Cours Statistiques et Analyse de Données

Cours 5, déjà...

FILIÈRE ISA - UV1

R. Billot et G. Coppin

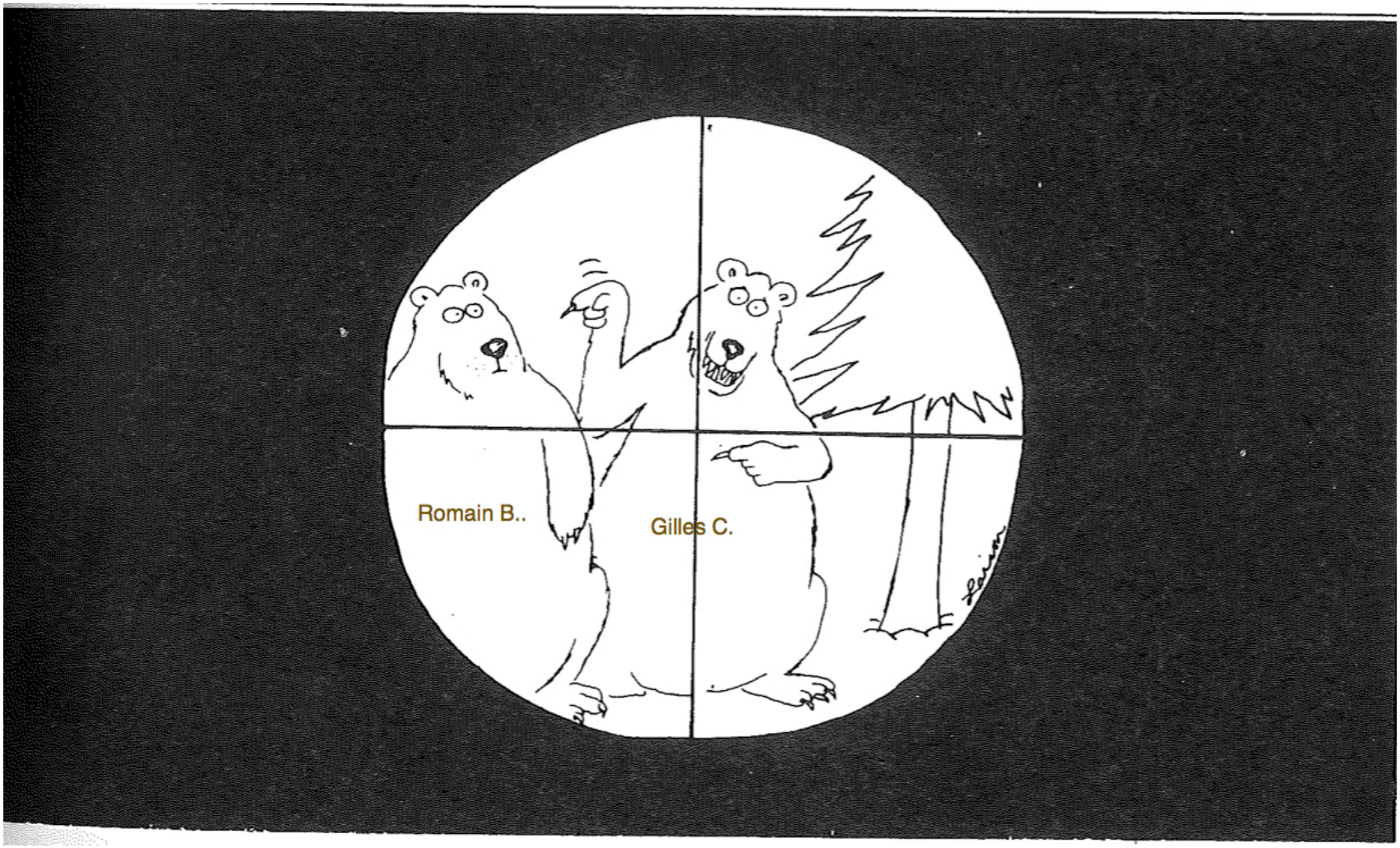
2016 - 2017

Objectifs du cours

Objs :

1. Régresser (linéairement)
2. Maîtriser les principes de l'ANOVA

Des enseignants solidaires

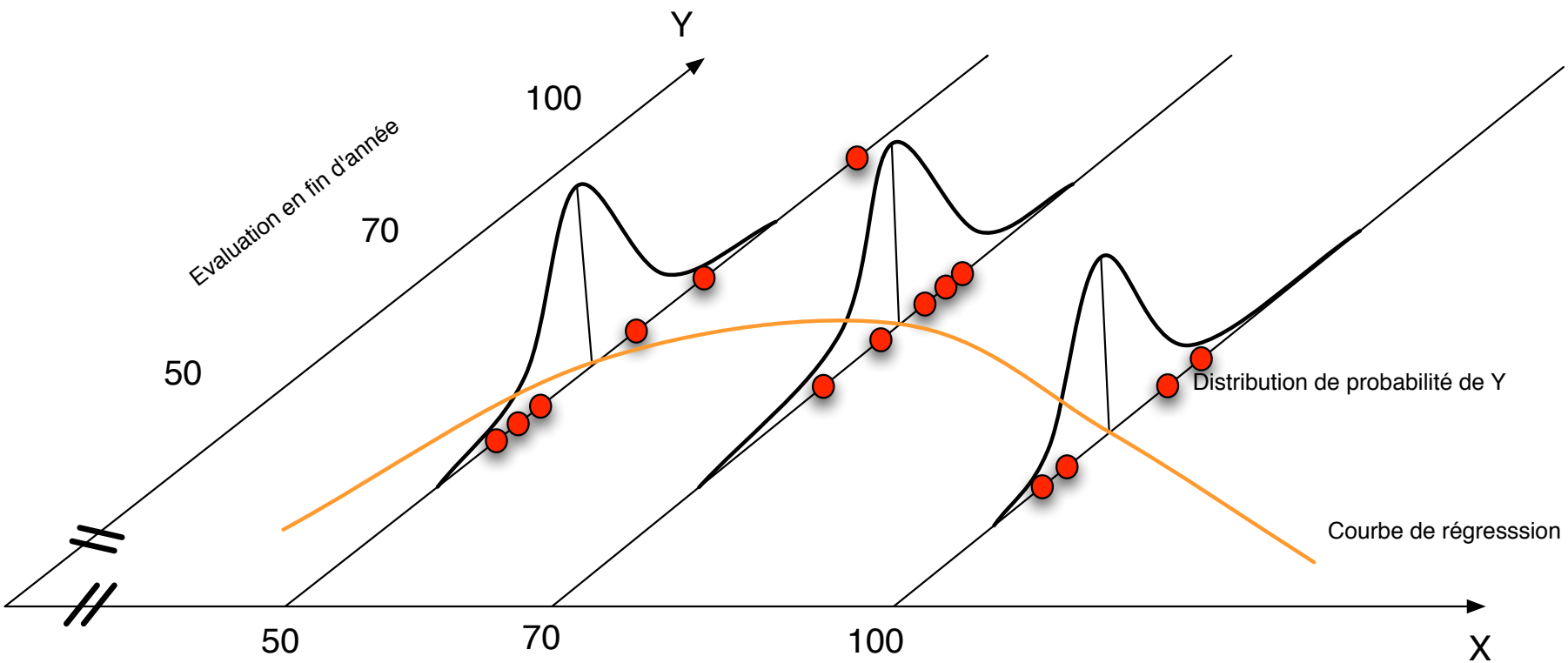


Modèles de régression

La régression consiste à bâtir un modèle qui exprime la dépendance statistique entre deux variables : une variable explicative X et une variable expliquée Y. Pour chaque valeur de X on obtient une distribution statistique de Y et la moyenne des ces distributions varie de façon systématique selon X.

