Scientific Method

Class 5

MASTER INFORMATIQUE - SIIA

G. Coppin

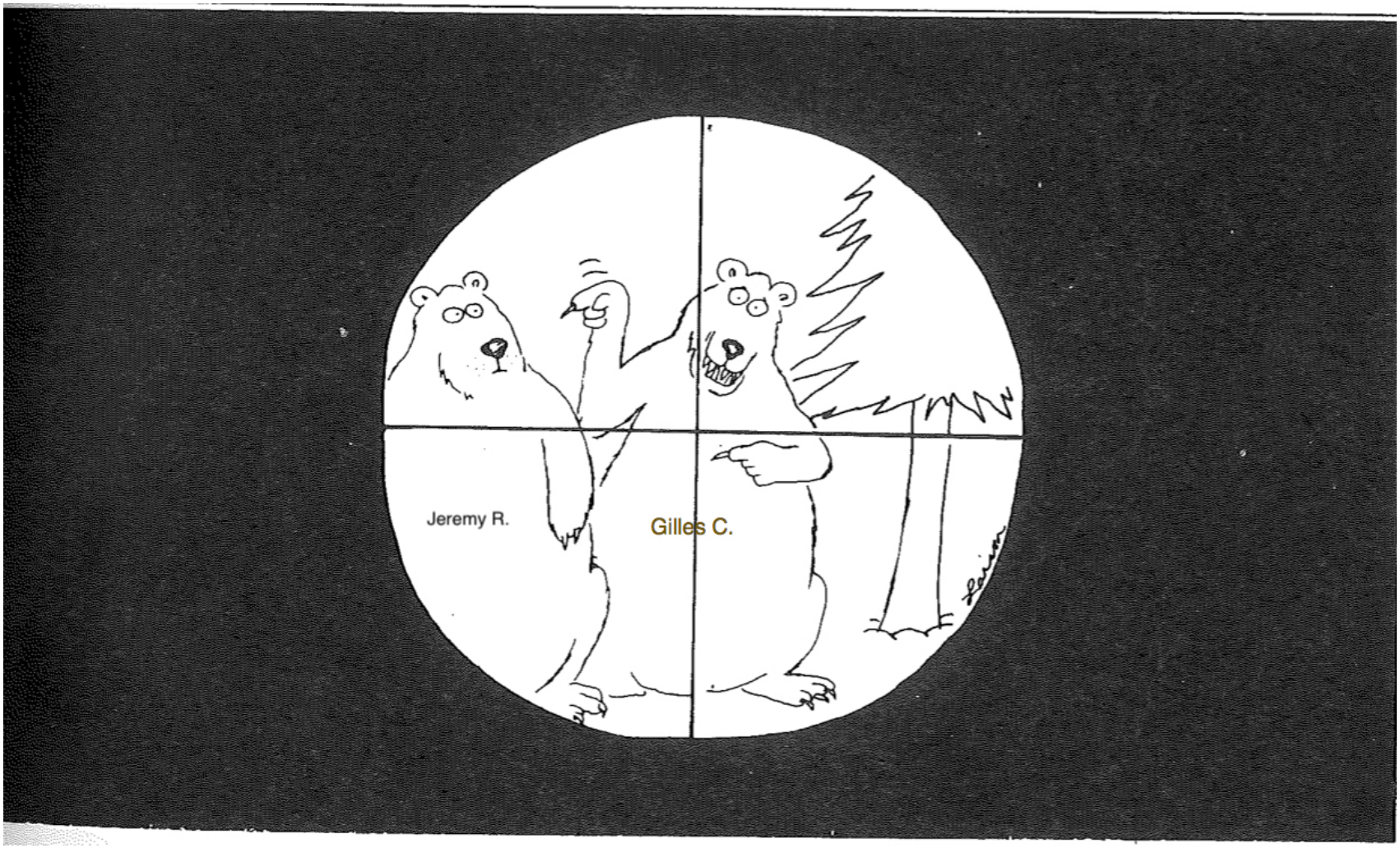
2021-2022

Course objectives

Objs :

1. Regress (linearly)
2. Master the principles of ANOVA

Solidarity teachers

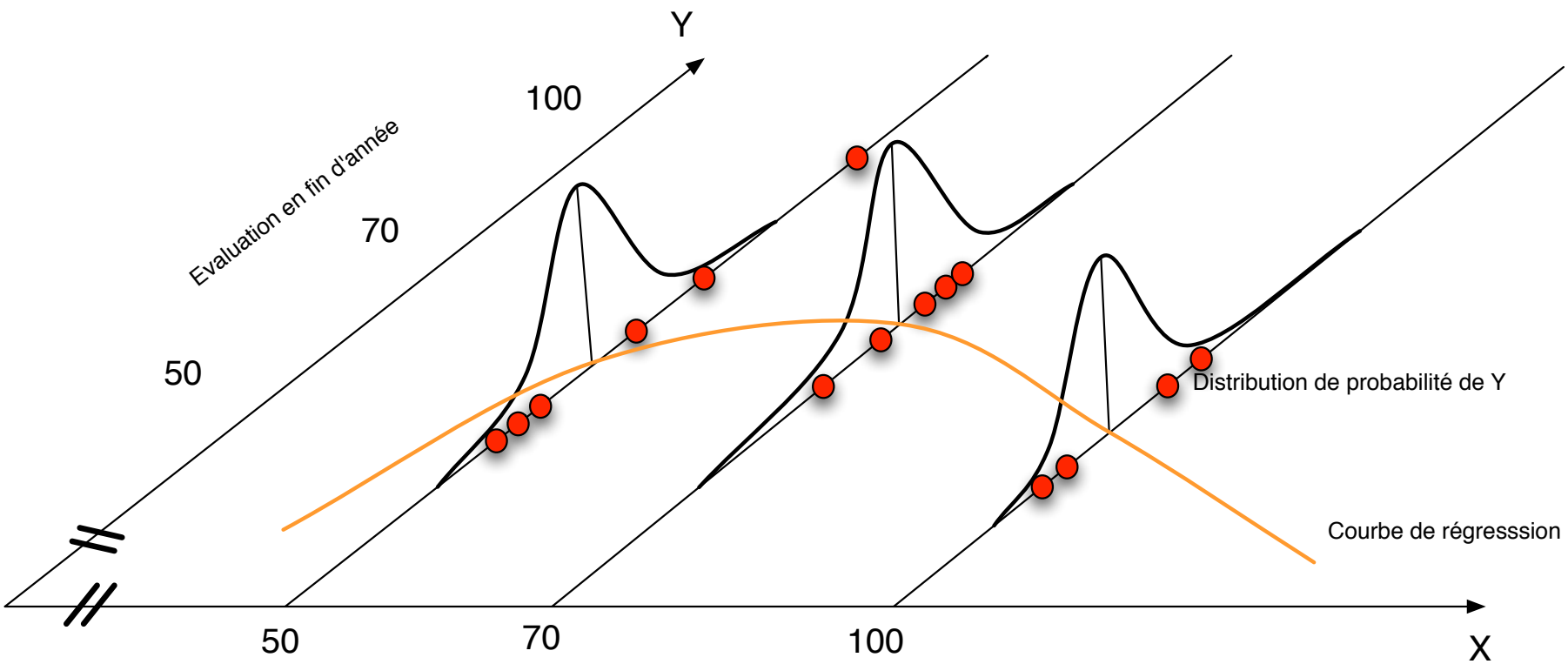


Regression models

Regression consists in building a model which expresses the statistical dependence between two variables : an explanatory variable X and an explained variable Y. For each value of X we obtain a statistical distribution of Y and the average of these distributions varies systematically along X.

