EXTENSIONS DU MODÈLE LINÉAIRE

Rappelons tout d'abord que dans le *modèle linéaire* examiné jusqu'ici, la linéarité requise concerne les coefficients et non les variables explicatives initiales; ainsi les modèles :

$$y = a + b.x + c.x^2 + d.x^3 + f.z + \varepsilon$$

$$y = a + b.\cos(t.\pi/6) + c.\sin(t.\pi/6) + d.\ln(w) + \varepsilon$$

sont des modèles linéaires, mais non le modèle :

$$y = x^a + b.z^c + \varepsilon$$

Il existe en effet bien d'autres formes de dépendance, en lesquelles les paramètres inconnus et à estimer n'ont pas nécessairement une incidence linéaire.

MODÈLES EXPONENTIELS ET DÉRIVÉS

Une liaison non linéaire usuelle est la liaison ***exponentielle***.

Ainsi la classique fonction de production de type ***Cobb-Douglas*** :

$$P = a.L^b.K^c.d^t$$

Ici, avec un terme d'évolution temporelle, dont le travail L, le capital K (*inputs*), et le temps t, sont les explicatives, et a, b, c et d les paramètres inconnus, est un modèle non linéaire.

Ou encore le ***modèle d'évolution temporelle*** à taux constant :

$$y(t) = y_0.(1+i)^t$$

de paramètres inconnus y_0 et i, où le temps t est l'unique exogène, dit ***modèle exponentiel*** (et solution de la relation différentielle $y' = \ln(1+i).y$).

Ces deux modèles peuvent néanmoins être ***linéarisés*** par ***passage aux logarithmes***. Le premier modèle, par exemple, se transforme en le suivant :

$$\ln(P) = \ln(a) + b.\ln(T) + c.\ln(K) + \ln(d).t$$

qui est un modèle linéaire équivalent; la variable à expliquer étant à présent ln(P), les explicatives ln(T), ln(K) et t, et les coefficients ln(a), b, c et ln(d).

Les modèles ainsi linéarisés peuvent alors être estimés par régression linéaire. On remarque toutefois que la perturbation aléatoire du modèle initial (que l'on n'avait pas écrite explicitement) doit avoir une forme multiplicative particulière, dite ***log-normale***, pour que celle du modèle transformé vérifie les hypothèses des mco.

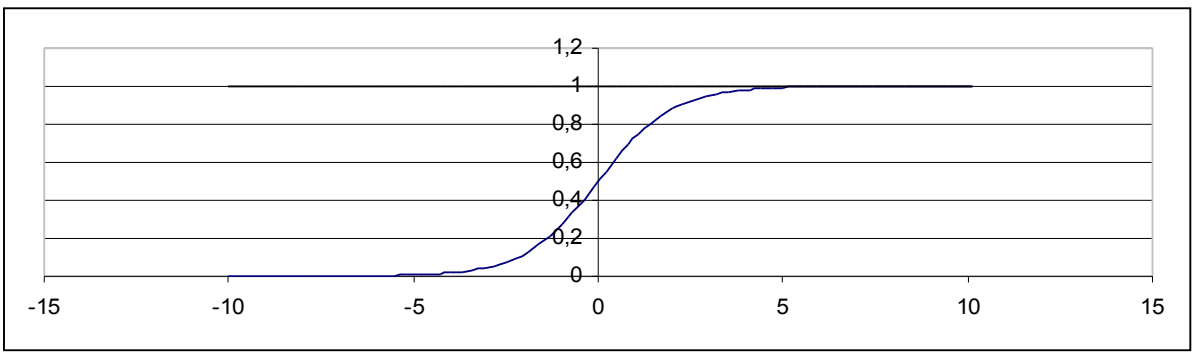
Il faut également noter que les coefficients sous leur forme initiale n'étant généralement pas des expressions linéaires des coefficients transformés, leurs estimations indirectes n'ont plus nécessairement les propriétés d'optimalité des estimations par les mco de ces derniers…

Modèle logistique

Un autre exemple de modèle de croissance usuel est le ***modèle logistique*** :

$$y = \frac{1}{a + b.e^{k.t}}$$ où a et b sont positifs et k négatif.