Exemple

On veut tester la dépendance entre l'évaluation à mi-année des performances d'un ensemble d'entreprises et l'évaluation de cette même performance à la fin d'année. On peut obtenir dans ce cas une courbe telle que :



Régression linéaire

On essaie via cette démarche de modéliser la dépendance entre deux variables par une équation linéaire. L'équation standard de la régression linéaire simple est donc la suivante :

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

où

- Y_i est la réponse au i^{eme} essai,
- β_0 et β_1 sont les paramètres du modèle,
- X_i est une constante (la valeur de la variable explicative pour le tirage en question) et
- ϵ_i un terme d'erreur ramené à une variable normale d'espérance nulle et de même variance σ^2.

On pose que ϵ_i et ϵ_j sont indépendants pour tout i et j.

Quelques explications

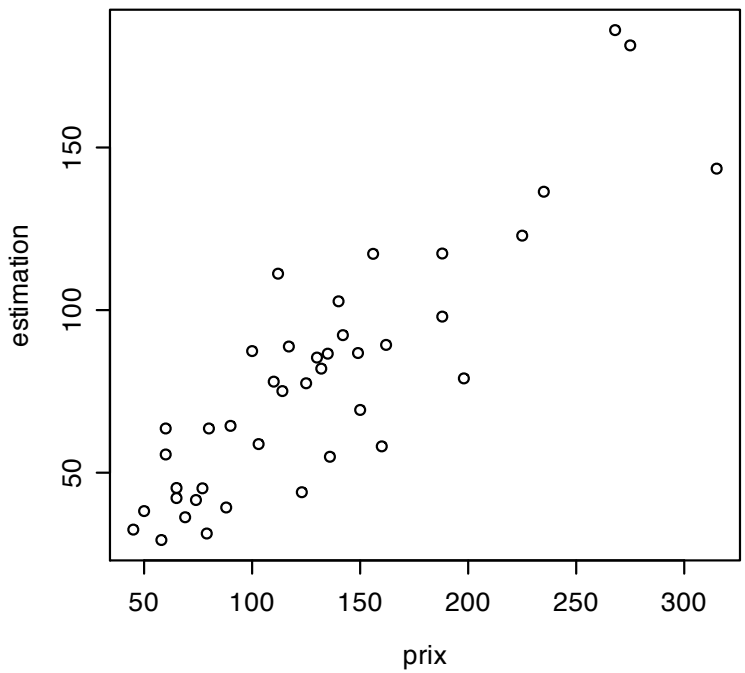
La régression est appelée *simple* parce que une seule variable explicative, *linéaire en paramètres et linéaire en variable*.

La régression linéaire concerne des variables X et Y quantitatives *i.e.* numériques.

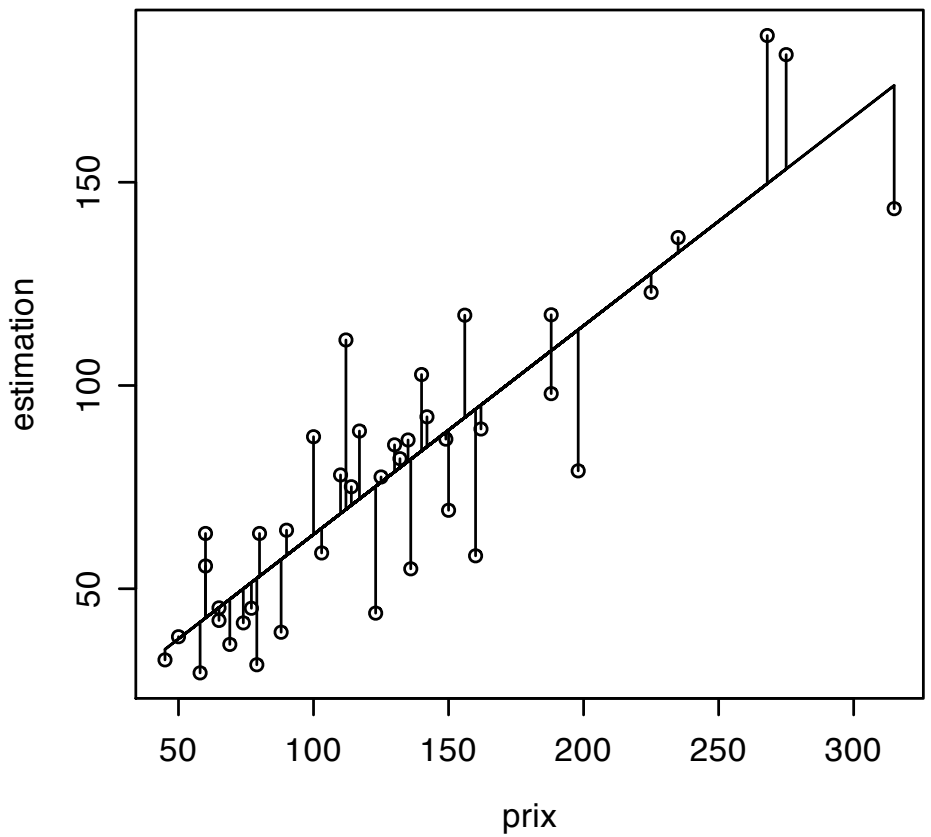
Dans un modèle de régression linéaire, les valeurs de X sont connues et contrôlées.