Example

We want to test the dependence between the mid-year evaluation of the performance of a set of companies and the evaluation of this same performance at the end of the year. In this case, we can obtain a curve such as :



Linear regression

We try via this approach to model the dependence between two variables by a linear equation. The standard equation of simple linear regression is therefore the following :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

or

- Y_i is the answer to the i^{th} test,
- β_0 and β_1 are the model parameters,
- X_i is a constant (the value of the explanatory variable for the edition in question) and
- ϵ_i an error term reduced to a normal variable with zero expectation and the same variance σ^2.

We assume that ϵ_i and ϵ_j are independent for all i and j.

Some explanations

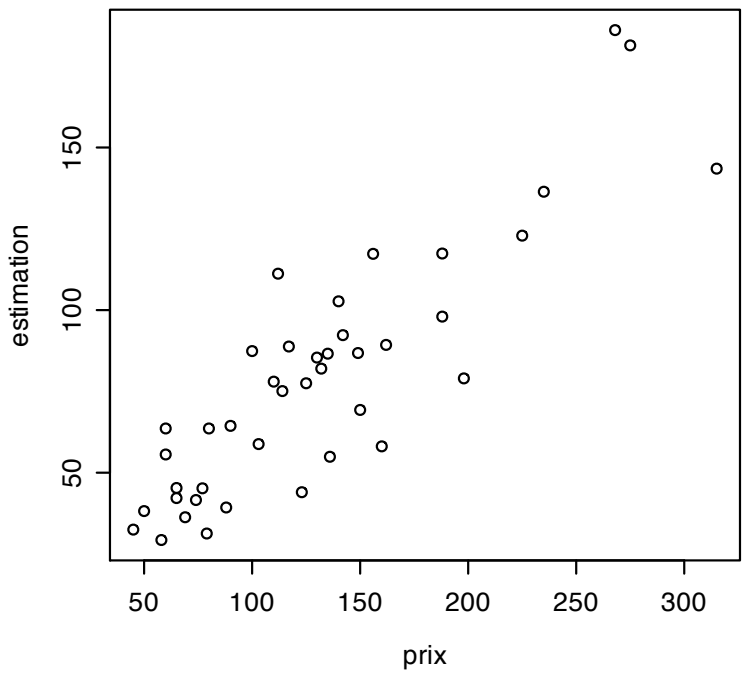
The regression is called *simple* because a single explanatory variable, *linear in parameters and linear in variable*.

Linear regression concerns variables X and Y quantitative *i.e.* numeric.

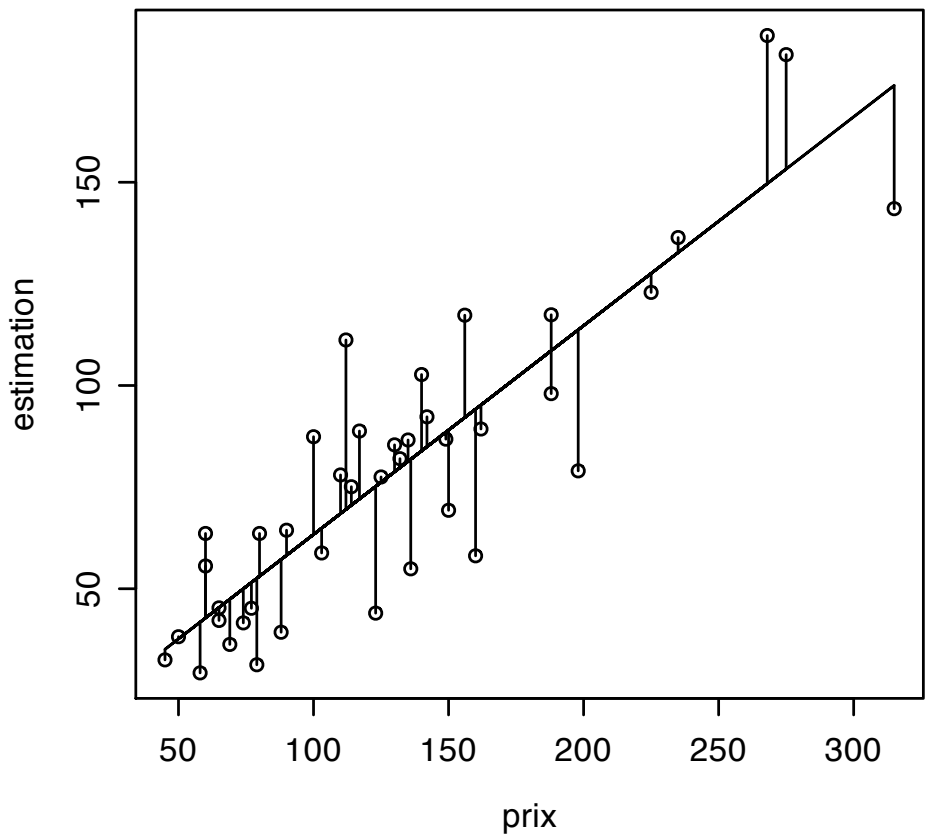
In a linear regression model, the values of X are known and controlled.

