Courbe de Lorenz et indice de Gini.

Grâce aux très nombreuses statistiques à notre disposition, on dispose d'un grand nombre de bases. Naturellement, le but, et même le seul intérêt, est d'en tirer le maximum d'informations et avoir les outils nécessaires pour comparer les résultats dans différents contextes.
On cherche ici à savoir si telle variable statistique est indépendante ou non de telle autre. Puis, en comparant différents contextes, toutes choses égales par ailleurs, si la dépendance de deux variables a un impact plus important dans certains contextes que dans d'autres.

Pour clarifier les idées, prenons l'exemple de l'étude de l'impact de la richesse sur le rachitisme des enfants. Dans un pays, on a établi une liste d'enfants avec les informations suivantes

1- Richesse
2- Rapport de taille (test de malnutrition)
3- Rang de naissance dans la famille
4- Age de la mère
5- Poids (critère peu précis)
6- Age de l'enfant en mois
7- Région de résidence : de 'a' à 'f'
8- Religion : de 'a' à 'c'
9- Habitat rural : 'Y' ou 'N'
10- Activité des parents : de 'a' à 'd'
11- Education des parents : de 'a' à 'd'
12- Naissance assistée ou non : 'Y' ou 'N'
13- Mère mariée : 'Y' ou 'N'

On s'est fixé pour but d'évaluer l'impact de ces critères sur la santé de l'enfant. En termes statistiques, cela revient à évaluer la non-indépendance de telle variable par rapport à la taille.

La courbe de Lorenz et l'indice de Gini.

La courbe de Lorenz est l'outil permettant d'apprécier l'équité en matière de richesse. L'indice de Gini est un nombre situé entre 0 et 1 qui est le rapport entre l'aire comprise entre la courbe de Lorenz et l'aire du demi carré du graphique. La documentation sur ces outils est abondante et complète. Il ne paraît pas utile de rentrer plus dans le détail.

Méthode d'étude de l'interdépendance de deux variables.

Soit un contexte C1. On a un grand nombre d'observations de type A et B. On cherche à établir une correspondance entre A et B. C'est à dire "peut-on conclure que la situation B est directement liée à la situation A ?".
Soit un autre contexte C2, toutes choses étant égales par ailleurs. On peu se poser la même question.
On calcule le coefficient d'interdépendance de l'implication A ==> B pour le contexte C 1 et pour le contexte C2.

Si on calcule la courbe de Lorenz en fonction de la variable A, on obtient une répartition du type "équitable". On peut comparer cette courbe entre les contextes (pays) C1 et C2.
La même chose pour la variable B.

On pourra en déduire des affirmations du genre "le rachitisme est plus équitable dans le contexte C1 que dans C2", " la répartition de richesse est plus équitable dans le contexte C2 que dans C1" etc. Cependant l'information cherchée est du type "quelle est l'impact de la variable A sur la situation B ?".
La seule méthode possible est de comparer les degrés d'implication.

C'est là que le dessin de courbes de Lorenz peut être un outil efficace. Si pour le contexte C1 et pour le contexte C2 on dessine sur un même graphique la courbe de la situation A et la courbe de situation B, quelles que soient les "hauteurs" de ces courbes, on pourra affirmer ou non que, dans ces deux contextes, l'implication A ==> B est vraie dans les mêmes proportions.

L'indice de Gini permet en plus de numériser cela. L'indice de Gini est homogène à une aire. Dans le contexte C1, on peut calculer l'indice qui correspondrait à la différence entre l'indice de la variable A et celui de la variable B. Cet indice peut être négatif ou positif. Un indice proche de 0 indique une forte corrélation A ==> B. Si cet indice est du même ordre pour le contexte C1 et pour le contexte C2, alors on pourra affirmer que l'implication A ==> B est vérifiée, indépendamment des contextes. Ce constat étant fait, on pourra étudier dans lequel des deux contextes C1 et C2 il vaut mieux agir en priorité.

