La régression linéaire

Par Yves DEMUR yves.demur@m4am.net

Version 2.3 du 2 janvier 2008

Disponibilité des versions sur http://yves.demur.free.fr avec des programmes de test en C

Comprendre un minimum la régression linéaire pour bien utiliser ses formules.

Table des matières

1 Résumé

La régression linéaire est un outil (mathématique) statistique qui permet de définir une loi linéaire entre deux variables intervenant dans un même phénomène. Les calculettes et tableurs proposent des formules prêtes à être utilisées mais elles ne conviennent qu'à un type de cas. La méthode statistique est basée sur le fait que le cumul d'erreurs s'annule « en moyenne ». Il faut donc savoir exprimer où se situe l'erreur pour appliquer l'effet statistique au bon endroit.

Le coefficient de corrélation est un indicateur sur la qualité des données. Il est dégradé par la dispersion des informations autour de la tendance. Il peut être utilisé pour calculer un autre indicateur qui évoque cette dispersion de manière plus palpable, à travers le coefficient de forme d'une ellipse.

2 Introduction

Lors d'expérimentations, lorsque l'on souhaite extraire les coefficients d'une loi linéaire, on peut utiliser la méthode dite « graphique ». On prend alors une feuille de papier millimétré et on y reporte les points de mesure. La relation linéaire s'exprime par le fait que l'ensemble des points prend la forme d'un nuage allongé relativement rectiligne. Ensuite on positionne une règle transparente de manière telle qu'il y ait autant de points d'un coté du bord de la règle que de l'autre. Après avoir tracé le trait, on y relève les coordonnées de deux points éloignés pour déduire l'équation de la droite et donc les coefficients de la loi linéaire recherchée.

Cette méthode dépend de la précision opératoire, liée à la qualité du papier et à celle de l'opérateur. Nous verrons plus loin que même avec une précision extrême, elle n'est pas la mieux adaptée dans tous les cas.

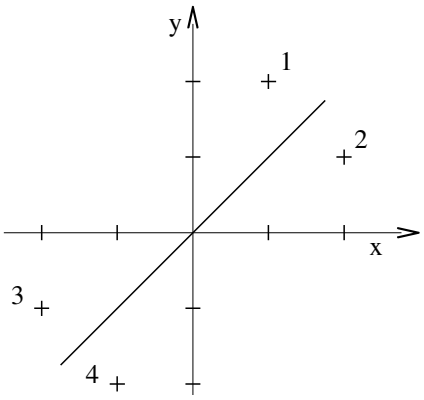
Pour lever les problèmes de mode opératoire, les outils statistiques permettent une précision bien plus rigoureuse. Mais la méthode graphique garde certains avantages :

- mise en évidence des points trop éloignés du nuage et donc à écarter,
- mise en évidence de l'aspect rectiligne du nuage de points (une courbure montre qu'une des variables, voire les deux, nécessite une fonction d'ajustement).

3 Étude d'un cas théorique simple

3.1 Problème

Considérons quatre points, représentant des couples de données (x, y), répartis symétriquement autour de la première bissectrice du repère (droite d'équation $y = x$) : $1\begin{bmatrix}1\\2\end{bmatrix}$ $2\begin{bmatrix}2\\1\end{bmatrix}$ $3\begin{bmatrix}-2\\-1\end{bmatrix}$ $4\begin{bmatrix}-1\\-2\end{bmatrix}$



Le calcul de régression linéaire préprogrammé (de calculette ou tableur), nous donne $y = 0.8\,x$ et $r = 0.8$.

En entrant x à la place de y et y à la place de x dans un nouveau calcul, on trouve $x = 0.8\,y$ et $r = 0.8$, d'où $y = 1.25\,x$.