Example

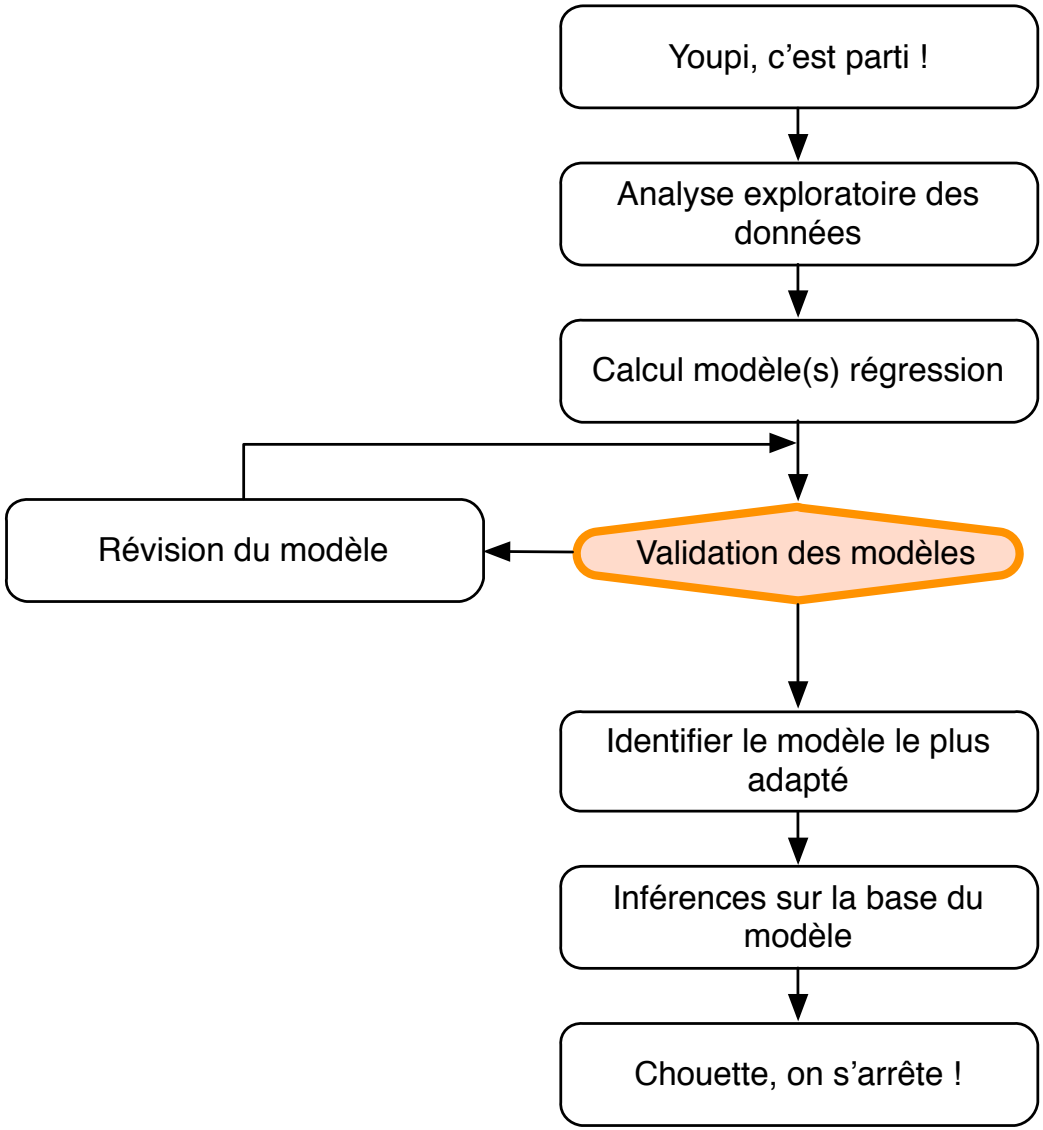
We are working on real estate price data (Parisian apartments) described by the estimates and the real selling price



Example (II)



Regression approach



Least squares line

Let d_i be the vertical distances between the points and the line. The sum of the squares of this line is the indicator of good approximation, that is to say

$$D = \sum_n d_i^2$$

If $\widehat{Y}_i$ is the height of the line at point X_i, then $d_i = \mid Y_i - \widehat{Y}_i \mid$ and

$$D = \sum_n (Y_i - \widehat{Y}_i)^2$$

We will simply look for the line $Y = b_0 + b_1 X$ which minimizes this sum.

Calculation of the coefficients

The coefficients b_0 and b_1 which minimize the distance D are as follows :

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\overline{XY}}{\sum X_i^2 - n\overline{X}^2}$$

and

$$b_0 = \frac{1}{n}(\Sigma Y_i - b_1 \Sigma X_i) = \overline{Y} - b_1 \overline{X}$$

The coefficient b_1 is linked to the correlation coefficient by

$$r = \frac{s_x}{s_y} b_1 = \frac{\widehat{\sigma}_x}{\widehat{\sigma}_y} b_1$$

$$r = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y}$$

Another example (II)

From these values we get

$$b_1 = 3.57$$

and

$$b_0 = 62.37$$

We can therefore estimate that the number of working hours increases by 3.57 hours per unit of output. This makes it possible to estimate the number of hours for a given batch size, thanks to the estimate function :

$$\widehat{Y} = 62.37 + 3.57X$$

For a set of 65 sink drainers (pink), we therefore expect $62.37 + 3.57 \times 65 = 294$ hours of work.

Residues

The residues are defined as : ?

$$e_i = Y_i - \widehat{Y}_i$$

By construction, the sum of the residuals is zero, and the sum of its squares is minimal.

Do not confuse the residuals $e_i = Y_i - \widehat{Y}_i$ with the errors $\epsilon_i = Y_i - E(Y_i)$.

Confidence intervals of the coefficients

b_0 and b_1 are estimators of β_0 and β_1. We can write :

$$b_1 = \sum k_i Y_i$$

with

$$k_i = \frac{X_i - \overline{X}}{\sum (X_i - \overline{X})^2}$$

We then show that $E(b_1) = \beta_1$ and $\sigma^2(b_1) = \sigma^2 \frac{1}{\sum (X_i - \overline{X})^2}$.

Two variables being fixed (β_0 and β_1), we lose two degrees of freedom and we have to approximate σ^2 by

$$s^2 = \widehat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

SSE for Sum of Square Errors.

Confidence intervals of the coefficients

In the case where Y_i are normal distributions, b_1 follows a normal distribution and $\frac{b_1-\beta_1}{\widehat{\sigma}(b_1)}$ a Student law centered at $n-2$ degrees of freedom. We can therefore estimate - as we did previously for the parameter estimates - the confidence interval of the coefficient b_1 by :

$$b_1 \pm t_{\alpha}^{n-2}\widehat{\sigma}(b_1)$$

Return to the example

In the case of our example, we calculate $\widehat{\sigma}(b_1) = 0.34$ and we have $t_\alpha^{n-2} = t_{0.05}^{23} = 2.069$. We can thus estimate that the coefficient is framed by

$$2.85 \leq \beta_1 \leq 4.29$$

We will add between 2.85 and 4.29 hours per unit of unblocking thing.

Estimate of the mean value $E(Y_h)$

We try to estimate the distribution $\widehat{Y}_h$ of the successive estimators of Y for a fixed value of X_h. The following results are obtained :