On figure la courbe pour a = b = - k = 1 (fonction logistique standard) :



La courbe logistique présente la forme symétrique d'un *S* aplati, traduisant une évolution accélérée, puis ralentie, vers la valeur asymptotique 1/a.

La grandeur y vérifie la condition différentielle:

$$y' = - k.(1-a.y).y$$

Bien que cela ne soit guère apparent, cette relation peut également être linéarisée. La transformation :

$$z = \ln\left(\frac{1}{y_t} - \frac{1}{y_{t+1}}\right)$$

conduit au modèle linéaire:

$$z = k.t + [\ln(b) + \ln(1 - e^k)] = k.t + c$$

On peut encore estimer par régression linéaire cette dernière relation (il faut cependant estimer a d'une autre manière, par exemple à l'aide du maximum observé de y).

MCO NON LINÉAIRES, MAXIMUM DE VRAISEMBLANCE

Bien d'autres formes de relations sont concevables, certaines ne pouvant pas être linéarisées, le modèle est alors ***intrinsèquement non linéaire*** (ainsi le modèle : y = a.x/(b+x) , dit *modèle de croissance de Monod*).

Si la linéarisation est parfois possible, comme dans les exemples précédents, et permet alors l'emploi des techniques élémentaires de régression, il est d'autres manières d'estimer les modèles *a priori* non linéaires.

Régression non linéaire

Soit le modèle

$$y = f(a, X)$$

où X désigne l'ensemble des explicatives, f la forme fonctionnelle retenue, et a l'ensemble des paramètres inconnus (on ne précise pas la perturbation).

Disposant d'observations y_i et x_i, on peut chercher directement la valeur A du vecteur des paramètres qui minimise la somme des carrés des erreurs, ou résidus :

$$SCR = \Sigma [y_i - f(a, x_i)]^2$$

On recourt pour cela aux techniques classiques d'optimisation (linéarisation locale, gradient, résolution itérative…) avec les problèmes habituels de convergence, d'optima locaux, etc., cette méthode est dite des ***mco non linéaires***.

Maximum de vraisemblance

Un autre point de vue, qui peut conduire en certains cas au calcul précédent, est celui du ***maximum de vraisemblance***. Le modèle complet s'écrit à présent :

$$y = f(a, X, \varepsilon)$$

incluant la perturbation aléatoire et précisant ses caractéristiques probabilistes. On cherche alors à maximiser par rapport aux paramètres la vraisemblance de l'ensemble des observations y_i conditionnées par celles x_i des explicatives.

En cas, par exemple, d'aléa additif : $\varepsilon_i = y_i - f(a, x_i)$, et sous les hypothèses de normalité traditionnelles, cela conduit à la minimisation précédente, mais ce n'est là qu'un cas particulier.

Sous des hypothèses très générales, les estimateurs du *maximum de vraisemblance* obtenus sont consistants, asymptotiquement efficaces et la méthode permet également d'estimer leurs variances asymptotiques.

Test de la log-vraisemblance

L'estimation du maximum de vraisemblance autorise en outre un test important qui généralise en quelque sorte le test d'une restriction linéaire déjà présenté dans le cas du modèle linéaire.

Soit H_a le modèle initial (en général non linéaire) et H_0, le modèle sous une certaine restriction, alors, si l'hypothèse H_0 est légitime, la quantité :

$$- 2.(LL_0 - LL_a) \quad \text{suit asymptotiquement une loi } chi^2(m)$$

c'est à dire une loi de chi-deux à m degrés de liberté, m étant le nombre de contraintes élémentaires définissant la restriction H_0 à H_a et LL notant ici le logarithme de la vraisemblance, ou ***log-vraisemblance***, maximisée par la méthode.

Les logiciels classiques opèrent ce test pour la significativité globale du modèle (hypothèse H_0 d'absence de toutes les variables hormis la constante) et pour la significativité de chaque variable considérée isolément, ce qui correspond au test traditionnel de Student des mco.

MODÈLES QUALITATIFS (CHOIX SIMPLE)

Une famille de modèles, dits à ***variable qualitative***, cherchent à expliquer la probabilité d'un certain événement ou d'une certaine situation en fonction d'explicatives convenables. Par exemple, la probabilité d'être propriétaire de son logement en fonction de variables socio-économiques, celle d'être atteint d'une certaine affection selon divers indicateurs biologiques, ou encore de voter pour tel parti en fonction de différents critères, etc.