Exemple

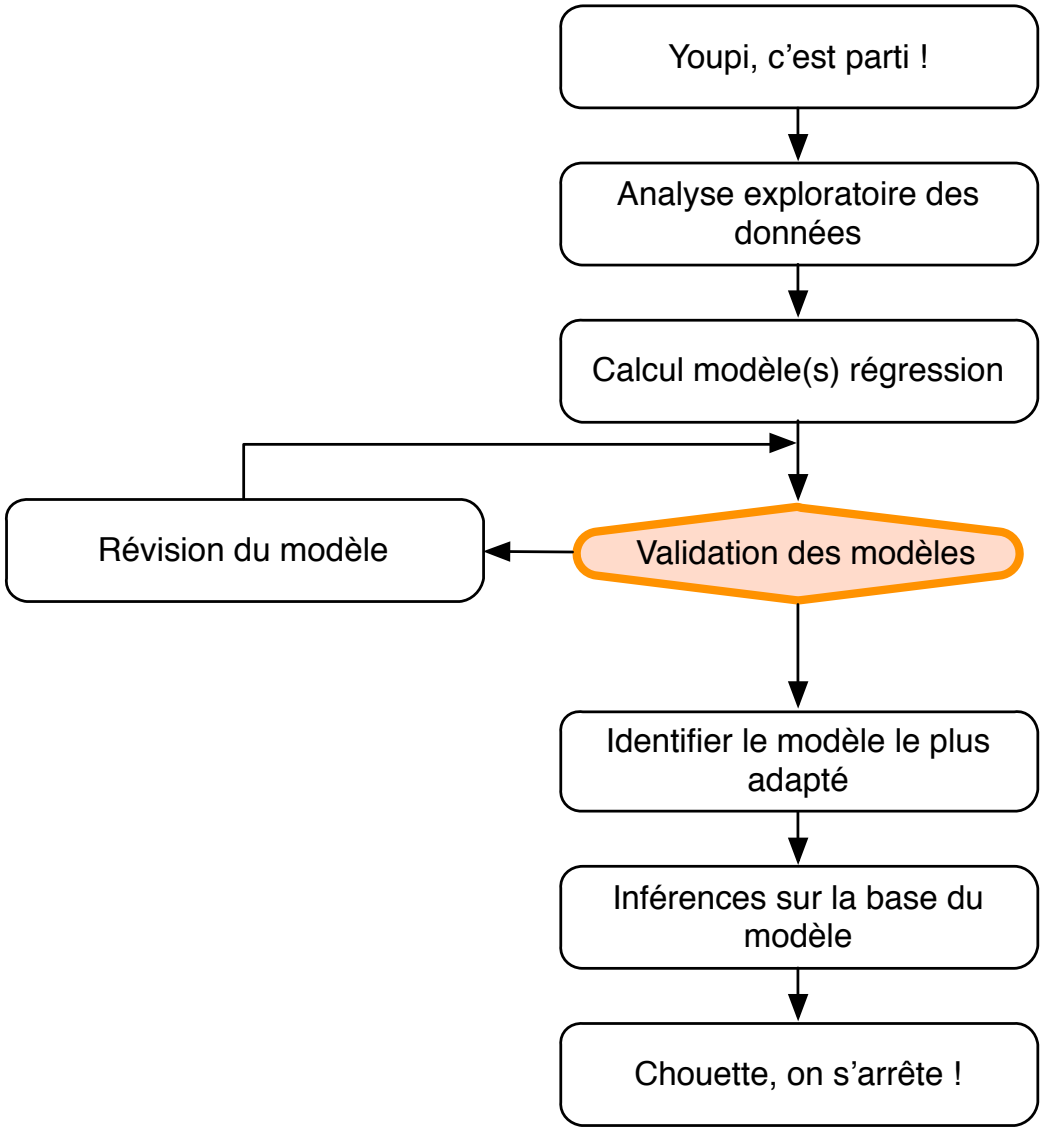
On travaille sur des données de prix immobiliers (appartements parisiens) décrits par les estimations et le prix de vente réel



Exemple (II)



Démarche de la régression



Droite des moindres carrés

Soient d_i les distances verticales entre les points et la droite. La somme des carrés de cette droite est l'indicateur de bonne approximation, soit

$$D = \sum_n d_i^2$$

Si $\widehat{Y}_i$ est la hauteur de la droite au point X_i, alors $d_i = \mid Y_i - \widehat{Y}_i \mid$ et

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

On va tout simplement chercher la droite $Y = b_0 + b_1 X$ qui minimise cette somme.

Calcul des coefficients

Les coefficients b_0 et b_1 qui minimisent la distance D sont les suivants :

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

et

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

Le coefficient b_1 est lié au coefficient de corrélation par

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

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

Autre exemple

Une compagnie de fabrication de débouche-éviers roses essaie d'optimiser la taille des lots de fabrication

Essai	Taille lot	Heures		
i	X_i	Y_i	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1	80	399	100	7.520
2	30	121	1.600	36.588
..	..	..	..	..
23	40	244	900	4.662
24	80	342	100	883
25	70	323	0	115
Total	1.750	7.807	19.800	307.203

Autre exemple (II)

A partir de ces valeurs on obtient :

$$b_1 = 3,57$$

et

$$b_0 = 62,37$$

On peut donc estimer que le nombre d'heures de travail augmente de $3,57$ heures par unité de production. Cela permet ainsi d'estimer le nombre d'heures pour une taille de lot donnée, grâce à la fonction d'estimation :

$$\widehat{Y} = 62,37 + 3,57X$$

Pour un lot de 65 débouche-éviers (roses), on s'attend ainsi à $62,37 + 3,57 \times 65 = 294$ heures de travail.

Résidus

On appelle résidus les différences entre valeurs observées Y_i et valeurs estimées $\widehat{Y}_i$.

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

Par construction, la somme des résidus est nulle, et la somme de ses carrés est minimale.

Il ne faut pas confondre les résidus $e_i = Y_i - \widehat{Y}_i$ avec les erreurs $\epsilon_i = Y_i - E(Y_i)$.

Intervalles de confiance des coefficients

b_0 et b_1 sont des estimateurs de β_0 et β_1. On peut écrire :

$$b_1 = \sum k_i Y_i$$

avec

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

On montre alors que $E(b_1) = \beta_1$ et $\sigma^2(b_1) = \sigma^2 \frac{1}{\sum (X_i - \overline{X})^2}$.

Deux variables étant fixées (β_0 et β_1), on perd deux degrés de liberté et on est obligés d'approximer σ^2 par

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

SSE pour Sum of Square Errors.

Intervalles de confiance des coefficients

Dans le cas où Y_i sont des lois normales, b_1 suit une loi normale et $\frac{b_1-\beta_1}{\widehat{\sigma}(b_1)}$ une loi de Student centrée à $n-2$ degrés de liberté. On peut donc estimer - comme on l'a fait précédemment pour les estimations de paramètres - l'intervalle de confiance du coefficient b_1 par :

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

Retour sur l'exemple

Dans le cas de notre exemple, on calcule $\widehat{\sigma}(b_1) = 0,34$ et on a $t_\alpha^{n-2} = t_{0.05}^{23} = 2.069$. On peut ainsi estimer que le coefficient est encadré par

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

On ajoutera entre 2.85 et 4.29 heures par unité de débouche-chose.