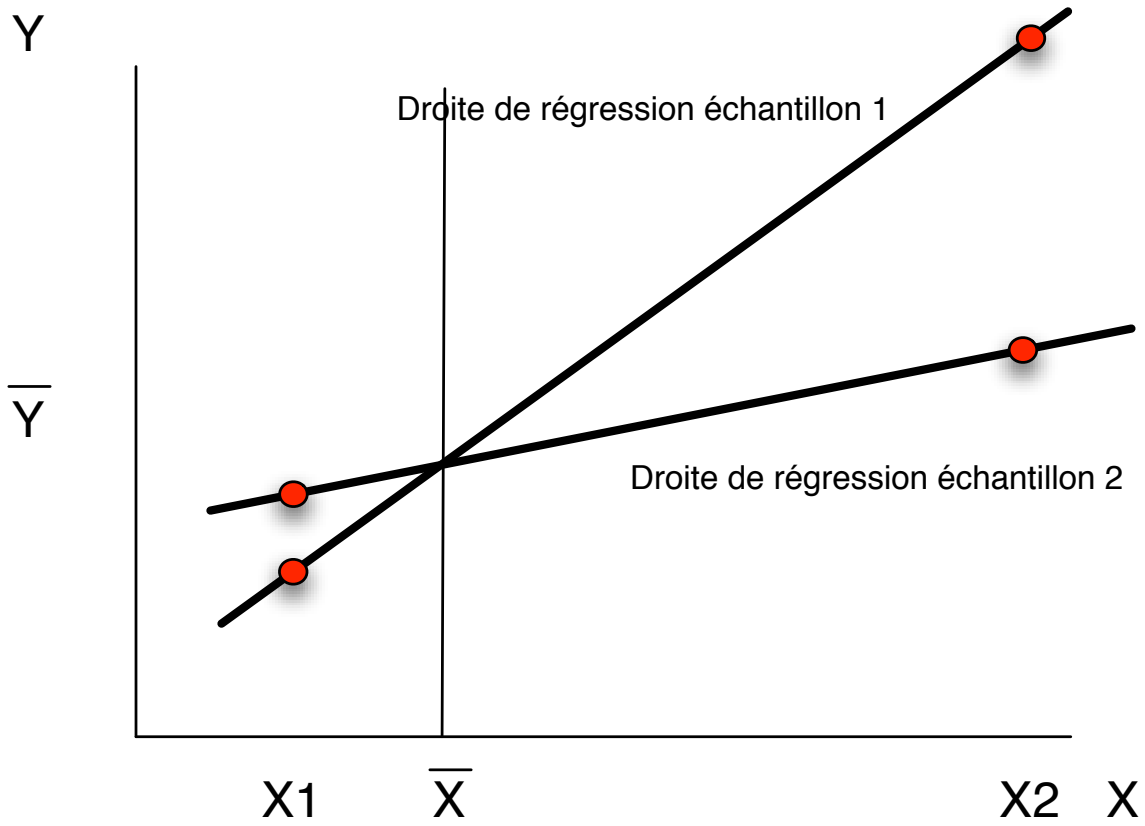
$$E(\widehat{Y}_h) = E(Y_h)$$

$$\sigma^2(\widehat{Y}_h) = \sigma^2[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}]$$

We then replace σ^2 by its estimate $MSE = s^2 = \widehat{\sigma}^2$.

Estimate of the mean value $E(Y_h)$

The further we are from the average, the greater the differences between successive estimates.



Prediction of an interval

Recall that : $b_1 = 3.57$ and $b_0 = 62.37$ and we have $\sigma^2 = 2.384$. So, for 65 stuffers :

$$\widehat{Y}_h = 62.37 + 3.57(65) = 294.4$$

$$\sigma^2(\widehat{Y}_h) = (\widehat{\sigma})^2(\frac{1}{25} + \frac{(65-70)^2}{19.800}) = 98.37$$

$$\sigma = 9.918$$

For a confidence interval of 0.90, we will have :

$$277.4 \leq E(Y_h) \leq 311.4$$

For 100 pieces, we would have :

$$359.52 \leq E(Y_h) \leq 479.42$$

Analysis of variance applied to regression

The following quantities are defined :

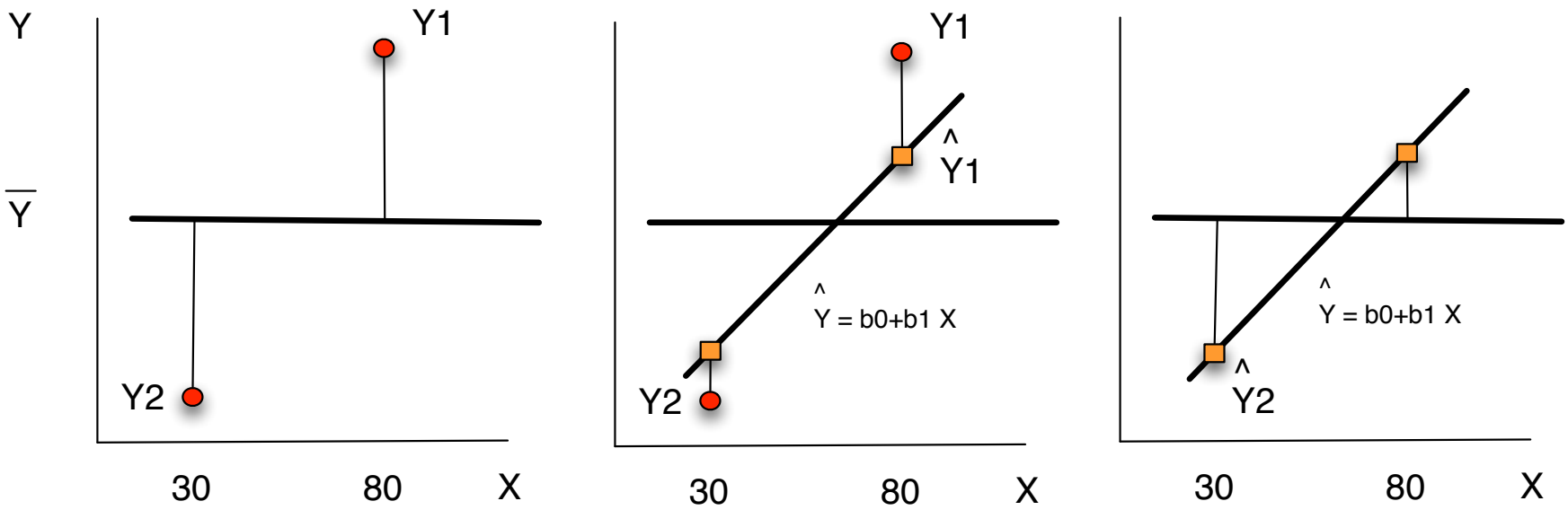
$$SSTO = \sum (Y_i - \overline{Y})^2$$

$$SSE = \sum (Y_i - \widehat{Y}_i)^2$$

$$SSR = \sum (\widehat{Y}_i - \overline{Y})^2$$

with SSTO (or SSY) : total sum of squares, SSE = error sum of squares, SSR : regression sum of squares.

Analysis of variance applied to regression (II)



Relations between error sums

We have :

$$Y_i - \overline{Y} = \widehat{Y}_i - \overline{Y} + Y_i - \widehat{Y}_i$$

hence :

$$\sum (Y_i - \overline{Y})^2 = \sum (\widehat{Y}_i - \overline{Y})^2 + \sum (Y_i - \widehat{Y}_i)^2$$

that is

$$SSTO = SSR + SSE$$

Degrees of freedom

SSTO has $n-1$ degrees of freedom : the $Y_i - \overline{Y}$ have a zero sum.

SSE has $n-2$ degrees of freedom : the parameters β_0 and β_1 are estimated and block two degrees of freedom.

SSR has 1 degree of freedom : the $\widehat{Y}_i$ are computed from the regression line, which has two degrees of freedom linked to the coefficients. We lose one for zero sum (same as SSTO).

Fisher's test for regression

We will test the hypothesis :

$$H_0 : \beta_1 = 0$$

For this, we see that :

$$E(MSE) = \sigma^2$$

and

$$E(MSR) = \sigma^2 + \beta_1^2 \sum (X_i - \overline{X})^2$$

If $\beta_1 = 0$, we see that MSE and MSR must be of the same order of magnitude.