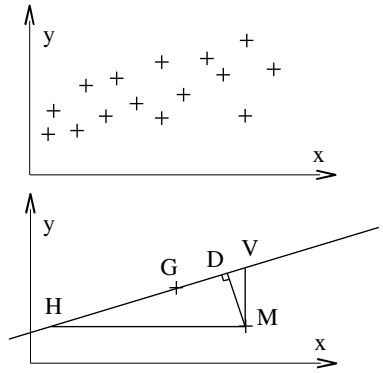
Problème : à partir de cette distribution de points, le calcul standard

- ne donne pas la droite qui intuitivement paraît la meilleure (la première bissectrice du repère),
- donne deux solutions différentes et un coefficient r inutile.

Il faut donc remonter à la source du calcul pour y voir plus clair.

3.2 Principe du calcul

On veut obtenir la « meilleure » équation de droite $y = ax + b$ à partir d'un nuage de n points.



Pour tracer la droite traduisant la tendance moyenne de corrélation entre x et y il suffit de connaître un point de cette droite et la pente.

L'optimisation sur b fait prendre comme point le barycentre : $G\begin{bmatrix}\bar{x} = \dfrac{\sum x}{n}\\ \bar{y} = \dfrac{\sum y}{n}\end{bmatrix}$

La pente a est calculée suivant un critère d'optimisation. On aura alors $b = \bar{y} - a\,\bar{x}$

Posons : $\overline{x^2} = \sum x^2/n \quad \overline{y^2} = \sum y^2/n \quad \overline{xy} = \sum xy/n$

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2 \quad \sigma_y^2 = \overline{y^2} - \bar{y}^2 \quad cov(x, y) = \overline{xy} - \bar{x}\,\bar{y}$$

Le coefficient de corrélation est défini par $r = \dfrac{cov(x, y)}{\sigma_x \times \sigma_y}$. Il caractérise le regroupement « en ligne » des points.

Pour calculer les a on fera judicieusement un changement d'origine du repère vers le point G.

On a le choix du critère d'optimisation permettant de calculer a :

1 Minimiser $\sum MV^2$ méthode la plus utilisée dite des moindres carrés, préprogrammée dans les calculettes et tableurs

$MV=|ax+b-y|$ On trouve $a_1=\frac{cov(x,y)}{\sigma_x^2}$

2 Minimiser $\sum MH^2$ $a_2=\frac{\sigma_y^2}{cov(x,y)}$

3 Minimiser $\sum MH\times MV$ (double aire HMV) $a_3^2=\frac{\sigma_y^2}{\sigma_x^2}$ $|a_3|=\frac{\sigma_y}{\sigma_x}$

Le signe est déduit de $cov(x,y)$ d'où $a_3=\frac{cov(x,y)}{|cov(x,y)|}\frac{\sigma_y}{\sigma_x}$

4 Minimiser $\sum MD^2$ $a_4=\frac{\sigma_y^2-\sigma_x^2+\sqrt{\left(\sigma_y^2-\sigma_x^2\right)^2+(2\mathrm{cov}(x,y))^2}}{2cov(x,y)}$

La méthode à suivre consiste donc à obtenir $n, \sum x, \sum y, \sum x^2, \sum y^2, \sum xy$ puis déduire $\bar{x}, \bar{y}, \sigma_x, \sigma_y, cov(x,y)$ et ensuite calculer a selon le critère choisi et enfin déduire $b = \bar{y} - a\bar{x}$.

Remarques

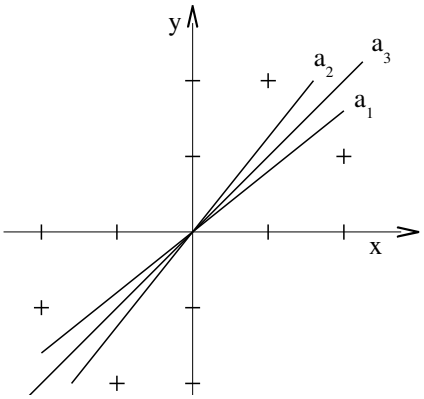
1) $\sqrt{a_1 a_2} = |a_3| = \frac{a_1}{r} = r\,a_2$ $\frac{a_1}{a_2}=r^2$ $r\leqslant 1$ a_1 et a_2 encadrent donc a_3

2) Si l'on fait un changement d'échelle $y'=ky$ on trouve $a'_1=ka_1$ $a'_2=ka_2$ $a'_3=ka_3$ $a'_4\neq ka_4$

Le coefficient a_4 dépend donc de l'échelle ; on remarque que la distance MD n'a aucune réalité physique si x et y ne sont pas exprimées avec la même unité. Son optimisation ne représente alors rien de rigoureux.