Estimation de la valeur moyenne $E(Y_h)$

On cherche à estimer la distribution $\widehat{Y}_h$ des estimateurs successifs de Y pour une valeur fixée de X_h. On obtient les résultats suivants :

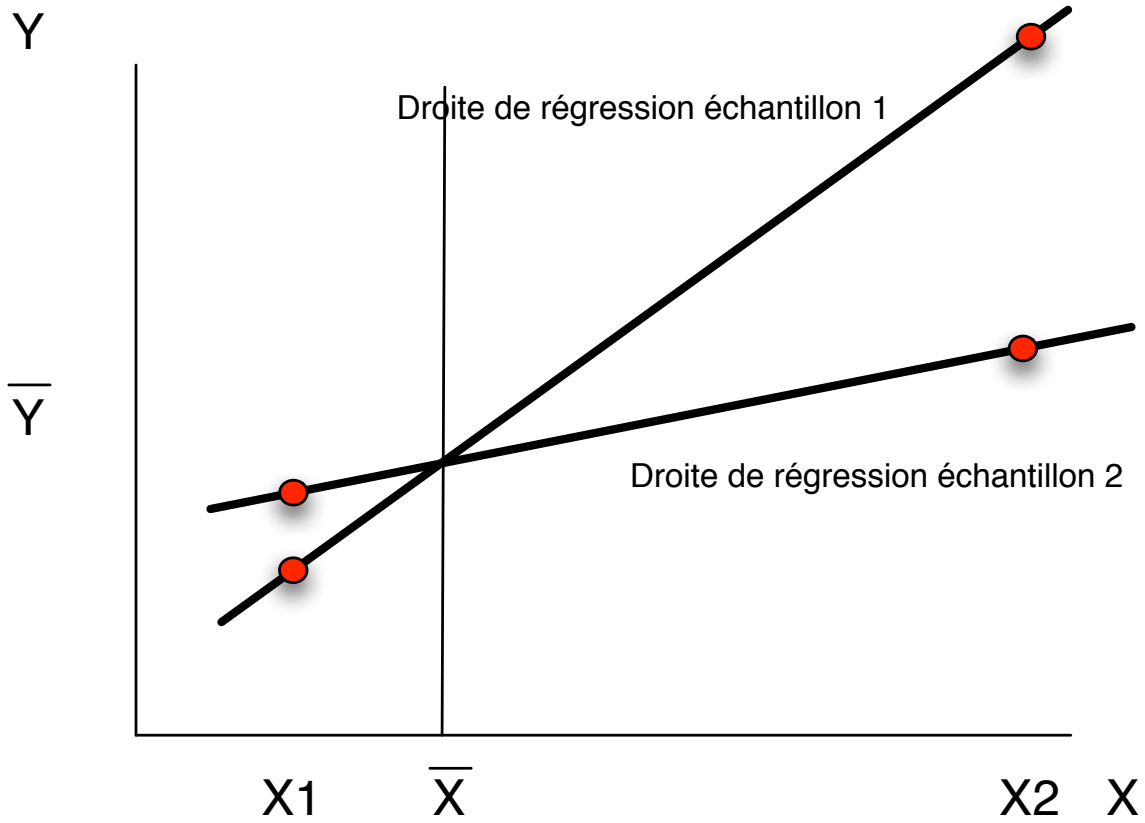
$$E(\widehat{Y}_h) = E(Y_h)$$

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

On remplace ensuite σ^2 par son estimation $MSE = s^2 = \widehat{\sigma}^2$.

Estimation de la valeur moyenne $E(Y_h)$

Plus on est loin de la moyenne, plus les écarts entre estimations successives sont grands.



Prédiction d'un intervalle

On rappelle que : $b_1 = 3,57$ et $b_0 = 62,37$ et on a $\sigma^2 = 2.384$. Alors, pour 65 débouche-trucs :

$$\widehat{Y}_h = 62,37 + 3,57(65) = 294,4$$

$$\sigma^2(\widehat{Y}_h) = (\widehat{\sigma})^2(\frac{1}{25} + \frac{(65-70)^2}{19.800}) = 98,37$$

$$\sigma = 9.918$$

Pour un intervalle de confiance à 0.90, on aura :

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

Pour 100 pièces, on aurait :

$$359,52 \leq E(Y_h) \leq 479,42$$

Analyse de la variance appliquée à la régression

On définit les grandeurs suivantes :

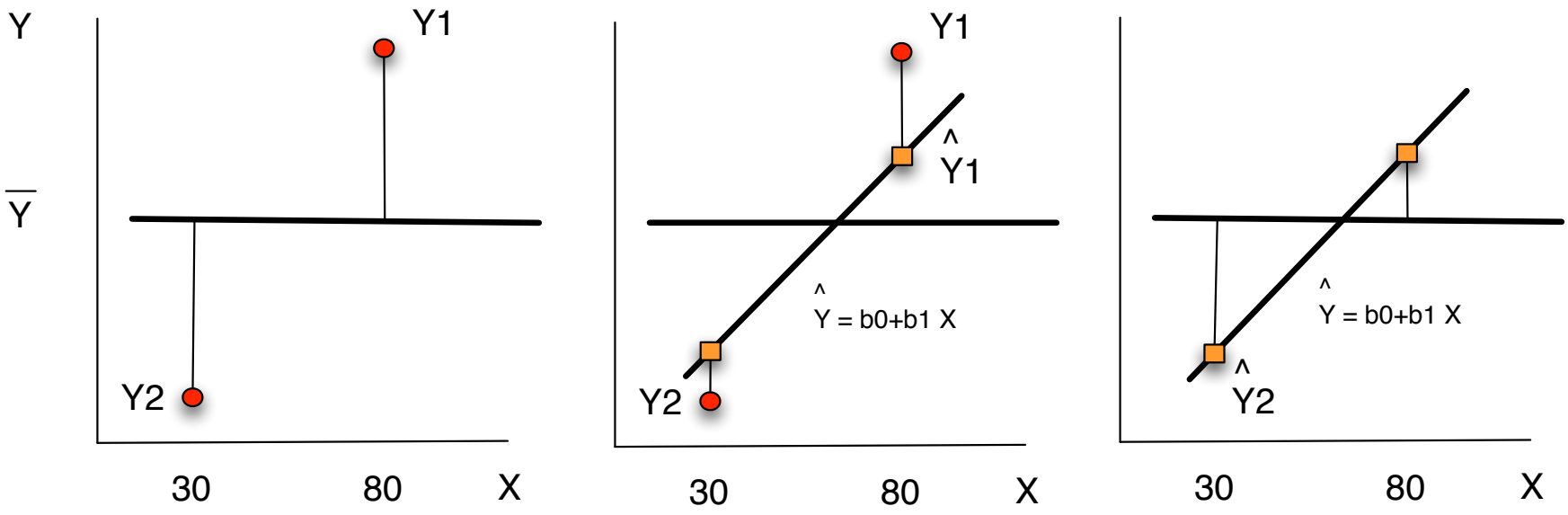
$$SSTO = \sum (Y_i - \overline{Y})^2$$

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

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

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

Analyse de la variance appliquée à la régression (II)



Relations entre sommes d'erreur

On a :

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

d'où :

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

soit

$$SSTO = SSR + SSE$$

Degrés de liberté

SSTO a $n-1$ degrés de liberté : les $Y_i - \overline{Y}$ ont une somme nulle.