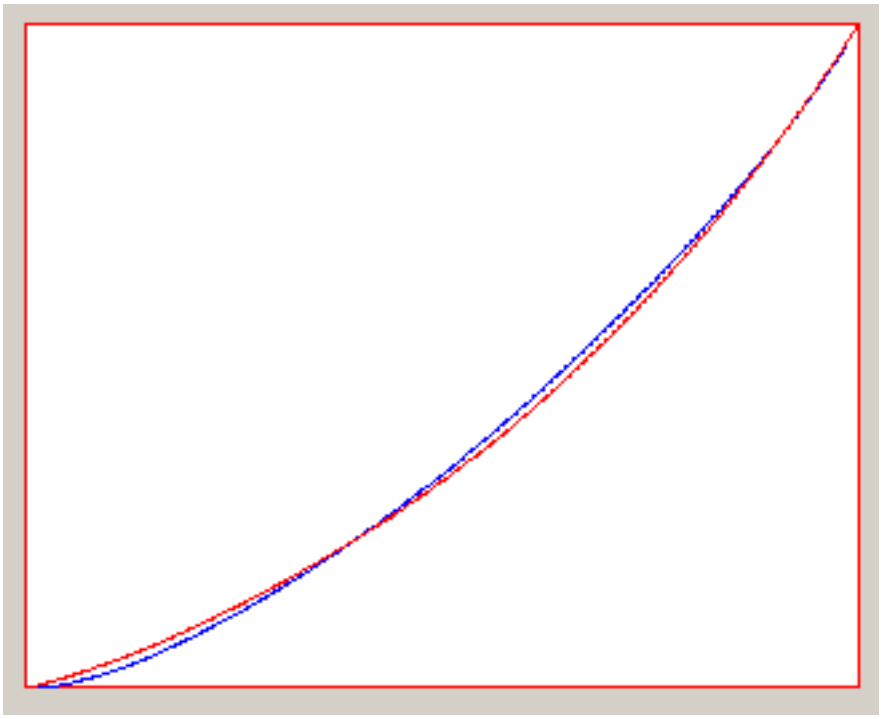
Il faut bien distinguer la comparaison des courbes ou des indices pour une même variable, ce qui est le but premier de la méthode de Lorenz, et la comparaison de la différence de deux courbes ou d'indice entre deux variables suivant les contextes.

Reprenons l'exemple de l'état de santé d'enfants.
Dans un pays C1, on constate une superposition quasi-parfaite de la courbe de richesse et de la courbe de santé. On en déduira que l'état de santé des enfants est directement lié à la répartition des richesses.
On fait le même type de statistique et de calcul dans un pays C2. Là, on constate une différence importante entre la courbe de richesse et la courbe de santé.
La conclusion est évidente, le pays C1 est pauvre et la répartition des richesse influe directement sur la qualité de santé des enfants. Par contre, dans le pays C2, il n'en est rien, les deux variables richesse et santé sont indépendantes.

L'intérêt de cette méthode est d'être très simple à mettre en œuvre, ce qui garanti sa fiabilité et très facile à interpréter, d'abord visuellement grâce aux courbes de Lorenz, puis par le calcul, par comparaison numérique des indices de Gini.

