LA COLINEARITE

La colinéarité est un problème très important en économie. Nous allons dans un premier temps en donner la définition et, sur un exemple évoquer les principaux problèmes causés par la colinéarité. Ensuite nous verrons plus théoriquement ces problèmes ainsi que le test de colinéarité de Belsley, Kuh et Welsch et nous terminerons par le traitement de la colinéarité en particulier avec la régression Ridge (bornée).

1 Définitions

1.1 La colinéarité parfaite

Prenant un modèle avec 4 variables explicatives

$$y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + \epsilon$$

Supposons que la variable X_2 dépend des variables X_3 et X_4 sous la forme exacte

$$X_2 = 3X_3 - 2X_4$$

On a bien sur ici des variables parfaitement colinéaires.Une colonne de la matrice X est combinaison de deux autres ici. Nous avons vu que cela était contraire aux MCO dans le premier chapitre car s'il y a colinéarité parfaite la matrice X^tX a un déterminant nul, en math vous avez vu que cela conduisait au fait que X^tX n'était pas de plein rang, certaines de ses valeurs propres (ici 1 car il y a une seule relation de colinéarité) sont nulles. Si son déterminant est nul, la matrice X^tX n'est pas inversible et donc le résultat des MCO n'existe pas. Inutile de se lamenter pour cela, pour résoudre ce problème il suffit de remplacer X_2 par sa valeur dans le modèle.

$$y = a_0 + a_1X_1 + a_2(3X_3 - 2X_4) + a_3X_3 + a_4X_4 + \epsilon$$

$$y = a_0 + a_1X_1 + (3a_2 + a_3)X_3 + (a_4 - 2a_2)X_4 + \epsilon$$

et le tour est joué, on peut faire les MCO sur ce modèle, en "oubliant" X_2.
Ce n'est pas ce problème qui nous inquiète vraiment, mais ce que l'on appelle la quasi-colinéarité.

1.2 La quasi-colinéarité

Dans le cas précédent, le vecteur X_2 est une combinaison parfaite des deux autres vecteurs. Mais supposons que la colinéarité ne soit pas parfaite, mais que X_2 soit "proche" de $3X_3 - 2X_4$, alors le déterminant de X^tX n'est pas nul mais très petit, certaines valeurs propres (ici une car il a qu'une seule relation) ne sont pas nulles mais presque nulles. Même avec un déterminant très petit les ordinateurs ont assez de puissance maintenant pour inverser la matrice X^tX et donner un résultat des MCO. Avant, dans les années 70 il nous arrivait

encore que les ordinateurs utilisés à l'Université ne soient pas assez puissants pour donner un résultat et indiquaient matrice non inversible, pas de résultat des MCO.
Nous avons donc maintenant un résultat, mais quelles sont réellement les conséquences de la quasi-colinéarité que l'on va appeler maintenant la **colinéarité.**

2 Conséquences de la colinéarité sur un exemple

Reprenons notre exemple statique sur la consommation des ménages US

```
Linear Regression - Estimation by Least Squares
Dependent Variable CM
Quarterly Data From 1955:01 To 2000:04
Usable Observations    184      Degrees of Freedom   179
Centered R**2     0.999360      R Bar **2   0.999346
Uncentered R**2   0.999898      T x R**2     183.981
Mean of Dependent Variable      3199.4326087
Std Error of Dependent Variable 1394.5165320
Standard Error of Estimate        35.6670535
Sum of Squared Residuals        227712.82831
Regression F(4,179)               69891.7616
Significance Level of F           0.00000000
Log Likelihood                    -916.20793
Durbin-Watson Statistic             0.548403

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                  78.773545    11.847804      6.64879  0.00000000
2.  RD                         0.873899     0.003907    223.65346  0.00000000
3.  TCHO                     -10.047432     2.176373     -4.61659  0.00000742
4.  SP                         0.298527     0.018802     15.87746  0.00000000
5.  INF                    -1781.517057   469.057938     -3.79807  0.00019957
```