SSE a $n-2$ degrés de liberté : les paramètres β_0 et β_1 sont estimés et bloquent deux degrés de liberté.

SSR a 1 degré de liberté : les $\widehat{Y}_i$ sont calculés à partir de la droite de régression, qui a deux degrés de libertés liés aux coefficients. On en perd un pour raison de somme nulle (idem SSTO).

Test de Fisher pour la régression

On va tester l'hypothèse :

$$H_0 \ : \beta_1 = 0$$

Pour cela, on constate que :

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

et

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

Si $\beta_1 = 0$, on voit que MSE et MSR doivent être du même ordre de grandeur.

Loi de Fisher-Snedecor

Si X et Y sont respectivement des lois de type χ^2_n et χ^2_p, alors on peut définir la loi suivante :

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

Cette loi sert de référence pour les analyses de variances.

Quel test pour la régression ?

Si H_0 est respectée, on doit retrouver avec F une distribution de Fisher, donc on va considérer la valeur de l'estimation de la statistique et analyser son positionnement par rapport à la valeur seuil (comme d'habitude ...). Si la statistique est inférieure à la valeur seuil $F(1-\alpha; 1, n-2)$, alors on ne peut pas rejeter H_0.

Dans le cas de l'exemple, on obtient $MSR = 252.378$ et $MSE = 2.384$, donc $F = 105$ que l'on doit comparer à la valeur seuil pour $n - 2 = 23$, soit $4,29$. Il semble que l'on ne puisse pas conserver H_0, donc que l'on ne puisse pas conserver l'hypothèse $\beta_1 = 0$, donc il existe une relation de dépendance linéaire entre X et Y.

Et dans 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 de détermination

Pour savoir quel est "l'impact" de la variable explicative X sur la détermination de Y, on peut se ramener au rapport entre SSR et SSTO. Plus ce rapport sera proche de 1, plus Y sera "gouverné" par les valeurs de X. On définit donc :

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

R^2 est compris entre 0 et 1.

Coefficient de détermination (II)

Attention !

- un coefficient R élevé ne signifie pas qu'on peut prédire correctement
- un coefficient R élevé n'indique pas forcément que la droite de régression est optimale
- un coefficient R proche de 0 ne signifie pas que X et Y sont indépendants

On a la relation suivante :

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

Limites de la régression linéaire

- la fonction de régression n'est pas linéaire
- les termes d'erreur n'ont pas une variance constante
- les termes d'erreur ne sons pas indépendants
- les modèles sont adaptés mais il y a des outliers
- le termes d'erreur ne sont pas distribués selon une loi normale
- il manque des variables explicatives

Analyse des individus extrêmes

- La régression linéaire est très sensible aux individus extrêmes, qui peuvent influencer fortement les coefficients.
- On peut étudier les individus extrêmes et leur influence par la technique du *jackknife* (analyse différentielle de la régression), les analyses de *leverage*, de *dfitts* et de *distance de Cook*.
- L'objectif est de décider si ces individus doivent être finalement écartés ("anormaux", "hors du modèle"), ou au contraire s'ils sont représentatifs.

Leverage

- La formule du *leverage* est :

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

  où h_i représente l'influence de l'individu i, $\frac{1}{n} \leq h_i \leq 2$.
- On évite généralement de conserver les individus pour lesquels $h_i \geq \frac{4}{n}$.

Dfitts

- On observe l'effet de l'absence de l'individu sur le résultat de la régression, en évaluant :

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

- On évite généralement de conserver les individus pour lesquels *dfitts* est de valeur absolue supérieure à 1 (petits échantillons) ou $2\sqrt{\frac{2}{n}}$ (autres tailles d'échantillons).

Distance de Cook

- L'effet de l'individu i peut également être évalué par :

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

- $D_{cook\ i}$ ne doit pas excéder une valeur tabulée $D_{cook\ ref}$ calculée en fonction du quantile d'une loi de Fisher.

Analyse des résidus

- Rappel : les résidus doivent être de lois normales de même variance et indépendants.
- La vérification est le plus souvent graphique. Elle est faite sur les couples $(e_{i\ std}, \widehat{Y})$, avec :

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

Transformations de linéarisation

Fonction	Transformation	Forme linéaire
$y = \alpha x^{\beta}$	$y' = Logy$ $x' = Logx$	$y' = Log + \beta x'$
$y = \alpha e^{\beta x}$	$y' = Logy$	$y' = Log\alpha + \beta x$
$y = \alpha + \beta Logx$	$x' = Logx$	$y = \alpha + \beta x'$
$y = \frac{x}{\alpha x - \beta}$	$y' = 1/x$	$y' = \alpha - \beta x'$
	$x' = 1/x$	
$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$y' = Log(\frac{y}{1-y})$	$y' = \alpha + \beta x$

