−0,085	n.s.		0,050	n.s.		0,050	n.s.	
	patron de l'industrie et du commerce	−0,033	n.s.		0,015	n.s.		−0,045	n.s.		−0,018	n.s.	
	cadre supérieur	0,045	n.s.		0,018	n.s.		0,054	n.s.		−0,028	n.s.	
	profession intermédiaire	−0,046	n.s.		0,008	n.s.		0,015	n.s.		−0,006	n.s.	
	ouvrier	−0,037	n.s.		−0,044	<,05	−0,3%	0,004	n.s.		0,014	n.s.	
	étudiant	−0,190	n.s.		−0,027	n.s.		0,089	n.s.		0,006	n.s.	
	retraité	−0,015	n.s.		−0,060	n.s.		−0,049	<,01	−0,8%	−0,061	<,01	−0,9%
	autre inactif	−0,011	n.s.		−0,020	n.s.		−0,025	n.s.		−0,042	<,01	−0,6%
Revenu *10 000 francs*	*> 10 000 francs*	−0,526	<,001	−6,8%	0,018	n.s.		0,184	n.s.		0,561	<,001	+10,2%
	nsp	0,045	n.s.		0,443	<,02	+3,4%	0,247	n.s.		0,403	<,01	+7,0%
Origine sociale *populaire*	*supérieure*	0,069	n.s.		0,293	n.s.		−0,081	n.s.		0,587	<,001	+10,7%
Compétence musicale *non-musicien*	*apprentissage musical*	-0,169	n.s.		0,434	[...5]	+3,3%	0,455	<,05	+8,1%	0,740	<,001	+14,0%
	auto-apprentissage	-0,140	n.s.		[...]		[...]%	0,306	n.s.		0,594	<,001	+10,8%

- 2 Log L :
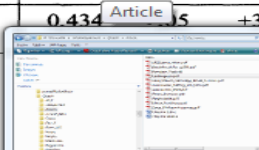 Modèle sans interactions : 9 757 (ddl : 64)
  Modèle avec interactions : 9 664 (ddl :100)
  diff. ddl: 36

Multinomial Regression

- One compares the probability of being in a category rather than being in the reference category.

- One variable's effect (gender, age…) differs depending on the category on which they apply

- With R : function multinom

    regm <- multinom(class~sexe+diplom, data=aa)

Duration model

- Examples
  - Godechot, Louvet (2010). Time for PhD to become PhD adviser.
  - Baudelot et al. (2016). Time for people to get kidney transplants
- Problem
  - Problem: for some people we can observe the whole process.
  - For others we can't
  - Censorship effect: Can we observe the event? Yes, No.
  - If no, this does not mean the event will not happen.

Figure 3. Waiting time before obtaining a kidney transplant for the whole ESRD population, by level of education

Percentage still in dialysis

INED
029A16

Legend:
- Primary
- Lower sec.
- Upper sec.
- Higher ed. ≤ 3 years
- Higher ed. > 3 years