Le coefficient du revenu disponible RD est de 0.87 et a un " t de Student "=223.6 très important.
Le coefficient du taux de chômage TCHO est de -10.04 et un t=-4.65 significatif
On va ajouter une variable explicative : la comsommation des ménages en (t-1) CM{1}=CM1

```
Linear Regression - Estimation by Least Squares
Dependent Variable CM
Quarterly Data From 1955:01 To 2000:04
Usable Observations    184      Degrees of Freedom   178
Centered R**2     0.999835      R Bar **2   0.999830
Uncentered R**2   0.999974      T x R**2     183.995
Mean of Dependent Variable      3199.4326087
Std Error of Dependent Variable 1394.5165320
Standard Error of Estimate        18.1570404
Sum of Squared Residuals        58682.704485
Regression F(5,178)              215857.2708
Significance Level of F           0.00000000
Log Likelihood                    -791.46143
Durbin-Watson Statistic             1.552945

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                  21.774807     6.535593      3.33173  0.00104959
2.  CM{1}                      0.791822     0.034970     22.64316  0.00000000
3.  RD                         0.185129     0.030483      6.07310  0.00000001
4.  TCHO                      -0.522327     1.185099     -0.44075  0.65993225
5.  SP                         0.077147     0.013682      5.63852  0.00000007
6.  INF                    -1250.424564   239.932682     -5.21156  0.00000051
```

Le coefficient de RD a fortement baissé ainsi que le t =6.07
Le coefficient de TCHO est passé à -0.5 et son t =-044 n'est plus significatif.
Les autres coefficients sont aussi modifiés.
Par contre le s qui était de 35.6670535 passe à 18.1570404, donc le second modèle est bien meilleur que le premier globalement.
Cela montre une très forte colinéarité entre CM1 et RD (le R^2 entre ces deux variables est R Bar **2 = 0.996871, il est calculé en estimant CM1 en fonction de RD).
La colinéarité est moins forte entre CM1 et TCHO mais les dégâts sur les coefficients sont importants également.

2.1 Conclusions

On vient de voir sur un exemple, que si dans un modèle on ajoute une variable très colinéaire aux autres alors deux phénomènes vont se passer:

- Les coefficients des anciennes variables colinéaires avec la nouvelle, se trouvent fortement modifiés quand on introduit cette variable. Economiquement les coefficients ne représentent plus la même chose.

- Les variances estimées sont souvent beaucoup plus grandes (voir l'écart-type estimé de RD) et par conséquent les "t de Student" sont plus petits, jusqu'à parfois devenir non significatifs alors qu'ils l'étaient avant.

  Il faut donc se méfier d'un mauvais t dans un modèle c'est parfois un signe de colinéarité.

- En conclusion, la colinéarité dans un modèle est un fléau pour la valeurs des coefficients mais souvent si la (ou les) variable ajoutée est intéressante elle diminue fortement le s du modèle c'est-à-dire que le modèle est mieux expliqué GLOBALEMENT.

  Que faire ? Il faut priviligier un bon modèle global en général, mais alors il faut savoir que l'on ne peut pas trop faire confiance aux valeurs numériques des coefficients (parfois même à leur signe), inutile de faire de grands effets de manche à propos de la valeur économique des coefficients surtout quand on ajoute des variables retardées (modèles dynamiques)!!!!!

3 Détection de la colinéarité

Nous allons présenter les principales méthodes de détection de la colinéarité, de la plus simple la matrice de corrélation à la plus complexe mais aussi la plus complète, le test de Belsley, Kuh et Welsch.
Rappels de mathématique: s'il y a colinéarité parfaite, le déterminant de la matrice X^tX est nul, la matrice a autant de valeurs propres nulles que de relations de colinéarité parfaite.
S'il y a quasi-colinéarité (colinéarité) le déterminant de la matrice X^tX est proche de 0, la matrice a autant de valeurs propres très petites que de relations de colinéarité.

3.1 La matrice de corrélation

C'est la matrice X^tX centrée et réduite. C'est une matrice symétrique qui contient des 1 sur sa diagonale et les corrélations des variables deux à deux ailleurs. Plus les corrélations sont proches de ± 1 plus il y a colinéarité de ces variables deux à deux.
C'est une technique simple de repérage de la colinéarité, mais son inconvénient est de ne repérer les corrélations des variables que deux par deux. L'exemple traditionnel que l'on propose pour voir les limites de cette méthode est le suivant: voici une matrice de corrélation

$$\begin{pmatrix} 1 & -.5 & -.5 \\ -.5 & 1 & -.5 \\ -.5 & -.5 & 1 \end{pmatrix}$$

Les corrélations de -.5 sont moyennes, pas de forte colinéarité entre les variables deux à deux et pourtant le déterminant de cette matrice est nul ce qui indique une corrélation parfaite. Comme cette corrélation parfaite ne peut être entre des variables deux à deux, c'est que la corrélation parfaite est entre les 3 variables (car la matrice est 3x3 donc il y a trois variables). Nous ne pouvions voir ce résultat avec les corrélations deux à deux.