Fisher-Snedecor Law

If X and Y are respectively laws of type χ^2_n and χ^2_p, then we can define the following law :

$$F(n;p) : \frac{X/n}{Y/p}$$

This law serves as a reference for variance analyzes.

Which test for regression ?

If H_0 is respected, we must find with F a Fisher distribution, so we will consider the value of the estimate of the statistic and analyze its positioning with respect to the threshold value (as usual ...). If the statistic is less than the threshold value $F(1-\alpha; 1, n-2)$, then we cannot reject H_0.

In the case of the example, we get $MSR = 252,378$ and $MSE = 2,384$, so $F = 105$ that we must compare to the threshold value for $n - 2 = 23$, that is to say $4,29$. It seems that we can't keep H_0, so we can't keep the hypothesis $\beta_1 = 0$, so there is a linear dependency relation between X and Y.

And in R ...

```
> summary(model)

Call:
lm(formula = V2 ~ V1)

Residuals:
   Min     1Q  Median     3Q    Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
V1             3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Coefficient of determination

To know what is the "impact" of the explanatory variable X on the determination of Y, we can come back to the relationship between SSR and SSTO. The closer this ratio is to 1, the more Y will be "governed" by the values of X. We therefore define :

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

R^2 is between 0 and 1.

Coefficient of determination (II)

Warning !

- a high R coefficient does not mean that we can predict correctly
- a high R coefficient does not necessarily indicate that the regression line is optimal
- a coefficient R close to 0 does not mean that X and Y are independent

We have the following relation :

$$r = +- \sqrt{R^2}$$

Limits of linear regression

- the regression function is not linear
- error terms do not have constant variance
- error terms are not independent
- the models are suitable but there are outliers
- the error terms are not distributed according to a normal distribution
- explanatory variables are missing

Analysis of extreme individuals

- Linear regression is very sensitive to extreme individuals, which can strongly influence the coefficients.
- We can study extreme individuals and their influence by the *jackknife* technique (differential regression analysis), *leverage*, *dfitts* and *distance analyzes. Cook.*
- The objective is to decide if these individuals should be finally discarded ("abnormal", "outside the model"), or on the contrary if they are representative.

Leverage

- The *leverage* formula is :

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{(n-1){s_x}^2}$$

  where h_i represents the influence of the individual i, $\frac{1}{n} \leq h_i \leq 2$.
- We generally avoid keeping the individuals for whom $h_i \geq \frac{4}{n}$.

Dfitts

- We observe the effect of the absence of the individual on the result of the regression, by evaluating :

$$dfitts = \frac{e_i \sqrt{h_i}}{MSE_{sans\ i}(1 - h_i)}$$

- We generally avoid keeping individuals for which *dfitts* has an absolute value greater than 1 (small samples) or $2\sqrt{\frac{2}{n}}$ (other sample sizes).

Cook's distance

- The effect of individual i can also be evaluated by :

$$D_{cook\ i} = \frac{e_{i\ std}^2 h_i}{2(1-h_i)}$$

- $D_{cook\ i}$ must not exceed a tabulated value $D_{cook\ ref}$ calculated according to the quantile of a Fisher law.

Analysis of residuals

- Reminder : the residuals must be normal distributions of the same variance and independent.
- Verification is most often graphical. It is done on the pairs $(e_{i\ std}, \widehat{Y})$, with :

$$e_{i\ std} = \frac{e_i}{s_{e_i}} = \frac{e_i}{\sqrt{MSE(1-h_i)}}$$

Linearization transformations

Function	Transformation	Linear form
$y = \alpha x^{\beta}$	$y' = Logy$ $x' = Logx$	$y' = Log + \beta x'$
$y = \alpha e^{\beta x}$	$y' = Logy$	$y' = Log\alpha + \beta x$
$y = \alpha + \beta Logx$	$x' = Logx$	$y = \alpha + \beta x'$
$y = \frac{x}{\alpha x - \beta}$	$y' = 1/x$	$y' = \alpha - \beta x'$
	$x' = 1/x$	
$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$y' = Log(\frac{y}{1-y})$	$y' = \alpha + \beta x$

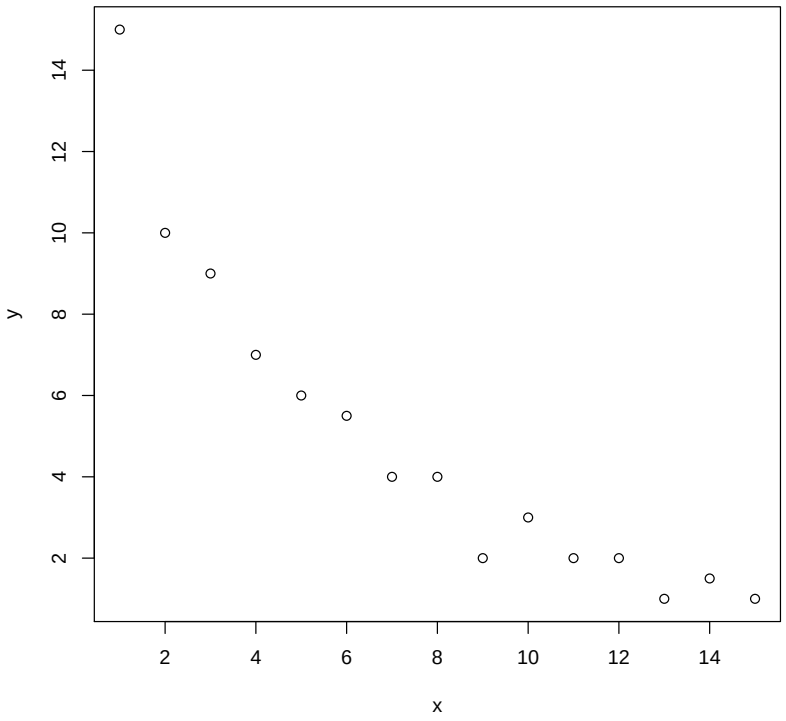
Example

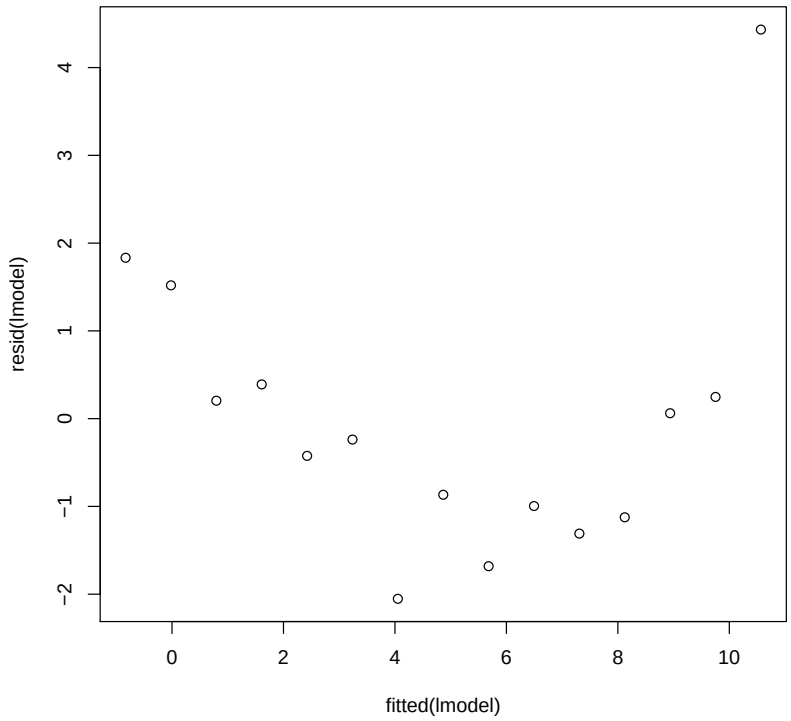
```
"x" "y"
"1" 1 15
"2" 2 10
"3" 3 9
"4" 4 7
"5" 5 6
"6" 6 5.5
"7" 7 4
"8" 8 4
"9" 9 2
"10" 10 3
"11" 11 2
"12" 12 2
"13" 13 1
"14" 14 1.5
"15" 15 1
```

Example

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0524 -1.0595 -0.2381  0.3190  4.4333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.3810     0.9238  12.320 1.52e-08 ***
x            -0.8143     0.1016  -8.014 2.19e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.7 on 13 degrees of freedom
Multiple R-squared: 0.8317,   Adjusted R-squared: 0.8187
F-statistic: 64.23 on 1 and 13 DF,  p-value: 2.193e-06
```