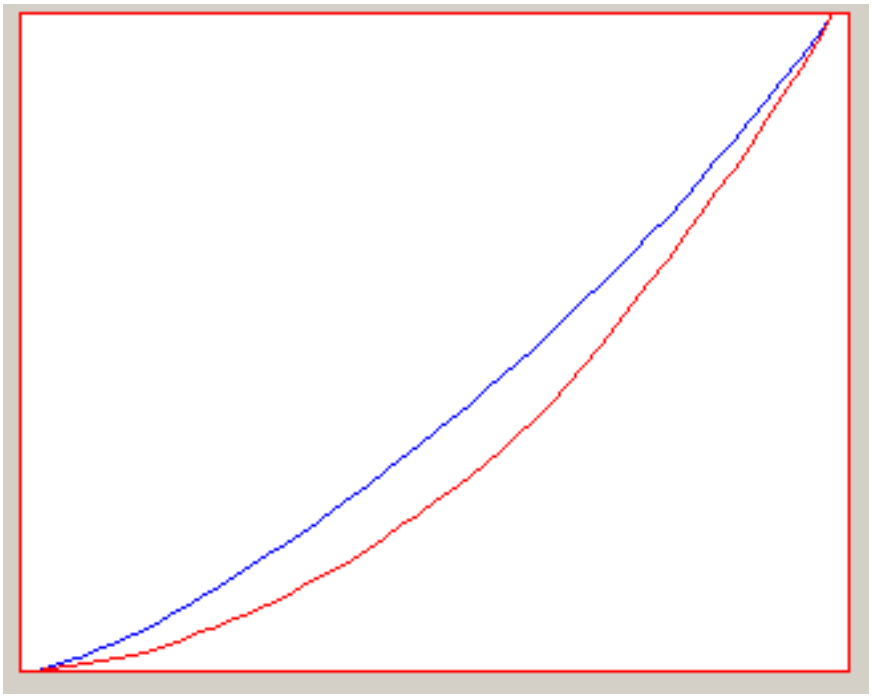
Dans les graphiques qui suivent, les mêmes variables ont été étudiées, dans le contexte C1 et dans le contexte C2.

Exemple de répartition contexte C1



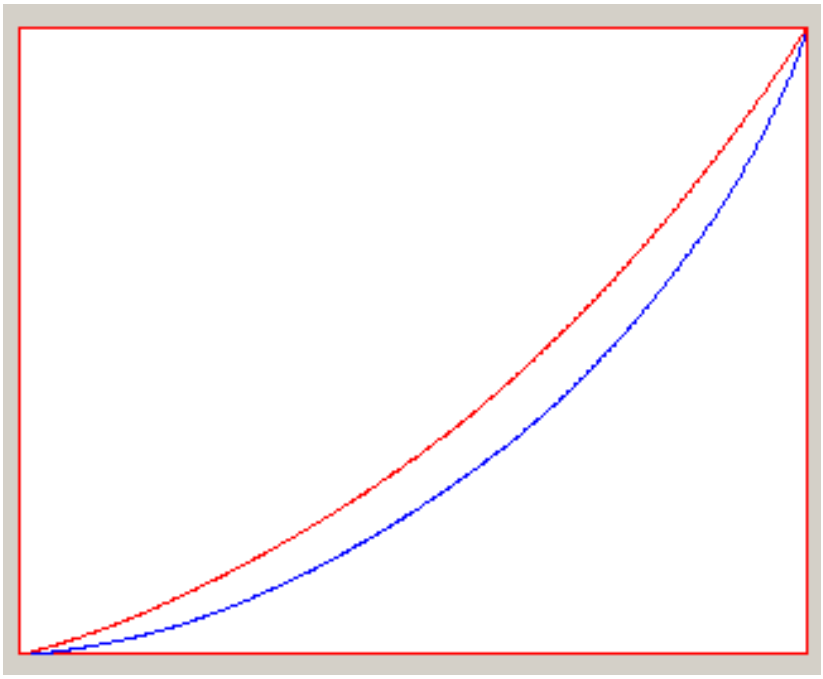
La différence des indices est 0.003, donc les deux variables sont très liées.