3.2 Le determinant de X^tX

La matrice X^tX est égale à sa transposée, ses k valeurs propres sont des $\lambda_i \geq 0$ et

$$\det(X^tX) = \prod_{i=1}^{k} \lambda_i$$

Si une valeur propre est nulle, il y a une relation de colinéarité parfaite et det(X^tX)=0
Si une valeur propre est petite, il y a une relation de colinéarité et det(X^tX) est petit.
Si deux valeur propres sont petites, il y a deux relations de colinéarité et det(X^tX) est petit.
Mais X^tX dépend des unités des variables il est donc difficile de savoir si ce déterminant est bien" petit". On peut supprimer l'influence des unités en prenant la matrice centrée réduite qui est la matrice de corrélation et en regardant si son déterminant est petit.
Comme il est difficile de voir quand il est réellement petit, on préfère utiliser la notion de conditionnement de matrice.

3.3 Le conditionnement de la matrice X

C'est la première partie de test de Belsley, Kuh et Welsch.

3.3.1 Le conditionnement de X^tX

Par définition c'est la rapport de la plus grande à la plus petite valeur propre de X^tX

$$Con(X^tX) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Si une valeur propre est petite le conditionnement est très grand, mais encore une fois que signifie petit ou grand?
Pour le savoir on va développer la première partie du test de Belsley, Kuh et Welsch.

3.3.2 Les indices de conditionnement de X

On travaille sur le modèle réduit pour ne plus avoir de problème d'unité, c'est-à-dire que les variables X_i sont divisées par $\sqrt{\sum X_{it}^2}$

$$\begin{aligned}
y_t &= a_0 + a_1\sqrt{\sum X_{1t}^2}\frac{X_{1t}}{\sqrt{\sum X_{1t}^2}} + a_2\sqrt{\sum X_{2t}^2}\frac{X_2}{\sqrt{\sum X_{2t}^2}} + a_3\sqrt{\sum X_{3t}^2}\frac{X_3}{\sqrt{\sum X_{3t}^2}} + \epsilon \\
y_t &= a_0 + a_1\sqrt{\sum X_{1t}^2}x_{1t} + a_2\sqrt{\sum X_{2t}^2}x_{2t} + a_3\sqrt{\sum X_{3t}^2}x_{3t} + \epsilon
\end{aligned}$$

On notera maintenant X^tX la matrice des variables réduites cette fois pour ne pas alourdir les notations.
Nous savons que l'on peut décomposer la matrice symétrique X^tX en produit

$$X^tX = V\Lambda V^t$$

où V est la matrice orthogonale des vecteurs propres et Λ la matrice des valeurs propres de X^tX.
On définit les matrices U et D diagonale

$$\begin{aligned}
\underset{(n,k)}{U} \text{ et } \underset{(k,k)}{D} \text{ telles que } X &= UDV^t \text{ et } U^tU = I \text{ et } V^tV = I \\
X^tX &= VDU^tUDV^t = VD^2V^t
\end{aligned}$$

car D est diagonale. On retrouve V la matrice orthogonale ($V^tV = I$) des vecteurs propres et $D^2 = \Lambda$ matrice des valeurs propres de X^tX.
Donc $D^2 = \Lambda$

- Définition des valeurs singulières de X:

  On note d_i les éléments de la matrice diagonale D, on vient de démontrer que $d_i = \pm\sqrt{\lambda_i}$.

  On appelle valeurs singulières de X les $d_i = +\sqrt{\lambda_i}$, il y en a k.

- Définition du conditionnement de la matrice non carrée X :

  Le conditionnement de X= cond X=$\frac{d\max}{d\min} = \sqrt{\text{cond } X^tX}$

- Définition des indices de conditionnement C_i de X : Il y en a k autant que de variables explicatives.

  $$C_i = \frac{d_i}{d_{\min}} \quad \text{i=1,....k}$$

  on va classer $d_k = d_{\max} > d_{k-1} > .... > d_2 > d_1 = d_{\min}$

  donc

  $$\begin{aligned}
  \frac{d\min}{d\min} &< \frac{d_2}{d\min} < \frac{d_3}{d\min} < ... < \frac{d_{k-1}}{d\min} < \frac{d\max}{d\min} \\
  C_1 &= 1 < C_2 < C_3 < ... < C_{k-1} < C_k = \text{Cond } X
  \end{aligned}$$