Example

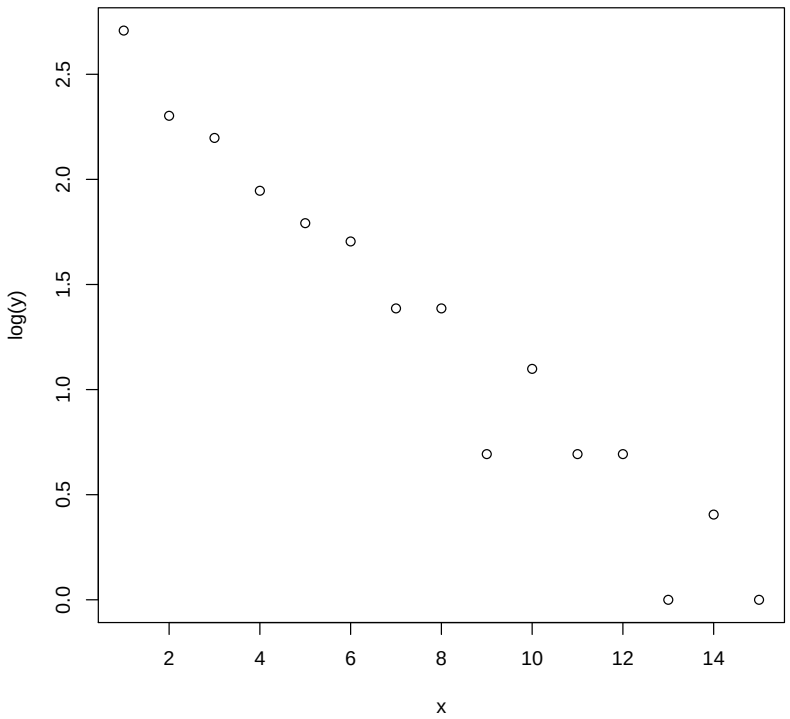


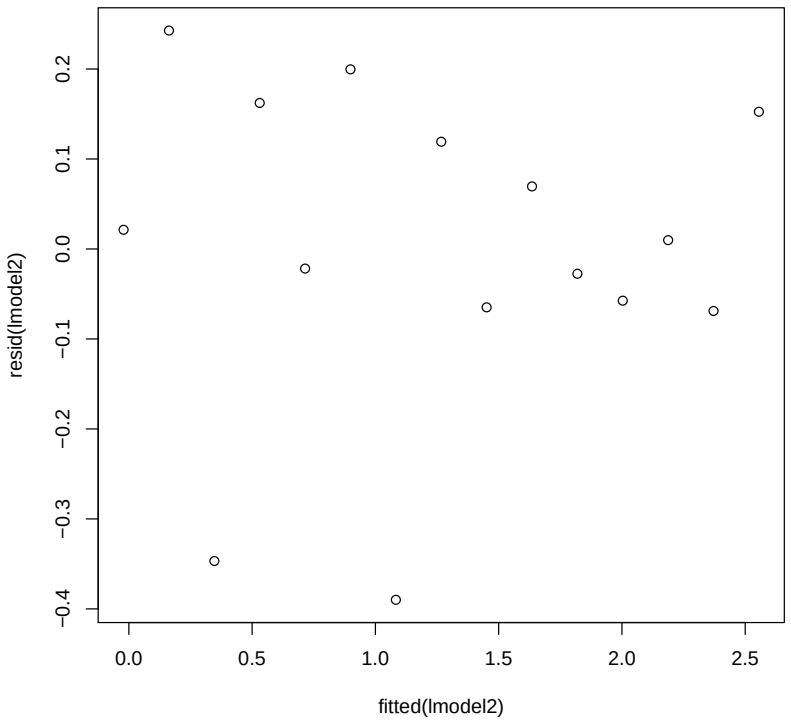


Example

```
lm(formula = log(y) ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38988 -0.06115  0.00982  0.13586  0.24275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73959    0.10151   26.99 8.42e-13 ***
x           -0.18406    0.01116  -16.49 4.28e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1868 on 13 degrees of freedom
Multiple R-squared: 0.9544,   Adjusted R-squared: 0.9508
F-statistic: 271.8 on 1 and 13 DF,  p-value: 4.282e-10
```

Example


log(y)
x
0.0
0.5
1.0
1.5
2.0
2.5
2
4
6
8
10
12
14


resid(lmodel2)
fitted(lmodel2)
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.0
0.5
1.0
1.5
2.0
2.5

Variance related transformations

Distribution	Variance $f(\mu)$	Transformation	Resulting Variance
Fish	μ	$\sqrt{(y)}$	0.25
Binomial	$\frac{\mu(1-\mu)}{n}$	$Arc\sin\sqrt{(y)}$	$\frac{0.25}{n}$