Exemple de répartition contexte C2



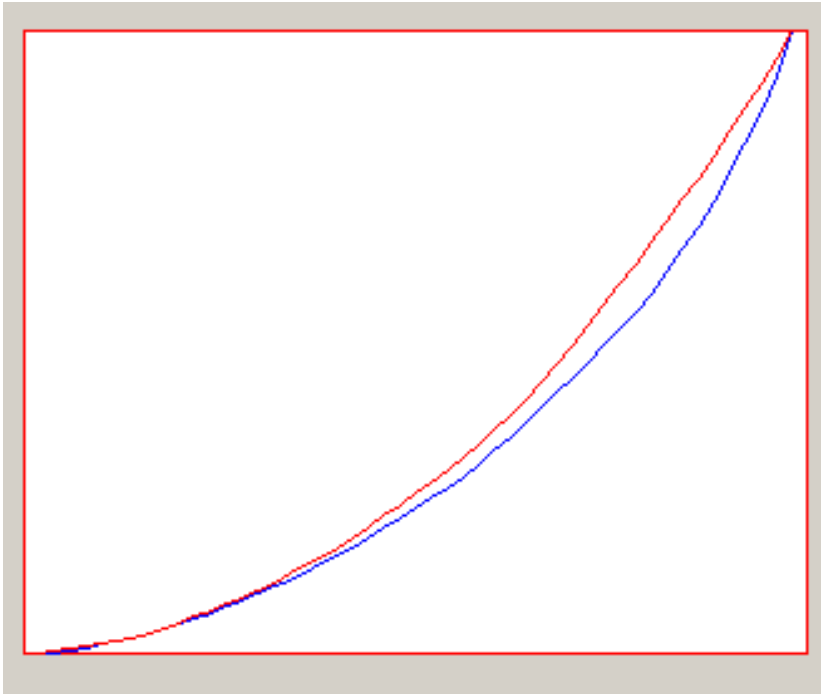
La différence des indices est 0.126, donc les deux variables sont indépendantes.

Exemple de répartition contexte C1



La différence des indices est 0.138, donc les deux variables sont indépendantes.

Exemple de répartition contexte C2

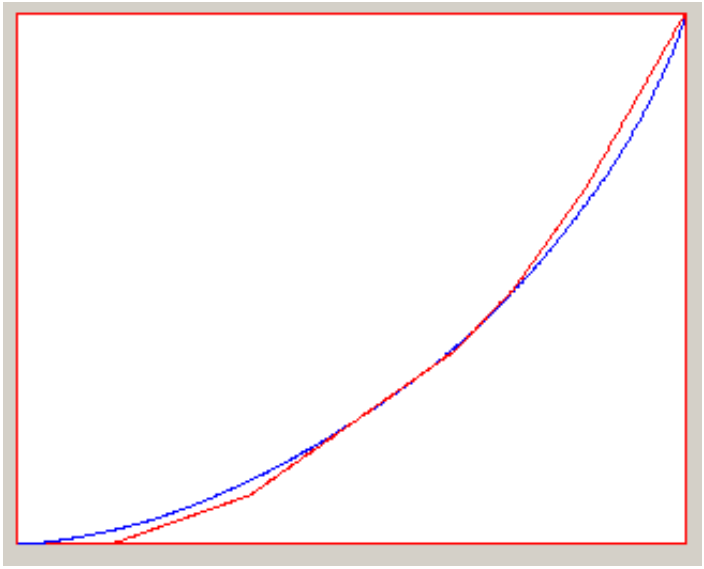


La différence des indices est 0.056, donc les deux variables sont indépendantes.

Dans le cas de ces deux contextes, on observe que les deux variables sont indépendantes.

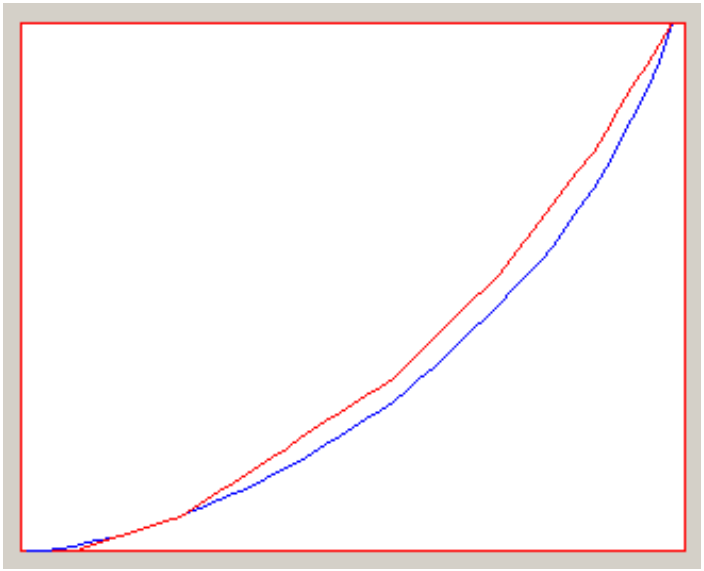
Dans les exemples précédents, il s'est agit de variables continues. Qu'en est-il si les variables sont discrètes?