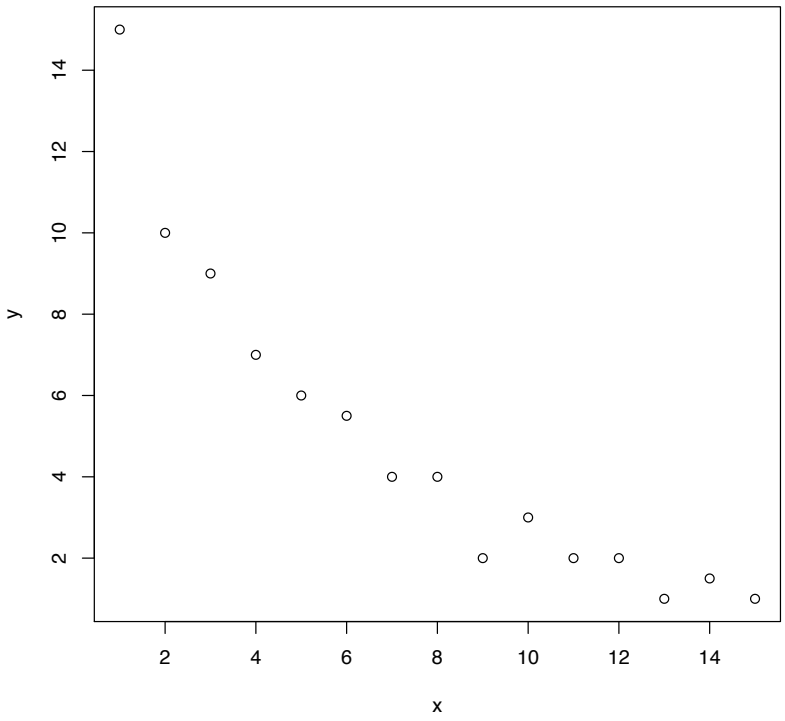
Exemple

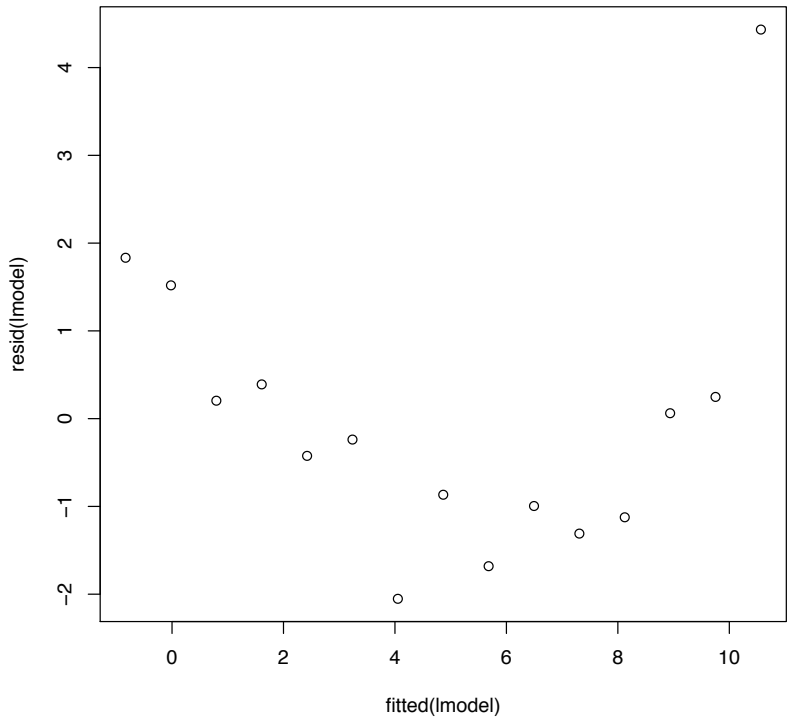
```
"x" "y"
"1" 1 15
"2" 2 10
"3" 3 9
"4" 4 7
"5" 5 6
"6" 6 5.5
"7" 7 4
"8" 8 4
"9" 9 2
"10" 10 3
"11" 11 2
"12" 12 2
"13" 13 1
"14" 14 1.5
"15" 15 1
```

Exemple

```
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
```

Exemple

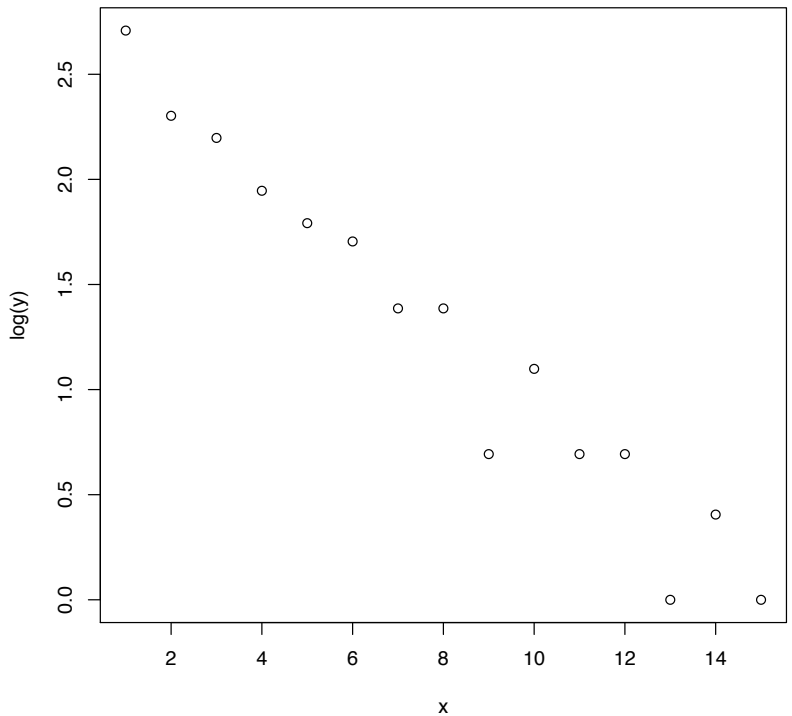


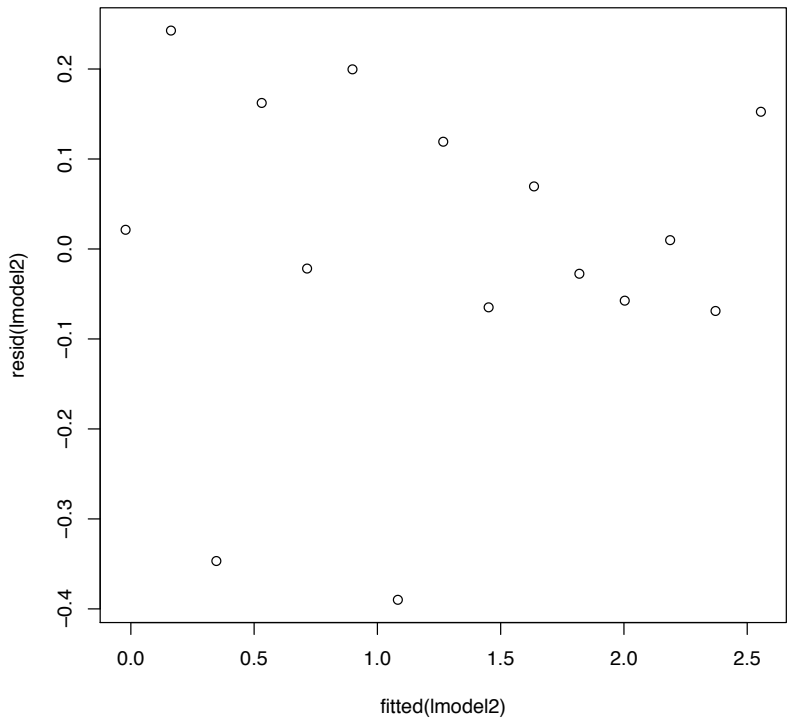


Exemple

```
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
```

Exemple


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

Transformations liées à la variance

Distribution	Variance $f(\mu)$	Transformation	Variance résultante
Poisson	μ	$\sqrt{(y)}$	0.25
Binomiale	$\frac{\mu(1-\mu)}{n}$	$Arc \sin \sqrt{(y)}$	$\frac{0.25}{n}$