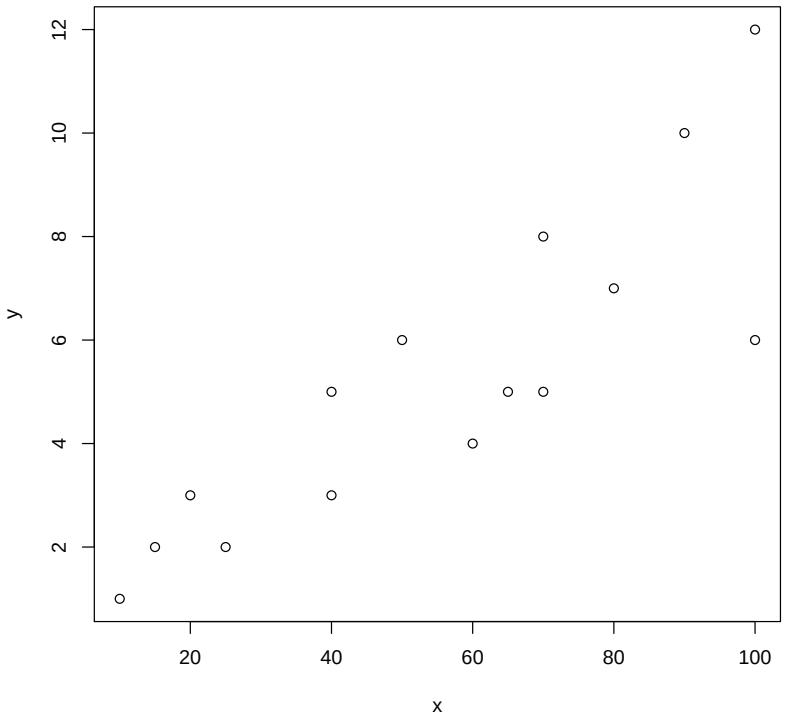
Example

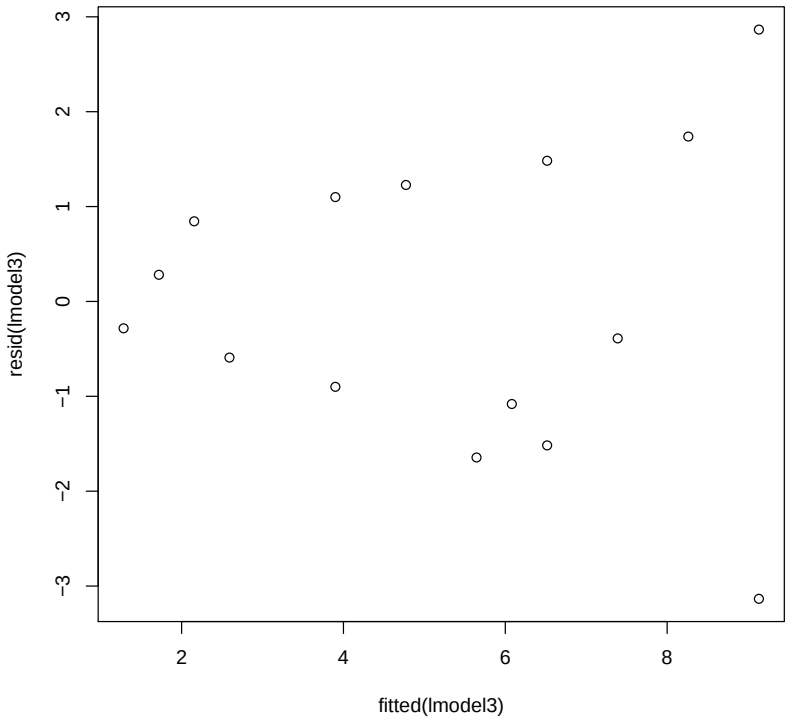
```
"x" "y"
"1" 10 1
"2" 15 2
"3" 20 3
"4" 25 2
"5" 40 3
"6" 40 5
"7" 50 6
"8" 60 4
"9" 65 5
"10" 70 5
"11" 70 8
"12" 80 7
"13" 90 10
"14" 100 6
"15" 100 12
```

Example

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1342 -0.9904 -0.2828  1.1639  2.8658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.41045    0.90598   0.453    0.658
x            0.08724    0.01442   6.048 4.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 13 degrees of freedom
Multiple R-squared: 0.7378,   Adjusted R-squared: 0.7176
F-statistic: 36.58 on 1 and 13 DF,  p-value: 4.113e-05
```

Example

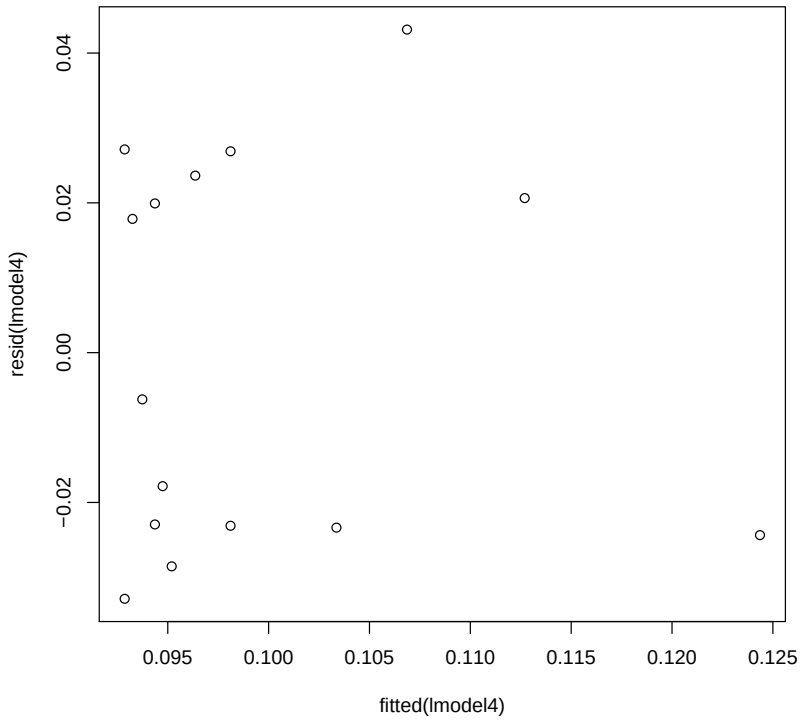

y
x
2
4
6
8
10
12
20
40
60
80
100


resid(lmodel3)
fitted(lmodel3)
−3
−2
−1
0
1
2
3
2
4
6
8

Example

```
Call:
lm(formula = z ~ t)