3) Pour forcer la droite à passer par l'origine (et obtenir $b=0$), on peut doubler chaque point par son symétrique par rapport à l'origine. Le calcul des $\sigma'_x, \sigma'_y, cov'(x,y)$ sur les $2n$ points donnera les mêmes valeurs que si on les avait calculées sur les n points originels mais en forçant $\bar{x}=\bar{y}=0$.

3.3 Application



Dans notre cas nous obtenons

$\bar{x}=\bar{y}=0$ $\sigma_x^2=\sigma_y^2=2.5$ $\overline{xy}=cov(x,y)=2$

$a_1=0.8$ $a_2=1.25$ $a_3=a_4=1$ $r=0.8$

Le coefficient a_1 , fourni par les calculs préprogrammés ne nous semble pas plus pertinent qu'un autre.

Il nous faut donc d'autres informations pour choisir notre coefficient a .

4 Signification du coefficient de corrélation

4.1 Étude de l'ellipse

On suppose une répartition de points de mesure uniforme formant une ellipse (répartition homogène elliptique, voir Fig 1). Les calculs de sommes sont alors remplacés par des intégrales de surface, mais les principes statistiques restent les mêmes.

Dans toute la suite on suppose $cov(x,y)$ positif.

On trouve : - droite de pente a_4 = grand axe de l'ellipse (axe de moment d'inertie minimal),

- droite a_1 = passe par le point de tangente verticale, droite a_2 par le point de tangente horizontale,

- droite a_3 = passe par l'intersection des deux tangentes.

$$\sigma_x^2=\frac{b}{4a}(a^2\cos^2\theta+b^2\sin^2\theta) \quad \sigma_y^2=\frac{b}{4a}(a^2\sin^2\theta+b^2\cos^2\theta) \quad cov(x,y)=\frac{b}{4a}(a^2-b^2)\cos\theta\sin\theta$$

$$r=\frac{cov(x,y)}{\sigma_x\sigma_y}=\frac{surface\,CFDG}{surface\,APBG}$$; F=foyer $GF^2=a^2-b^2$ a_3 et a_4 confondues si $\theta=\frac{\pi}{4}$

Remarque

Par la méthode graphique sur un nuage de points classique mais de forme elliptique, le coup d'oeil nous entraîne à tracer la droite a_4 puisque c'est celle qui est le mieux centrée sur l'ellipse. Comme a_4 dépend de l'échelle, on voit que la méthode graphique donne un résultat qui dépend des échelles du graphe, d'après une des remarques du paragraphe présentant le principe du calcul. Donc, la façon dont le tracé a été fait influe sur le résultat par le choix du papier et des échelles dans un premier temps, puis par la précision du travail graphique.

4.2 Ellipse équivalente

Pour toute répartition quelconque de points on peut définir une ellipse homogène équivalente ayant mêmes écarts types même covariance et même centre de gravité.

Les droites a_1 a_2 a_3 et a_4 seront alors les mêmes.

Connaissant σ_x et σ_y et r nous allons essayer de tracer cette ellipse.

Traçons d'abord l'ellipse réduite. Elle est l'ellipse équivalente de la distribution x, y' avec $y'=\frac{\sigma_x}{\sigma_y}y$ (voir Fig 2). Elle est centrée sur le centre de gravité des points, la pente de son grand axe est 1. Les droites a'_3 et a'_4 sont confondues et ont pour pente 1, on a aussi $r=a'_1=1/a'_2$.