Ci dessous, le résultat avec l'une des courbes calculées avec des valeurs de 0 à 5.
Contexte C1 :



On constate que les deux courbes ne sont pas tout à fait superposables, comme elles l'étaient dans le premier graphique. Pourtant, la différence d'indices de Gini est 0.002. Cet exemple montre que la comparaison des valeurs numérique doit être confirmée par une comparaison visuelle.

Contexte C2 :



On constate dans ce cas que les deux variables sont nettement indépendantes.

Discussion.

On peut se poser beaucoup de questions sur le résultat de cette méthode, comment fixer la limite entre l'indépendance et la non-indépendance de variables, etc.
Dans tous les cas, la courbe de Lorenz aura la même allure, partant de (0;0), terminée en (1;1) ayant une concavité tournée vers le haut. Cette courbe peut être approchée, avec une bonne précision par une courbe dont la fonction est Y = A + B exp(CX). On pourrait donc envisager de comparer les courbes à l'aide de cette fonction. Cela aurait l'intérêt de comparer trois paramètres au lieu d'un seul, comme on le pratique avec l'indice de Gini. Cela nécessiterait une étude complémentaire.

Par contre, l'examen de plusieurs courbes correspondant aux mêmes variables dans des contextes différents permettent de classer à coup sûr ces différents contextes.

Différents essais : voir annexe.

Vérification de la cohérence de l'indice de Gini avec les paramètres A, B et C de la fonction représentée par la courbe de Lorenz. Cette méthode consiste à établir une fonction du type puissance en calculant ses paramètres A, B, C, D, tels qu'elle minimise les écarts. Les valeurs V1, V2 et V3 sont les coefficients A, B, C de la fonction Y = A + B exp(CX), le résultat Res est l'indice de Gini calculé directement.

```
Résultat (fonction puissance) : A= 0.6494 B= -3.5859 C= -6.2572 D= -0.0567
Res = exp(A) * V1^B * V2^C * V3^D
Les constantes suivantes ont été ajoutées aux valeurs lues :
 1.00 1.42 1.00 0.00
Valeur calculée 0.26 (pour 0.26 théorique e=0.00)
Valeur calculée 0.26 (pour 0.26 théorique e=0.00)
Valeur calculée 0.22 (pour 0.22 théorique e=-0.00)
Valeur calculée 0.35 (pour 0.35 théorique e=0.00)
Valeur calculée 0.24 (pour 0.23 théorique e=0.00)
Valeur calculée 0.38 (pour 0.40 théorique e=-0.02)
Valeur calculée 0.28 (pour 0.28 théorique e=0.00)
Valeur calculée 0.27 (pour 0.28 théorique e=-0.00)
Valeur calculée 0.41 (pour 0.40 théorique e=0.01)

Calcul sur 9 groupes.
 Ecart-type = 0.01
 Ecart relatif (valeur absolue) = 1.70%
 Ecart relatif (valeur signée) = 0.02%
Ajustement puissance sur 9  groupes   emq=0.01  ep=0.01
```

Ces résultats confirment le bien fondé de cette méthode, puisqu'il y a homogénéité des comparaisons. Il y a lieu de préciser que tous ces essais ont été réalisés à partir de la même base (source incontestable), mais avec des variables différentes et des nombres d'individus différents.