Residuals:
      Min        1Q    Median        3Q       Max
-0.032863 -0.023237 -0.006238  0.022138  0.043138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08936    0.01067   8.378 1.34e-06 ***
t            0.34997    0.28112   1.245    0.235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02699 on 13 degrees of freedom
Multiple R-squared: 0.1065,   Adjusted R-squared: 0.03779
F-statistic:  1.55 on 1 and 13 DF,  p-value: 0.2351
```

Example

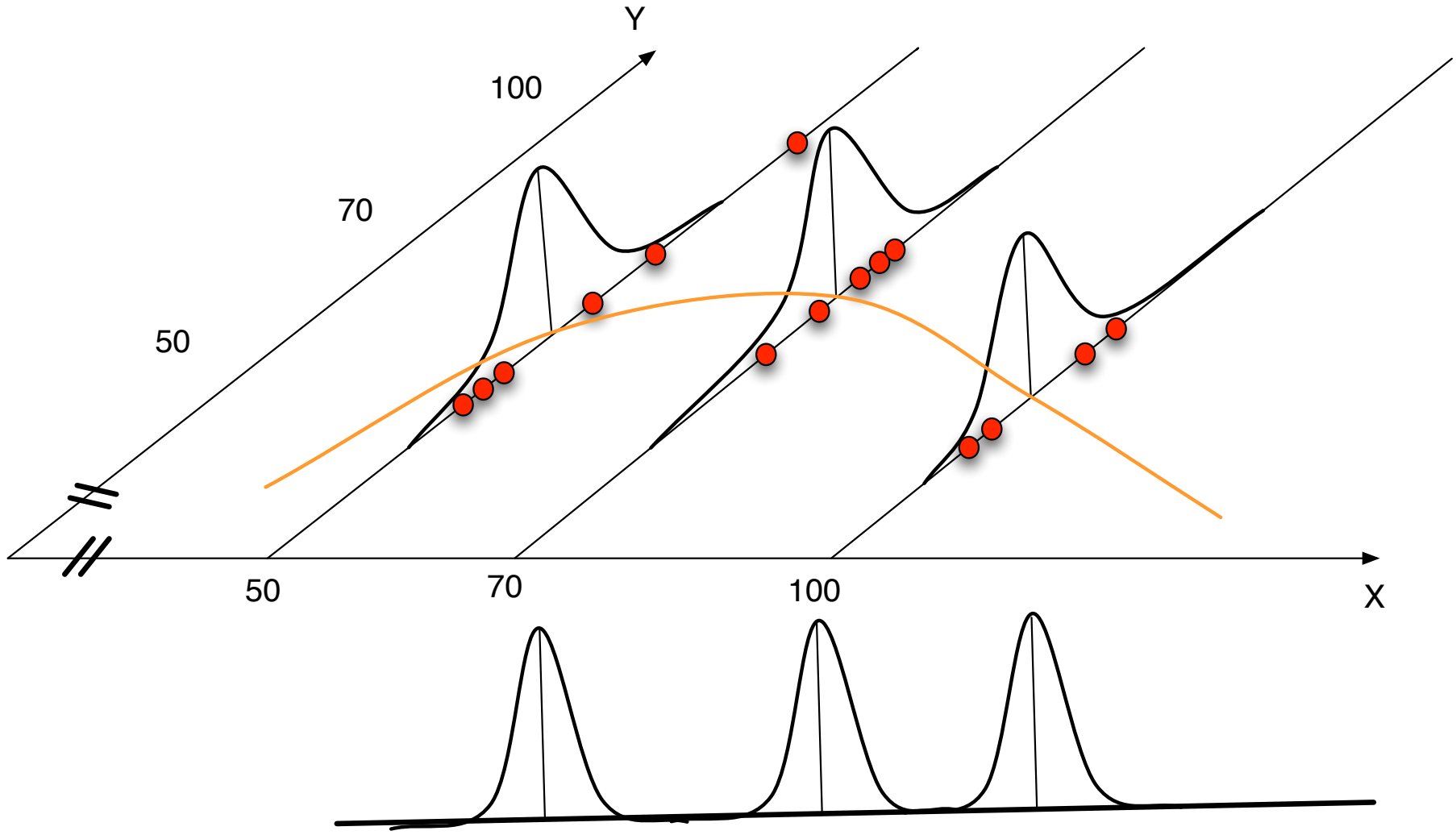

resid(lmodel4)
0.04
0.02
0.00
−0.02
0.095
0.100
0.105
0.110
0.115
0.120
0.125
fitted(lmodel4)

Analysis of variance for ... comparison of means

We want to know what is the influence of factors on the behavior of populations. These factors or explanatory variables are qualitative variables, the observed or explained variable is numeric. The modalities of the different factors are generally called levels.

One- or two-way analysis of variance (ANOVA).

ANOVA and regression



ANOVA model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + .. + \beta_n X_{ij,n} + \epsilon_{ij}$$

with $X_{ijk} = 1$ if factor k and 0 otherwise.

Formalization of the problem

We have k samples of sizes $n_1, n_2, .., n_k$ corresponding to the different modalities of the factor $A_1, A_2, .., A_k$.

Factor	A_1	A_2	..	..	A_k
	y_1^1	y_2^1	..	..	y_k^1
	y_1^2	y_2^2	..	..	y_k^2
	..	..	..	..	..
	$y_1^{n_1}$	$y_2^{n_2}$	..	..	$y_k^{n_k}$
Average	$\overline{y_1}$	$\overline{y_2}$	..	..	$\overline{y_k}$

We want to know if $H_0 : m_1 = m_2 = ... = m_k$

$$y_i^j = m_i + \epsilon_i^j = \mu + \alpha_i + \epsilon_i^j$$

α_i effect of the level i of the factor and ϵ_i^j of distribution $N(0, \sigma)$.

Variance decomposition

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_i^j$$

with $Y_i^j - \overline{Y} = Y_i^j - \overline{Y}_i + \overline{Y}_i - \overline{Y}$, we get :

$$\sum_i \sum_j (Y_i^j - \overline{Y})^2 = \sum_i \sum_j (Y_i^j - \overline{Y}_i)^2 + \sum_i n_i (\overline{Y}_i - \overline{Y})^2$$

that is

$$SSY(SSTO) = SSE + SSR$$

Variance comparison

By writing $SSR = \frac{1}{n}\sum_{i=1}^{k} n_i SSR_i$, we get that :

$$\frac{nSSR}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i SSR_i}{\sigma^2}$$

Knowing that $\frac{n_i SSR_i}{\sigma^2}$ is a variable of type $\chi^2_{n_i-1}$, the variable $\frac{nSSR}{\sigma^2}$ is it of type χ^2_{n-k}.
Likewise, the variable $\frac{nSSE}{\sigma^2}$ is a variable of type χ^2_{k-1}.

Fisher-Snedecor law for comparison

If the means were identical (H_0 : the factor levels have no influence), we should have an identical influence between the intragroup effects (SSR) and the intergroup effects (SSE). This should therefore result in a ratio of 1 between the two quantities, namely :

$$\frac{MSE}{MSR} = \frac{SSE/k-1}{SSR/n-k} = F(k-1;n-k) \simeq 1$$

The F law is a Fisher-Snedecor law (at $k-1$ and $n-k$ degrees of freedom). If the value obtained for F is too extreme, we will refute the hypothesis of identical means.

Analysis of variance : example

We are trying to find out if there are differences between the housing tax rates depending on the region. We have the following table :

area	number	average	variance
center	13	4.38	3.63
is	10	17.66	4.38
idf	26	11.76	15.04
north	9	25.95	50.40
west	14	18.89	9.59
southeast	18	19.76	8.63
southwest	10	20.51	20.69

ANOVA : example (II)

We calculate :

- the inter-group variance $SSE = 1706$
- the intra-group variance $SSR = 1320$

With $k - 1 = 6$ and $n - k = 93$, we have $df = 99$ and $F = 20.03$

ANOVA : example (II)

We calculate :

- the inter-group variance $SSE = 1706$
- the intra-group variance $SSR = 1320$

With $k - 1 = 6$ and $n - k = 93$, we have $df = 99$ and $F = 20.03$

Which excludes equality of averages

Contrasts

We can still want to know if within the means, we have identical pairs ($m_i = m_j$). For this, we rely on Scheffé's formula which indicates that

$$m_i - m_j - SSY\widehat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \leq x_i - x_j \leq m_i - m_j + SSY\widehat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

takes place with a probability

$$P(F_{k-1;n-k} \leq \frac{SSY^2}{k-1}) = 1-\alpha$$

We calculate $S = \sqrt{(k-1)F_\alpha(k-1;nk)}$ and if $\mid x_i - x_j \mid > S\overline{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$ then the means m_i and m_j are different.