3.3.3 Travaux de Belsley, Kuh et Welsch

Grâce à des simulations ils ont montré que si on trouve des indices de conditionnement

$$
\begin{aligned}
30 &< C_i < 100 \text{ il y a colinéarité} \\
C_i &> 100 \text{ il y a forte colinéarité}
\end{aligned}
$$

En pratique en économié des indices sont souvent supérieurs à 1000.
Exemple 1: avec k=5 variables explicatives
les indices
C1 = 1
C2 = 4
C3 = 10
C4 = 20
C5 = 25 =cond X $\Longrightarrow$ le plus grand indice étant inférieur à 30 il n'y a pas de colinéarité

Exemple 2 : avec k=4
C1 = 1
C2 = 10
C3 = 20 le second indice étant inférieur à 30 pas de seconde relation de colinéarité
C4 = 150 = cond X $\Longrightarrow$ le plus grand indice étant supérieur à 100 il y a une forte relation de colinéarité

Exemple 3 avec k=4
C1 = 1
C2 = 10 inférieur à 30 donc pas de troisième relation
C3 = 100 le second indice étant supérieur à 100 il y a une seconde relation de colinéarité
C4 = 250 = cond X $\Longrightarrow$ le plus grand indice étant supérieur à 100 il y a une forte relation de colinéarité

On commence toujours par le plus grand indice, s'il est <30 pas de relation de colinéarité dans le modèle.
S'il est >30 il y a une relation et on remonte pour voir si le précédent est aussi >30 si oui il y a une seconde relation et on remonte encore, sinon on s'arrête.

Exemple 4 avec la consommation des ménages US avec comme variables explicatives CM1, RD, TCHO, SP, INF
1
2.7
5.3
11.1
14.9 $\Longrightarrow$ pas d'autre relation de colinéarité
283.8 $\Longrightarrow$ une très forte relation de colinéarité

Ce test est très intéressant car il donne une mesure du phénomène de colinéarité et fournit le nombre de relation de colinéarité. Ce que l'on aimerait bien également c'est connaitre les variables responsables de la colinéarité. Nous allons trouver en général la réponse dans la seconde partie du test de Belsley,Kuh et Welsch (B.K.W)

4 Test de Belsley-Kuh-Welsch partie 2

Nous allons grâce à cette seconde partie du test essayer de trouver les variables responsables de la colinéarité.

4.1 Construction du tableau

L'étude se base sue les variances des $\widehat{a_i}$. Nous avons vu que $X^tX = VD^2V^t$ donc la matrice de variances-covariances de $\overrightarrow{a}$ peut s'écrire

$$V_{\overrightarrow{a}} = \sigma^2(X^tX)^{-1} = \sigma^2(VD^2V^t)^{-1} = \sigma^2VD^{-2}V^t$$

car la matrice V est la matrice orthogonale des valeurs propres. La variance v($\widehat{a_i}$) est le i^{ième} terme sur la diagonale de la matrice $V_{\overrightarrow{a}}$ soit en posant v_{ij} le terme général de la matrice V et d_j les valeurs singulières de X

$$v(\widehat{a_i}) = \sigma^2 \sum_{j=1}^{k} \frac{v_{ij}^2}{d_j^2}$$

On va caractériser l'apport de chaque valeur singulière d_j à la variance $v(\widehat{a_i})$ par le rapport

$$\sigma^2 \frac{v_{ij}^2}{d_j^2} / v(\widehat{a_i}) = \frac{v_{ij}^2}{d_j^2} / \sum_{j=1}^{k} \frac{v_{ij}^2}{d_j^2} = \pi_{ij}$$

π_{ij} est le pourcentage de la variance $v(\widehat{a_i})$ dépendant de la valeur singulière d_j d'où la construction du tableau

	$\widehat{a}_1$	...	$\widehat{a}_i$	...	$\widehat{a}_k$	$Index$
d_1	π_{11}	...	π_{i1}	...	π_{k1}	$d_{\max}/d_1$
...	...	...	...	...	...	...
d_j	π_{1j}	...	π_{ij}	...	π_{kj}	$d_{\max}/d_j$
...	...	...	...	...	...	...
d_k	π_{1k}	...	π_{ik}	...	π_{kk}	$d_{\max}/d_k$
$\sum$	1	1	1	1	1	

Ce tableau permet donc d'établir la part de la variance d'un estimateur imputable à chaque valeur singulière de X.