On peut aussi se poser la question suivante : les variables étudiées contiennent deux informations, l'une représentée par la courbe bleue, l'autre par la courbe rouge, est-on en droit d'établir un lien logique entre ces deux représentation ?
Lors du calcul de chacune des courbes de Lorenz, ces deux informations sont dissociées. En effet les deux calculs sont faits successivement et chaque liste est triée en fonction de la variable étudiée. Le rang d'un individu concerné sera très probablement différent dans chaque cas.

Pour fixer les idées, considérons que l'on traite la variable "richesse" et la variable "rapport de taille de l'enfant" (les variables 1 et 2). Le but de l'étude est naturellement d'évaluer l'impact de la richesse de la famille sur la croissance de l'enfant.

Faisons l'hypothèse suivante :
Un enfant est issu d'une famille riche, mais suite à je ne sais quoi, il n'a pas grandi comme il aurait dû.
Un autre enfant est issu d'une famille pauvre, mais par volonté de sa mère, il a eu une croissance tout à fait normale.
Ces deux exemples sont des exceptions, au moins on les considère comme telles pour le raisonnement. En probabilité on sait que 0.7% des résultats sont "hors-norme", mais sur un grand nombre, il n'y a pas de raison de les rejeter.
Ceci est l'hypothèse limite. Dans le cas général, et dans le contexte du présent fichier, on constate que le critère richesse influe sur le caractère croissance. C'est ce qui reste à démontrer.

Soit un petit nombre d'individus dont le critère richesse est compris entre les bornes R1 et R2. Soit un autre petit nombre d'individus dont le critère croissance est compris entre les bornes C1 et C2. On choisira ces nombres d'individus égaux dans les deux cas.
La méthode des courbes de Lorenz place ces deux groupes à la même proportion dans l'ensemble des individus de la liste (position en abscisse). Le rapport d'inégalité, c'est à dire le rapport vis à vis de la somme totale des nombres représentant la valeur étudiée est donnée par la valeur en ordonnée des moyennes constatées. Si ces moyennes sont comparables, alors les individus correspondant aux bornes R1-R2 sont le mêmes que ceux correspondant aux bornes C1-C2.
Pour une meilleure compréhension, j'ai parlé de "petit nombre" et de "moyenne". Dans le cas général, les variables sont continues, chaque élément de courbe correspond à un individu. En vertu du postulat de la moyenne, la position d'un individu pour la variable étudiée sera la même pour l'autre variable. En termes mathématiques, cela pourrait se dire ainsi : "la probabilité que un individu soit situé à la même abscisse est maximale".
La position relative des deux courbes de Lorenz reflète donc bien, en ordonnée, la position de chaque individu correspondant à une abscisse.

Conclusion.

La méthode des courbes Lorenz est très utilisée dans plusieurs domaines. L'indice de Gini est une numérisation simple de cette notion d'équité vis à vis d'une situation.
La comparaison de deux variables pour un même contexte permet d'apprécier aisément le rapport de cause à effet, et la comparaison des résultats pour différents contextes est un outil de décision facile à utiliser et sûr.

Pierre Dolez le 26/12/2015

Annexe 1 : extrait du listing résultat d'essais

Nombre = 500 ; wt = Wealth ; score = zscore ; fic=Prod500RT ; Visu = Bon
Zscore indice de Gini = 0.256 (courbe bleue)
Y= A + B * exp(C * X) A = -0.303 B = 0.287 C = 1.523 (emq= 0.009)
Wealth (richesse) indice de Gini = 0.259
Y= A + B * exp(C * X) A = -0.260 B = 0.256 C = 1.600 (emq= 0.002)

Nombre = 50 ; wt = Wealth ; score = zscore ; fic=Prod50RT ; Visu = pas Bon
Zscore indice de Gini = 0.224 (courbe bleue)
Y= A + B * exp(C * X) A=-0.420 B= 0.409 C= 1.277 (emq= 0.007)
Wealth (richesse) indice de Gini = 0.350
Y= A + B * exp(C * X) A=-0.141 B= 0.129 C= 2.255 (emq= 0.011)