Time (years)
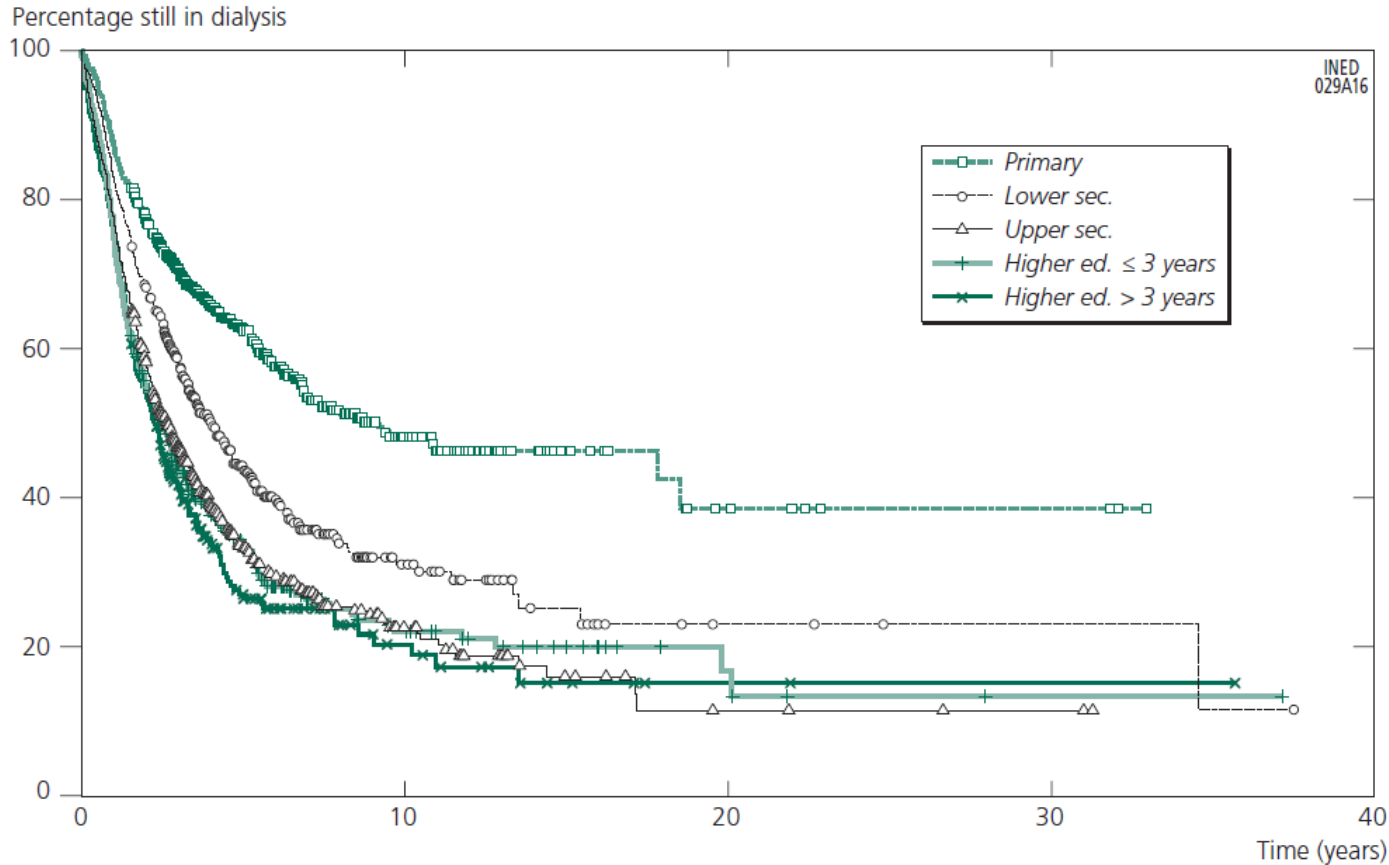
Interpretation: These Kaplan-Meier curves show the proportion of the ESRD population still in dialysis that has not yet received a kidney transplant. Five years after beginning dialysis, 62% of patients with primary education are still in dialysis.

Source: Quavi-REIN survey.

Table 2. Odds of being a transplant recipient and being registered on the waiting list