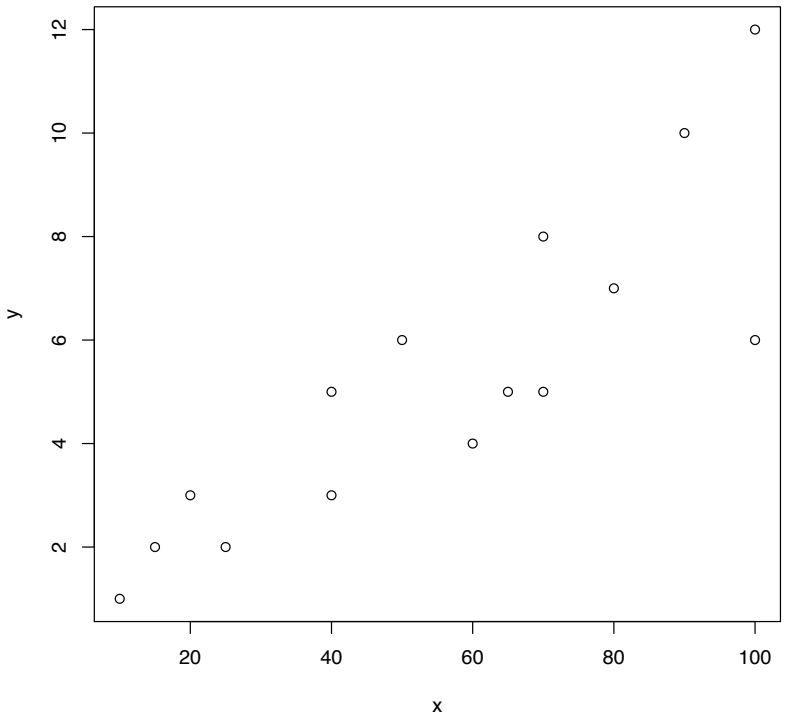
Exemple

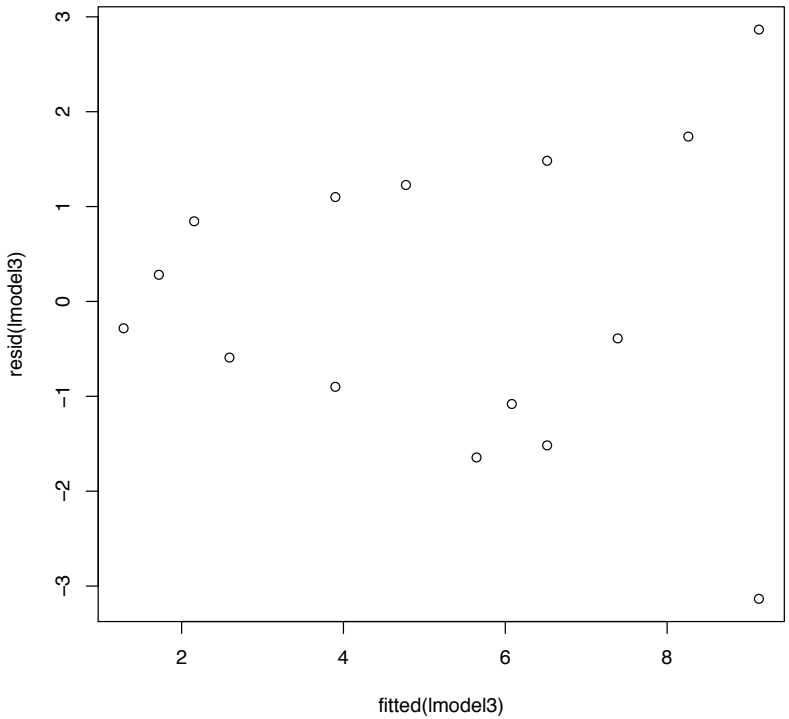
```
"x" "y"
"1" 10 1
"2" 15 2
"3" 20 3
"4" 25 2
"5" 40 3
"6" 40 5
"7" 50 6
"8" 60 4
"9" 65 5
"10" 70 5
"11" 70 8
"12" 80 7
"13" 90 10
"14" 100 6
"15" 100 12
```