La variable à expliquer y est donc une probabilité, mais la variable véritablement observée est *a priori* la variable dichotomique indiquant la présence ou l'absence du caractère étudié pour les différentes observations (et, naturellement, les explicatives envisagées). Dans le cas de données abondantes et d'explicatives prenant un petit nombre de valeurs, les fréquences empiriques associées aux différentes combinaisons peuvent néanmoins fournir une approximation des probabilités correspondantes, mais cela n'est pas indispensable.

Modèle de probabilité linéaire

Une première idée est de chercher à estimer un modèle tel que :

$$p = a + b.x + c.z \quad \text{(dans le cas de deux exogènes)}$$

il est clair qu'un tel modèle est naïf et peu adapté, les valeurs ajustées de p pouvant a priori parcourir l'ensemble des réels et non simplement l'intervalle [0, 1].

Modèles probit et logit

Une idée destinée à remédier à l'inconvénient précédent est de passer par l'intermédiaire d'une fonction F dont l'ensemble des valeurs prises est l'intervalle [0, 1], et d'estimer un modèle de la forme :

$$p = F(a + b.x + c.z) \quad \text{ou encore} \quad F^{-1}(p) = a + b.x + c.z$$

dans le cas de deux explicatives.

Les ***fonctions de répartitions*** associées aux distributions de probabilité (c'est à dire donnant la probabilité d'être plus petit que t pour la loi retenue) conviennent.

Si on prend pour F la fonction de répartition associée à la loi normale $\boldsymbol{N}(0, 1)$, le modèle est le ***modèle probit***. La fonction F, ni son inverse ne s'expriment sous une forme résolue, mais peuvent être calculées numériquement.

Si on retient pour F la ***fonction de répartition logistique*** (cad la fonction logistique standard rencontrée plus haut dans un autre contexte) :

$$p = F(t) = 1/(1 + e^{-t}) = e^{t}/(1 + e^{t})$$

on obtient le ***modèle logit***, longtemps apprécié pour la possibilité de calculer F^{-1} explicitement pour linéariser le modèle :

$$t = F^{-1}(p) = \ln[p/(1-p)]$$

quantité appelée ***logit*** de p, tandis que le rapport $p/(1-p)$ est parfois appelé ***odd ratio***. Et on a donc la relation :

$$\text{odd ratio}\ (p) = p/(1-p) = e^{a}.(e^{b})^{x}.(e^{c})^{z}$$

On va présenter ces deux modèles d'une manière quelque peu différente, quoique mathématiquement équivalente, dans le cadre général des ***modèles à variable latente***, non observable.

MODÈLES À VARIABLE LATENTE

Modèles probit et logit

Conservant pour fixer les idées deux exogènes, x et z, on suppose que la réalisation de l'événement étudié (y = 1) dépend d'une ***variable latente*** : y*, non observable :

$$y^* = a + b.x + c.z + \varepsilon$$

dépendant elle-même des exogène et d'une perturbation aléatoire ε, et telle que

$$y = 1 \quad \text{si} \quad y^* > 0$$
$$y = 0 \quad \text{sinon}$$

par suite, en supposant la loi de la perturbation ε centrée et symétrique, et en notant encore F sa fonction de répartition :

$$P(y = 1) = P(y^* > 0) = P(\varepsilon > - a - b.x - c.z) = P(\varepsilon < a + b.x + c.z)$$
$$= F(a + b.x + c.z)$$

La variable y* n'intervenant que par son signe, on voit que les coefficients a, b et c, et ε, ne sont définis qu'à un facteur d'échelle près, qu'on peut fixer librement, par exemple via l'écart-type de ε.

Comme il a été dit, si ε suit la loi normale centrée réduite, on a le *modèle probit*, et si ε suit la distribution logistique déjà indiquée, le *modèle logit* (ces deux modèles sont en fait assez proches, comme les deux lois de distribution qui les définissent).

Estimation, interprétation

Les modèles précédents sont estimés par la méthode du maximum de vraisemblance.

Comme dans le cas des mco classiques, les écart-type estimés des coefficients sont calculés et des tests appropriés permettent de juger de la significativité de chacun d'entre eux, et on examine ensuite le signe de ceux qui paraissent à retenir (qui indique le sens de l'influence de la variable considérée).