Nombre = 100 ; wt = Wealth ; score = zscore ; fic=Prod100RT ; Visu = pas Bon
Zscore indice de Gini = 0.235 (courbe bleue)
Y= A + B * exp(C * X) A=-0.377 B= 0.361 C= 1.364 (emq= 0.011)
Wealth (richesse) indice de Gini = 0.350
Y= A + B * exp(C * X) A=-0.133 B= 0.123 C= 2.266 (emq= 0.008)

Nombre = 500 ; wt = Weight ; score = zscore ; fic=Prod500PT ; Visu = pas Bon
Zscore indice de Gini = 0.256 (courbe bleue)
Y= A + B * exp(C * X) A=-0.303 B= 0.287 C= 1.523 (emq= 0.009)
Wealth (richesse) indice de Gini = 0.397
Y= A + B * exp(C * X) A=-0.104 B= 0.098 C= 2.402 (emq= 0.006)

Nombre = 500 ; wt = AgeMere ; score = zscore ; fic=Prod500AmT ; Visu = Bon
Zscore indice de Gini = 0.256 (courbe bleue)
Y= A + B * exp(C * X) A=-0.303 B= 0.287 C= 1.523 (emq= 0.009)
AgeMere indice de Gini = 0.275
Y= A + B * exp(C * X) A=-0.251 B= 0.239 C= 1.666 (emq= 0.008)

Nombre = 100 ; wt = AgeMere ; score = zscore ; fic=Prod500AmT ; Visu = Moyen
Zscore indice de Gini = 0.235 (courbe bleue)
Y= A + B * exp(C * X) A=-0.377 B= 0.361 C= 1.364 (emq= 0.011)
AgeMere : indice de Gini = 0.279
Y= A + B * exp(C * X) A=-0.242 B= 0.236 C= 1.678 (emq= 0.004)

Nombre = 500 ; wt = Région ; score = zscore ; fic=Prod500RegT ; Visu = Pas bon
Zscore indice de Gini = 0.256 (courbe bleue)
Y= A + B * exp(C * X) A=-0.303 B= 0.287 C= 1.523 (emq= 0.009)
Région indice de Gini = 0.395
Y= A + B * exp(C * X) A=-0.103 B= 0.085 C= 2.596 (emq= 0.012)

Nombre = 500 ; wt = Région ; score = Richesse ; fic=Prod500RegR ; Visu = Pas bon
Région indice de Gini = 0.259 (courbe bleue)
Y= A + B * exp(C * X) A=-0.260 B= 0.256 C= 1.600 (emq= 0.002)
Wealth (richesse) indice de Gini = 0.395
Y= A + B * exp(C * X) A=-0.103 B= 0.085 C= 2.596 (emq= 0.012)

Références : https://en.wikipedia.org/wiki/Gini_coefficient

Although the Gini coefficient is most popular in economics, it can in theory be applied in any field of science that studies a distribution. For example, in ecology the Gini coefficient has been used as a measure of biodiversity, where the cumulative proportion of species is plotted against cumulative proportion of individuals.[76] In health, it has been used as a measure of the inequality of health related quality of life in a population.[77] In education, it has been used as a measure of the inequality of universities.[78] In chemistry it has been used to express the selectivity of protein kinase inhibitors against a panel of kinases.[79] In engineering, it has been used to evaluate the fairness achieved by Internet routers in scheduling packet transmissions from different flows of traffic.[80]

The Gini coefficient is sometimes used for the measurement of the discriminatory power of rating systems in credit risk management.[81]

The discriminatory power refers to a credit risk model's ability to differentiate between defaulting and non-defaulting clients. The formula , in calculation section above, may be used for the final model and also at individual model factor level, to quantify the discriminatory power of individual factors. It is related to accuracy ratio in population assessment models.