Exemple

```
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
```

Exemple

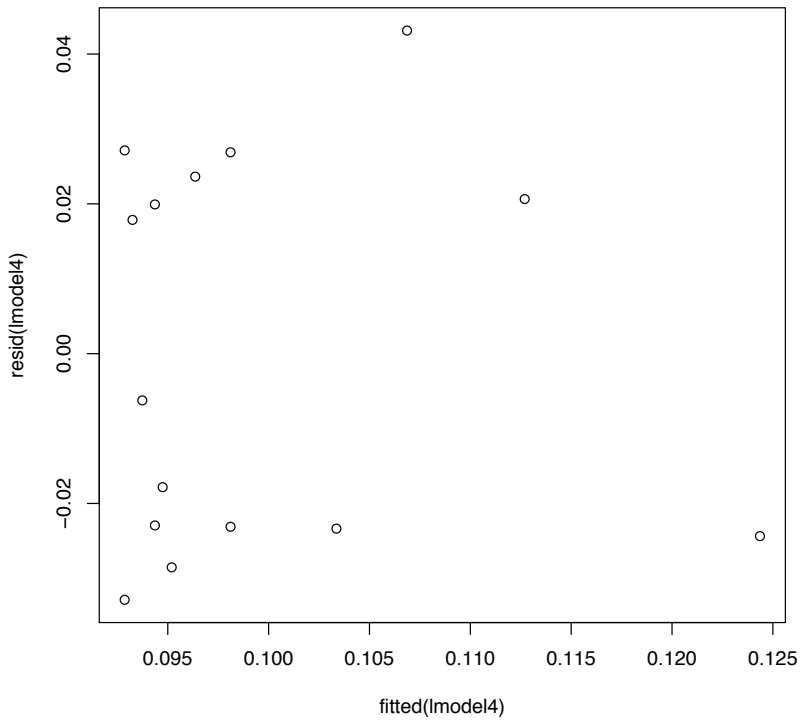

y
x
2
4
6
8
10
12
20
40
60
80
100


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

Exemple

```
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
```

Exemple

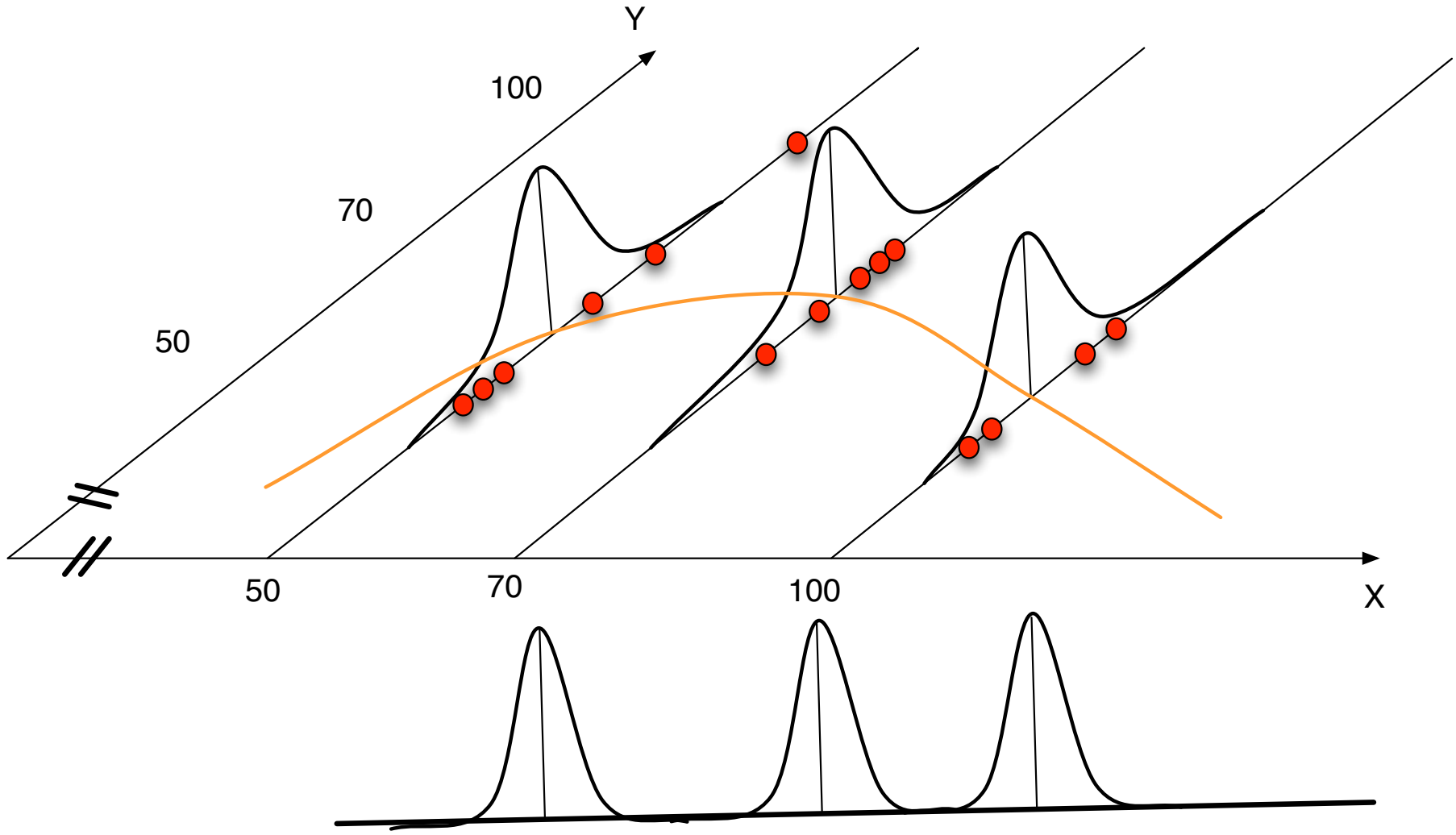

resid(lmodel4)
0.04
0.02
0.00
−0.02
0.095
0.100
0.105
0.110
0.115
0.120
0.125
fitted(lmodel4)

Analyse de variance pour ... comparaison de moyennes

On cherche à savoir quelle est l'influence de facteurs sur le comportement de populations. Ces facteurs ou variables explicatives sont des variables qualitatives, la variable observée ou expliquée est numérique. Les modalités des différents facteurs sont généralement appelées niveaux.

Analyse de variance (ANOVA) à un ou deux facteur(s).

ANOVA et régression



Modèle de l'ANOVA

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

avec $X_{ijk} = 1$ si facteur k et 0 sinon.

Formalisation du problème

On dispose de k échantillons de tailles $n_1, n_2, .., n_k$ correspondant aux différentes modalités du facteur $A_1, A_2, .., A_k$.

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

On veut savoir si $H_0 : m_1 = m_2 = ... = m_k$

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

α_i effet du niveau i du facteur et ϵ_i^j de loi $N(0, \sigma)$.

Décomposition de la variance

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

avec $Y_i^j - \overline{Y} = Y_i^j - \overline{Y}_i + \overline{Y}_i - \overline{Y}$, on obtient :

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

soit

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

Comparaison de variances

En écrivant $SSR = \frac{1}{n}\sum_{i=1}^{k} n_i SSR_i$, on obtient que :

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

Sachant que $\frac{n_i SSR_i}{\sigma^2}$ est une variable de type $\chi^2_{n_i-1}$, la variable $\frac{nSSR}{\sigma^2}$ est elle de type χ^2_{n-k}.
De même, la variable $\frac{nSSE}{\sigma^2}$ est une variable de type χ^2_{k-1}.

Loi de Fisher-Snedecor pour la comparaison

Si les moyennes étaient identiques (H_0 : les niveaux des facteurs n'ont pas d'influence), on devrait avoir une influence identique entre les effets intragroupe (SSR) et les effets intergroupe (SSE). Ceci devrait donc se traduire par un rapport de 1 entre les deux quantités, soit :

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

La loi F st une loi de Fisher-Snedecor (à $k-1$ et $n-k$ degrés de liberté). Si la valeur obtenue pour F est trop extrême, on réfutera l'hypothèse de moyennes identiques.

Analyse de variance : exemple

On cherche à savoir s'il existe des différences entre les taux de taxe d'habitation en fonction de la région. On a le tableau suivant :

zone	effectif	moyenne	variance
centre	13	4.38	3.63
est	10	17.66	4.38
idf	26	11.76	15.04
nord	9	25.95	50.40
ouest	14	18.89	9.59
sud-est	18	19.76	8.63
sud-ouest	10	20.51	20.69

ANOVA : exemple (II)

On calcule :

- la variance inter-groupes $SSE = 1706$
- la variance intra-groupe $SSR = 1320$

Avec $k - 1 = 6$ et $n - k = 93$, on a $df = 99$ et $F = 20.03$

ANOVA : exemple (II)

On calcule :

- la variance inter-groupes $SSE = 1706$
- la variance intra-groupe $SSR = 1320$

Avec $k - 1 = 6$ et $n - k = 93$, on a $df = 99$ et $F = 20.03$

Ce qui exclut une égalité des moyennes

Contrastes

On peut quand même vouloir savoir si au sein des moyennes, on a des paires identiques ($m_i = m_j$). Pour cela, on s'appuie sur la formule de Scheffé qui indique que

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

a lieu avec une probabilité

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

On calcule $S = \sqrt{(k-1)F_\alpha(k-1;n-k)}$ et si
$\mid x_i - x_j \mid > S\overline{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$ alors les moyennes m_i et m_j sont différentes.