Par le calcul on trouve que cette ellipse homogène réduite équivalente est inscrite dans un carré de demi côté $2\sigma_x\left(\frac{1+r}{1-r}\right)^{1/4}$, qu'elle a pour demi grand axe $a'=2\sigma_x(1+r)(1-r^2)^{-1/4}$, pour demi petit axe $b'=2\sigma_x(1-r^2)^{1/4}$, pour distance focale $GF'=2\sigma_x\left(\frac{4r^2(1+r)}{1-r}\right)^{1/4}$.

L'ellipse homogène équivalente se déduit de l'ellipse réduite par l'affinité d'axe Gx, de direction y et de rapport $\frac{\sigma_y}{\sigma_x}$.

Cette transformation déforme l'ellipse et son grand axe a_4 se sépare de la droite a_3 (voir Fig 1).

La visualisation d'une telle ellipse en surimpression sur le nuage de points permet de se rendre compte de la qualité de la distribution des points en appréciant l'allongement de l'ellipse. Cet allongement, caractérisable par le rapport grand axe / petit axe, peut remplacer le coefficient r comme indicateur.

4.3 Lien entre r et les proportions de l'ellipse

Le coefficient r est directement lié au rapport $u=\frac{a'}{b'}$ de l'ellipse réduite puisque le calcul donne $u=\left(\frac{1+r}{1-r}\right)^{1/2}$.

L'allongement de l'ellipse réduite ne dépend pas des échelles de représentation de x et y contrairement à celui de l'ellipse mise à l'échelle en y . Il est donc plus représentatif de la distribution que le rapport de l'ellipse non réduite.

Le coefficient u permet, comme r mais de manière plus visible, de caractériser la dispersion autour de la tendance.

r	1	0.995	0.99	0.95	0.9	0.8	0.7
r²	1	0.99	0.98	0.902	0.81	0.64	0.49
u		20	14.1	6.25	4.36	3	2.38
1/u	0	5%	7%	16%	23%	33%	42%
1-r	0	0.5%	1%	5%	10%	20%	30%

Dans notre cas théorique simple, l'obtention d'un coefficient de corrélation de 0.8 ne nous évoque rien tandis qu'un rapport u de 3 nous dit que l'ellipse est peu plate (bien ventrue !) et donc que la dispersion autour de la tendance est relativement importante.

5 Choix du critère d'optimisation

5.1 Situation de la dispersion

Dans le paragraphe qui présente le principe du calcul, nous avons vu qu'il existe plusieurs possibilités d'optimisation pour déduire la pente de la droite.

La méthode statistique s'appuie sur le fait qu'une dispersion s'annule en moyenne. Il faut donc savoir d'où est issue la dispersion qui nous intéresse (et que l'on souhaite annuler, si l'on a suffisamment de points).

Dans notre phénomène physique étudié, les sources de dispersion sont multiples :

- incertitudes sur la mesure de x ,
- incertitudes sur la prise en compte de x dans le phénomène linéaire,
- (le phénomène linéaire lui même est supposé parfait),
- incertitudes sur la prise en compte de y dans le phénomène linéaire,
- incertitudes sur la mesure de y .

Si au lieu de tracer des points on traçait des rectangles d'incertitude pour chaque relevé, l'allongement de ces rectangles nous indiquerait quelle est l'incertitude qui a le plus d'influence. C'est elle qu'il faut annuler « en moyenne » en priorité.

Des rectangles très allongés dans le sens des y feront utiliser le critère a_1 .

Des rectangles très allongés dans le sens des x feront utiliser le critère a_2 .

Des rectangles relativement équilibrés feront préférer le critère a_3 .

Le choix est guidé par la mise en évidence de la dispersion et de l'endroit où elle est générée.

On peut aller plus loin en remarquant que a_3 est la moyenne géométrique de a_1 et a_2 et donc se lancer dans une pondération $a=\left(a_1^p \times a_2^q\right)^{1/(p+q)}$ en évaluant les poids relatifs p et q à partir des proportions du rectangle d'incertitude.