4.1.1 Cas particulier de vecteurs orthogonaux

Si les termes v_{ij} et v_{ji} de la matrice des vecteurs propres sont nuls alors les deux variables explicatives X_i et X_j de X sont orthogonales.
Si tous les vecteurs x_i de X sont orthogonaux, la matrice V des vecteurs propres est diagonale et alors

$$\pi_{ij} \ = 0 \text{ si } i \neq j \quad \text{et} \quad \pi_{ij} \ = 1 \text{ si } i = j$$

Si on trouve dans le tableau une colonne i contenant des valeurs nulles $\pi_{ij} = 0$ sauf $\pi_{ii} = 1$ et si de plus dans les autres colonnes $\pi_{ji} = 0$ pour tout j $\neq i$ alors le vecteur X_i est orthogonal aux autres vecteurs de X.

4.1.2 Interprétation du tableau

Nous nous intéressons à la colinéarité c'est-à-dire au contraire de l'orthogonalité.
Dans le cas de la colinéarité les π_{ij} associés à des valeurs singulières petites seront grands (ne pas oublier que par construction les π_{ij} sont ≤ 1), sauf dans le cas ou π_{ij} est lui-même très petit, à la limite lorsque $v_{ij} = 0$ c'est-à-dire lorsque les vecteurs x_i et x_j sont orthogonaux.
B.W.W suggèrent de considérer π_{ij} grand s'il est >0.5. Ceci conduit à la découverte des variables colinéaires.
En effet, sur la ligne j correspondant à une valeur singulière petite (indice de conditionnement >30), si deux ou plusieurs π_{ij} sont >0.5 cela implique que les v_{ij} correspondants sont importants et ainsi des liaisons linéaires existent entre ces variables.

Exemple 1

$Indice$	$\widehat{a}_1(X_1)$	$\widehat{a}_2(X_2)$	$\widehat{a}_3(X_3)$	$\widehat{a}_4(X_4)$
1	1	0	0	0
2	0	1	0	0
5	0	0	0.01	0.1
500	0	0	0.99	0.9

L'indice de conditionnement de 500 correspond à une forte colinéarité entre les variables X_3 et X_4 (deux indices >0.5). c'est la seule liaison de colinéarité car un seul indice est >30.

Exemple 2

$Indices$	$\widehat{a}_1(X_1)$	$\widehat{a}_2(X_2)$	$\widehat{a}_3(X_3)$	$\widehat{a}_4(X_4)$	$\widehat{a}_5(X_5)$
1	0	0	0	0	0
5	0	0	0	0	0
8	0.002	0	0.003	0.003	0
83	0	0.001	0.824	0.815	0.001
468	0.998	0.999	0.173	0.182	0.999

L'indice 468 indique une forte colinéarité entre les variables X_1, X_2 et X_5 (trois valeurs >0.5). Un second indice élevé (>30) indique une colinéarité entre les variables X_3 et X_4.

Exemple 3

Reprenons l'exemple de la consommation des ménages US en fonction de la consommation décalée d'une période CM1, du RD, TCHO, SP, INF. Le tableau donne

```
TABLEAU DE DECOMPOSITION DE LA VARIANCE
***************************************
indice      tableau de decomposition de la variance
           cons  CM1   RD    TCHO  SP    INF
   1.0     0.002 0.000 0.000 0.001 0.001 0.007
   2.7     0.002 0.000 0.000 0.004 0.044 0.098
   5.3     0.035 0.000 0.000 0.041 0.024 0.768
  11.1     0.433 0.001 0.001 0.012 0.196 0.054
  14.9     0.382 0.000 0.000 0.804 0.253 0.061
 283.8     0.147 0.999 0.999 0.138 0.482 0.012
```

Comme nous l'avons vu un seul indice est important (283.8). On constate sur cette ligne que 3 variables ont une part de la variance >0.5, il s'agit de CM1, RD et dans une moindre part SP. La relation de colinéarité lie ces trois variables.

Comme vous le voyez, ce test est très simple d'utilisation. La seule difficulté est de le programmer. Souvenons-nous du fait que les variables doivent être réduites pour faire ces deux tests de colinéarité, réduites mais non centrées car vous voyez que la constante fait partie du tableau et qu'elle peut dans certains cas (pas ici) faire partie de la relation de colinéarité.
Conclusion : un seul indice >30 donc une seule relation de colinéarité liant CM1, RD et plus faiblement SP.