	Descriptive statistics (%)	Odds of being a transplant recipient				Odds of being on the waiting list				Odds of being a transplant recipient / on the waiting list			
		Crude prop. (%)	Crude OR	OR (log. reg.)	OR (duration)	Crude prop. (%)	Crude OR	OR (log. reg.)	OR (duration)	Crude prop. (%)	Crude OR	OR (log. reg.)	OR (duration)
	1	2	3	4	5	6	7	8	9	10	11	12	13
Overall or Constant	100	59		6.15		61		26.71*		89		3.4	
Level of education													
Primary	29	40	Ref.	Ref.	Ref.	42	Ref.	Ref.	Ref.	89	Ref.	Ref.	Ref.
Lower secondary	21	58	2.09***	1.20	1.14	63	2.34***	1.43*	1.29**	87	0.81	0.76	0.90
Upper secondary	23	68	3.22***	1.59**	1.43***	71	3.34***	1.81**	1.50***	90	1.17	0.98	1.03
Higher ed. ≤3 years	16	70	3.42***	1.37·	1.36***	75	4.09***	1.70**	1.35***	88	0.91	0.81	0.95
Higher ed. > 3 years	11	72	3.76***	1.91**	1.63***	70	3.17***	1.70*	1.61***	95	2.42*	1.76	1.18
Sex													
Male	63	57	Ref.	Ref.	Ref.	60	Ref.	Ref.	Ref.	88	Ref.	Ref.	Ref.
Female	37	61	1.18·	1.07	1.03	63	1.13	1	1.07	91	1.3	1.15	1.01
Age at ESRD[a]													
1st degree or *Lower third*	m=52	*85*	*21.46***	0.81**	0.87***	90	*41.47***	0.77**	0.88***	93	2.20***	1.02	0.95·
2nd degree or *Middle third*	(e.t.=18)	*71*	*9.35***	1.006***	1.004***	75	*12.88***	1.007***	1.004***	90	*1.47*	0.999	1.000
3rd degree or *Upper third*		*21*	*Réf.*	0.999***	0.999***	19	*Réf.*	0.999***	0.999***	86	*Réf.*	0.999	0.999
Disease													
Diabetes mellitus	17	27	Ref.	Ref.	Ref.	32	Ref.	Ref.	Ref.	81	Ref.	Ref.	Ref.
Genetic	20	77	8.83***	4.21***	2.11***	80	8.47***	3.98***	1.72***	92	2.62***	2.31**	1.28*
Glomerulonephritis	45	60	3.98***	1.78***	1.50***	63	3.60***	1.45*	1.30**	89	1.84*	1.49	1.22
Unknown	19	65	4.92***	3.25***	1.77***	67	4.29***	2.97***	1.53***	92	2.72***	2.03*	1.31*
Region													
Nord-Pas-de-Calais	6	38	Ref.	Ref.	Ref.	44	Ref.	Ref.	Ref.	84	Ref.	Ref.	Ref.
Alsace	3	57	2.10**	3.47***	1.66*	59	1.88*	4.03***	1.70**	89	1.52	1.82	0.87
Auvergne	3	54	1.90*	2.44*	1.50	65	2.38**	5.00***	1.55*	77	0.62	0.75	0.77
Basse-Normandie	2	74	4.55***	7.31***	2.67***	72	3.30**	5.46***	1.93**	97	6.11·	7.59·	1.41
Burgundy	2	50	1.60	2.67	2.13**	51	1.35	2.60	1.95*	94	3.15	2.58	0.95
Brittany	6	62	2.63***	4.74***	2.83***	64	2.29***	6.05***	2.36***	89	1.48	1.78	1.22

Cox models

- Probability of observing an event conditionally to the duration of observation

- With R

  library(survival)

  result <- survreg(Surv(y,y>0,type='left') ~ x + z, dist='gaussian')

References

Ai, Chunrong, and Edward C Norton. 2003. "Interaction Terms in Logit and Probit Models." *Economics Letters* 80 (1): 123–29.

Baudelot, Christian, Yvanie Caillé, Olivier Godechot, Sylvie Mercier, and Paul Reeve. 2016. "Renal Diseases and Social Inequalities in Access to Transplantation in France." *Population* 71 (1): 23–51.

Coulangeon, Philippe. 2003. "La stratification sociale des goûts musicaux: Le modèle de la légitimité culturelle en question." *Revue française de sociologie* 44 (1): 3–33.

Galton, Francis. 1886. "Regression towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–63.

Godechot, Olivier. 2016. "Financialization Is Marketization! A Study of the Respective Impacts of Various Dimensions of Financialization on the Increase in Global Inequality." *Sociological Science* 3: 495–519.

Godechot, Olivier, and Marc Gurgand. 2000. "Quand les salariés jugent leur salaire." *Économie et statistique* 331 (1): 3–24.

Godechot, Olivier, and Alexandra Louvet. 2010. "Comment les docteurs deviennent-ils directeurs de thèse? Le rôle des réseaux disponibles." *Sociologie*, no. 1: 3–23.

Herpin, Nicolas. 2003. "La taille des hommes: son incidence sur la vie en couple et la carrière professionnelle." *Économie et statistique* 361 (1): 71–90.

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26 (1): 67–82.