Cette pondération a été testée dans un programme qui utilise les dispersions en x et y pour calculer p et q . Par tâtonnement, j'ai fini par trouver des valeurs qui aboutissent à un résultat pertinent, mais sans justification théorique :

$p=(\Delta y/\sigma_y)^{2.5}$ et $q=(\Delta x/\sigma_x)^{2.5}$

5.2 Cas de choix

Les exemples évoqués ci dessous sont des cas où deux variables relevées sont liées par une loi que l'on « sait » linéaire et que l'on souhaite déterminer « en moyenne ». C'est par l'analyse de ces exemples dans leur réalité que l'on arrive à décider quel critère choisir. J'ai volontairement pris des exemples aboutissant à des choix différents, du point de vue de la régression linéaire et de la notion de « en moyenne ».

5.2.1 Exemple théorique

Dans le cas de l'exemple théorique vu plus haut, nous n'avons aucune information sur l'origine de la dispersion. Il est donc impossible de décider quel critère d'optimisation choisir pour annuler « en moyenne » cette dispersion. Le critère a_3 semble celui le plus « moyen ». Si le coefficient de corrélation était bon (l'ellipse équivalente réduite aurait un rapport de longueurs bien éloigné de 1), les critères donneraient un résultat quasiment identique.

5.2.2 Concentration et opacité

Un verre d'eau dans lequel on a versé du café filtrera d'autant plus la lumière qu'il contient du café. On peut donc mesurer les opacités en utilisant diverses concentrations. L'ensemble des couples de mesure (concentration, opacité) peut être utilisé pour calculer la loi reliant ces deux grandeurs. On peut alors créer un appareil mesurant la qualité d'un café en mesurant son opacité.

Dans ce cas, on peut supposer que la concentration est bien maîtrisée tandis que la traversée de lumière (pour la mesure) et le relevé de l'opacité sont sujets à dispersion. Nous retenons donc le critère a_1 .

5.2.3 Température extérieure et consommation de chauffage

On relève tous les jours la température moyenne annoncée par Météo-France pour sa région ainsi que la consommation de gaz utilisé pour le chauffage de sa maison. On souhaite déterminer la loi reliant ces deux grandeurs. La connaissance de cette relation permettra de faire des prévisions de consommation à partir des courbes d'évolution moyenne des températures pendant l'hiver.

Dans ce cas, nous évaluons que l'incertitude la plus important porte sur la température extérieure effective subie par la maison. On peut même ajouter l'effet d'autres phénomènes climatiques (vent, humidité) qui influent sur le transfert thermique extérieur de la maison, ce qui aboutit à la notion de température efficace. La dispersion sur le volume de gaz (précision du compteur, dispersion sur la quantité d'énergie apportée par une unité de volume de gaz) est considérée comme faible devant l'erreur entre la température extérieure efficace et celle annoncée par Météo-France. Nous retenons donc le critère a_2 .

5.2.4 Jet de graviers

Une personne placée derrière une clôture a jeté une poignée de graviers sur une terrasse. La zone d'éparpillement des graviers a une forme allongée. On suppose que cet allongement est dans la direction du lancer. A partir du relevé de la position de chaque gravier, on veut déterminer à quel endroit de la clôture se situait le lanceur. Le relevé se fait en coordonnées cartésiennes à partir d'un référentiel où l'axe y est la clôture. Les couples de valeurs sont donc les coordonnées (x, y) de chaque gravier.

Dans ce cas, nous avons un besoin de détermination de direction géométrique à partir de la répartition des points. Si leur nuage avait la forme d'une ellipse, il faudrait retenir son grand axe et c'est par conséquent le critère a_4 qui nous le donne. On remarque que ce critère présente l'avantage d'être indépendant de l'orientation du repère pris pour le relevé (le grand axe est lié à la forme de l'ellipse uniquement). Cela conforte notre choix par rapport aux autres critères.

6 Conclusion

La régression linéaire ne peut s'appliquer qu'aux phénomènes pouvant se traduire par une loi du type $g(u,v)=a\,f(u,v)+b$

La méthode graphique présente l'avantage de montrer les points aberrants et permet de les éliminer, mais a l'inconvénient de donner un résultat qui dépend de l'opérateur et des échelles. Elle permet aussi de tenir compte des incertitudes en traçant le rectangle de largeur $2\Delta x$ et de hauteur $2\Delta y$ autour de chaque point.