L'interprétation des coefficients eux-même demande plus de prudence : l'influence d'une variable sur la probabilité d'apparition étudiée p n'est en effet pas linéaire; dans le cas du modèle logistique par exemple, elle l'est seulement sur son logit : ln[p/(1-p)].

Dans ce cas, sont généralement calculées sous le noms d'*odds ratio estimés* les exponentielles des coefficients estimés, qui du fait de la relation exponentielle indiquée plus haut, mesurent l'incidence multiplicative de l'augmentation d'une unité du facteur considéré sur, précisément, l'odd ratio : p/(1-p).

L'incidence d'une variable sur la probabilité p elle-même dépend en revanche du point considéré... Aussi, au prix de calculs appropriés, sont parfois calculés les effets marginaux (ou bien les élasticités) en un point particulier, souvent le point moyen de la population.

Certains logiciels calculent des ***taux de bien classés*** : une observation étant considérée comme telle si sa probabilité théorique, obtenue par le modèle estimé, dépasse 0,5 et si elle présente le caractère étudié (y = 1), ou bien si elle est inférieure à 0,5 et ne le présente pas (y = 0).

Une variante plus subtile examine toutes les paires associant une observation positive et une observation négative, et calcule la proportion de paires pour lesquelles les probabilités calculées sont dans le sens logiquement attendu, c'est à dire la plus grande pour l'observation positive.

Dans les deux cas, il est encourageant de trouver un bon taux de bien classés, mais il serait déraisonnable d'attendre un nombre très proche de 1, compte tenu notamment du caractère aléatoire du modèle…

Modèle tobit (ou de Tobin)

Il arrive que l'on souhaite expliquer une variable y ne pouvant prendre que des valeurs positives (par exemple les dépenses automobiles d'un ménage n'ayant pas nécessairement une voiture), ou simplement non observables en-dessous d'un certain seuil, à l'aide de diverses explicatives. Comme précédemment, il est clair qu'un modèle tel que :

$$y = a + b.x + c.z + d.w + \varepsilon$$

risque de ne pas convenir, la fonction de x, z et w envisagée pouvant *a priori* prendre des valeurs non bornées et de signe quelconque.

L'idée de Tobin est que la relation précédente définit en fait une ***variable latente*** : y*, non observable lorsqu'elle est négative, et égale à y dans le cas contraire. On dit aussi, d'une autre manière, que y est ***censurée à gauche***.

Le modèle peut à nouveau être estimé par la méthode du maximum de vraisemblance à partir des observations de y et des explicatives.

Une modélisation plus large (*modèle de Cragg*) autorise des fonctions différentes pour la censure (la ***fonction de sélection***) et l'output y.

Exemple

Une étude cherche à expliquer la probabilité d'occuper un emploi, et le temps travaillé éventuel, pour les femmes mariées américaines en 1975. Les variables considérées sont :

- LFP dummy variable égale à 1 si la personne travaille
- WHRS nombre d'heures travaillées en 1975
- KL6 nombre d'enfants de moins de 6 ans à la maison
- K618 nombre d'enfants de 6 à 18 ans à la maison
- WA âge de la personne
- WA2 WA au carré
- WE nombre d'années d'étude
- WMED nombre d'années d'étude de la mère
- WFED nombre d'années d'étude du père
- UN taux de chômage local
- CIT dummy variable égal à un pour les grandes villes, 0 sinon
- PRIN revenu du ménage hors du revenu éventuel de la femme

On présente une partie des calculs et résultats obtenus par Gretl, SAS et Eviews (sorties composites).

Modèle **logit**

The LOGIT Procedure

```
Convergence obtenue après 5 itérations

Modèle : Estimation Logit utilisant les 753 observations 1-753
Variable dépendante: LFP

  Moyenne de LFP = 0,568
  Nombre de cas 'correctement prédis' = 509 (67,6%)
  f(beta'x) à la moyenne des variables indépendantes = 0,245
  Pseudo-R{\super 2} de McFadden = 0,11988
```

```
  Log de vraisemblance = -453,15
  Test du ratio de vraisemblance: Chi-deux(10) = 123,446 (p. critique 0,000000)
  Critère d'information d'Akaike (AIC) = 928,3
  Critère bayesien de Schwarz (BIC) = 979,165
  Critère d'Hannan-Quinn (HQC) = 947,896

  VARIABLE              COEFFICIENT      ERR. STD      T          PENTE
                                                                 (à la moyenne)
  const                -1,90685          2,57704      -0,740
  KL6                  -1,42442          0,201186     -7,080     -0,348319
  K618                 -0,0823107        0,0696801    -1,181     -0,0201278
  WA                    0,0738531        0,118054      0,626      0,0180596
  WA2                  -0,00150858       0,00134965   -1,118     -0,000368899
  WE                    0,265813         0,0450665     5,898      0,0650003
  WFED                 -0,0162468        0,0288467    -0,563     -0,00397291
  WMED                  0,00683259       0,0302391     0,226      0,00167080
  UN                   -0,0179777        0,0263328    -0,683     -0,00439617
  CIT                   0,0409095        0,178216      0,230      0,0100038
  PRIN                 -3,55837E-05      8,14429E-06  -4,369     -8,70144E-06

Odds Ratio Estimates

              Point          95% Wald
  Effect      Estimate     Confidence Limits

  KL6          0.241       0.162       0.357
  K618         0.921       0.803       1.056
  WA           1.077       0.854       1.357
  WA2          0.998       0.996       1.001
  WE           1.304       1.194       1.425
  WFED         0.984       0.930       1.041
  WMED         1.007       0.949       1.068
  UN           0.982       0.933       1.034
  CIT          1.042       0.735       1.477
  PRIN         1.000       1.000       1.000

Association of Predicted Probabilities and Observed Responses

               Prédicted
                 0     1
  Observed  0  167   158
            1   86   342

  Percent Concordant      73.2    Somers' D    0.467
  Percent Discordant      26.5    Gamma        0.468
  Percent Tied             0.2    Tau-a        0.229
  Pairs                139100    c            0.734
```

MCO pour les seules salariées

The REG Procedure

Model: MODEL1
Dependent Variable: WHRS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	17169593	3433919	6.03	<.0001
Error	422	240141427	569056		
Corrected Total	427	257311020			

```
  Root MSE            754.35768    R-Square     0.0667
  Dependent Mean     1302.92991    Adj R-Sq     0.0557
  Coeff Var            57.89703

Parameter Estimates

                      Parameter      Standard
  Variable            Estimate       Error         t Value   Pr > |t|

  Intercept        2118.23514      339.59069       6.24      <.0001
  KL6              -341.03762       99.79527      -3.42      0.0007
  K618             -114.26877       30.70754      -3.72      0.0002
  WA                 -7.76401        5.52256      -1.41      0.1605
  WE                -16.26121       17.03608      -0.95      0.3404
  PRIN               -0.00430        0.00365      -1.18      0.2390
```

Modèle **tobit**

The TOBIT Procedure

```
Convergence obtenue après 21 itérations

Modèle : Estimation Tobit utilisant les 753 observations 1-753
Variable dépendante: WHRS

  Moyenne de la variable dépendante = 740,576
  Écart-type de la var. dép. = 871,314
  Observations censurées: 325 (43,2%)
  sigma (scale) = 1262,68
  Log de vraisemblance = -3893,65
  Critère d'information d'Akaike (AIC) = 7801,29
  Critère bayesien de Schwarz (BIC) = 7833,66
  Critère d'Hannan-Quinn (HQC) = 7813,76

  VARIABLE       COEFFICIENT        ERR. STD        T       p. critique
  const           1111,67           483,919         2,297   0,02161 **
  KL6            -1060,61           127,447        -8,322   <0,00001 ***
  K618           -106,985           43,8260        -2,441   0,01464 **
  WA             -36,3008           7,84445        -4,628   <0,00001 ***
  WE              127,705           25,3298         5,042   <0,00001 ***
  PRIN           -0,0220749         0,00475317     -4,644   <0,00001 ***
  Scale           1262,680          47,4008

Test pour la normalité des résidus
  Hypothèse nulle: l'erreur est normalement distribué
  Statistique de test: Chi-deux(2) = 65,0002
  avec p. critique = 7,68042e-015
```

----===ooOoo===----

(21.04.2009)