La méthode statistique, dans le cas d'un r proche de 1 (i.e. d'une distribution bien allongée) donne des résultats précis et peu dépendants du critère utilisé. Dans le cas d'une dispersion sensible dans le nuage de points $(r^2<0.98)$ il est préférable de se méfier des calculs préprogrammés et de se concentrer sur la (les) source(s) de la dispersion afin de bien choisir le critère statistique d'optimisation. Une recherche des points aberrants est aussi opportune.

Les dénominations de « variable déterminante » et « variable déterminée » rencontrées dans certaines présentations sur la régression linéaire ne justifient nullement pourquoi il faut utiliser le critère a_1 . La seule explication (non exprimée) serait que toutes les dispersions se reportent sur la variable déterminée. Or l'exemple du chauffage d'une maison nous montre bien que la dispersion peut être partout et que le rôle des variables n'est pas à prendre en considération : la température efficace, variable déterminante, cumule des dispersions bien plus grandes que la consommation de gaz, variable déterminée. Dans cet exemple il faut bien sûr utiliser le critère a_2 si l'on veut rester cohérent avec le principe statistique.

7 Figures

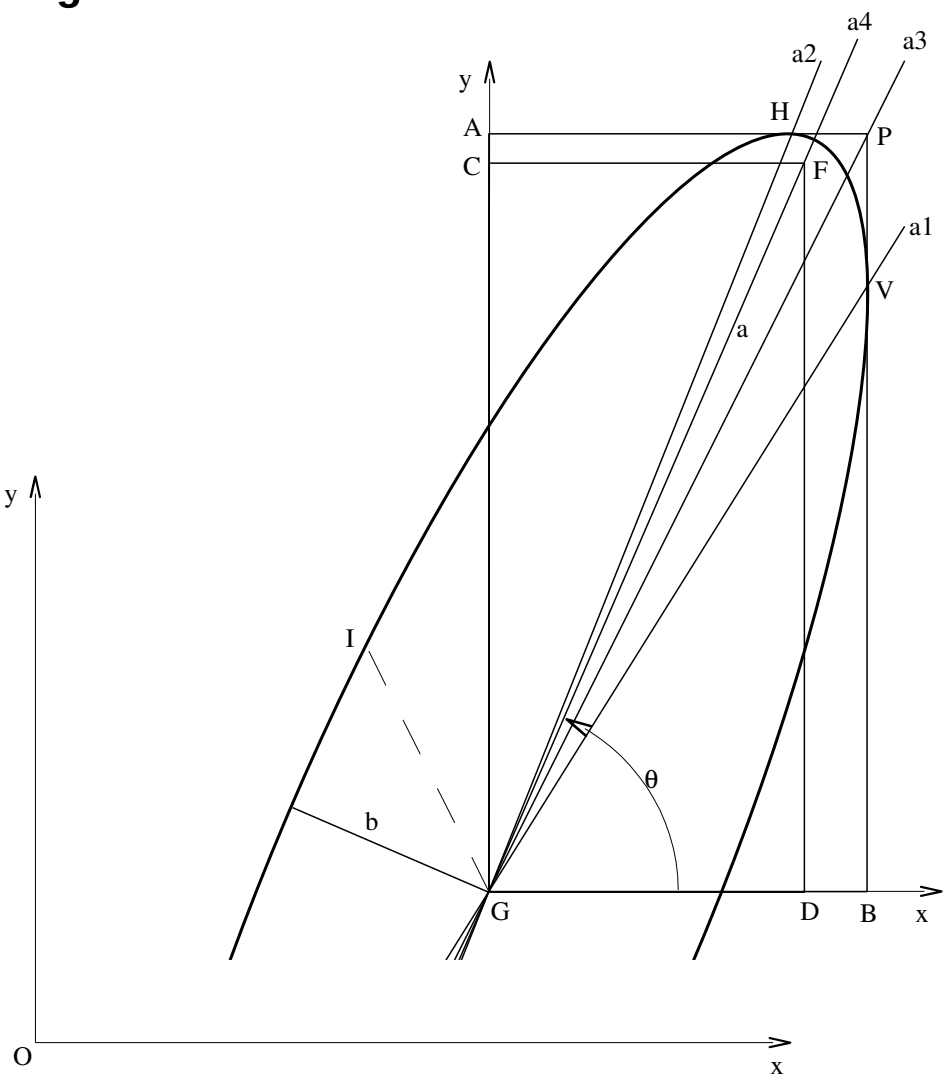


FIG 1
Ellipse équivalente

$r = 0.8$

$\sigma_x^2 = \frac{6.25}{3}\, cm^2$

$\sigma_y^2 = \frac{25}{3}\, cm^2$

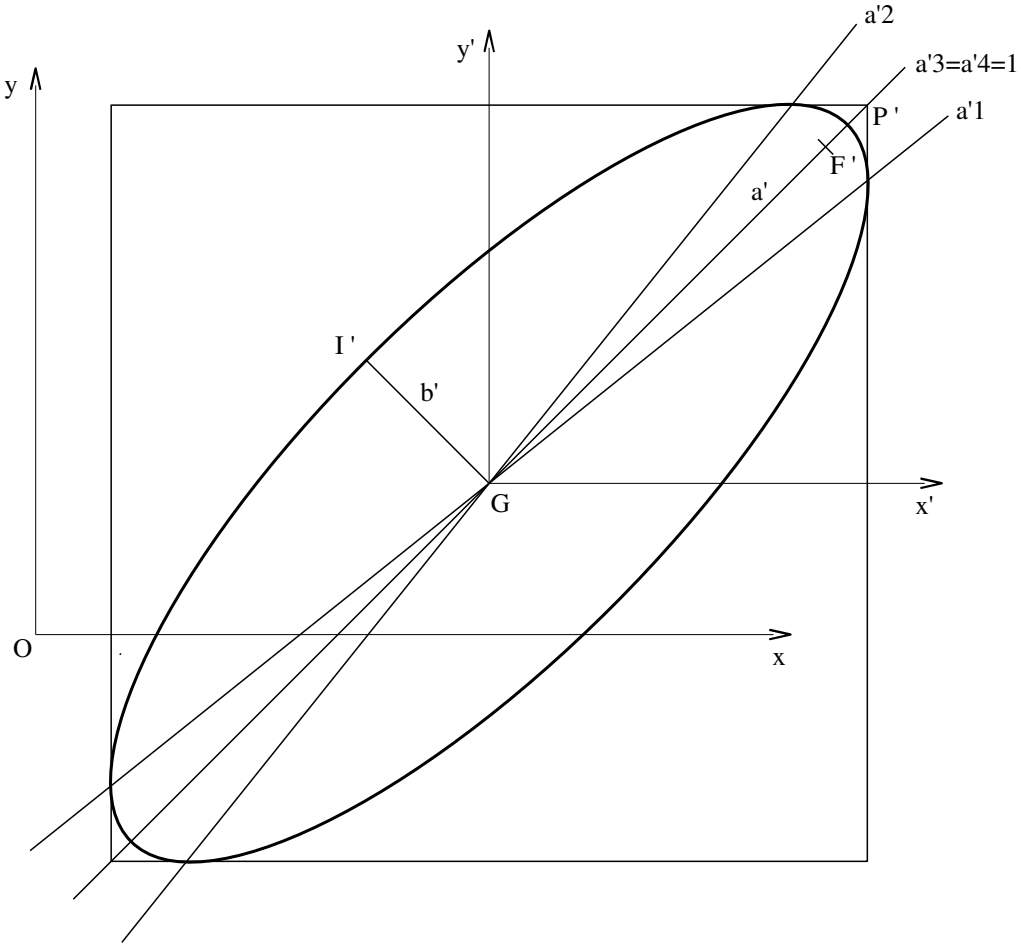


FIG 2
Ellipse réduite

$a' = 3\sqrt{5}\ \ cm$

$b' = \sqrt{5}\ \ cm$

$u = 3$