U. Paris Nanterre

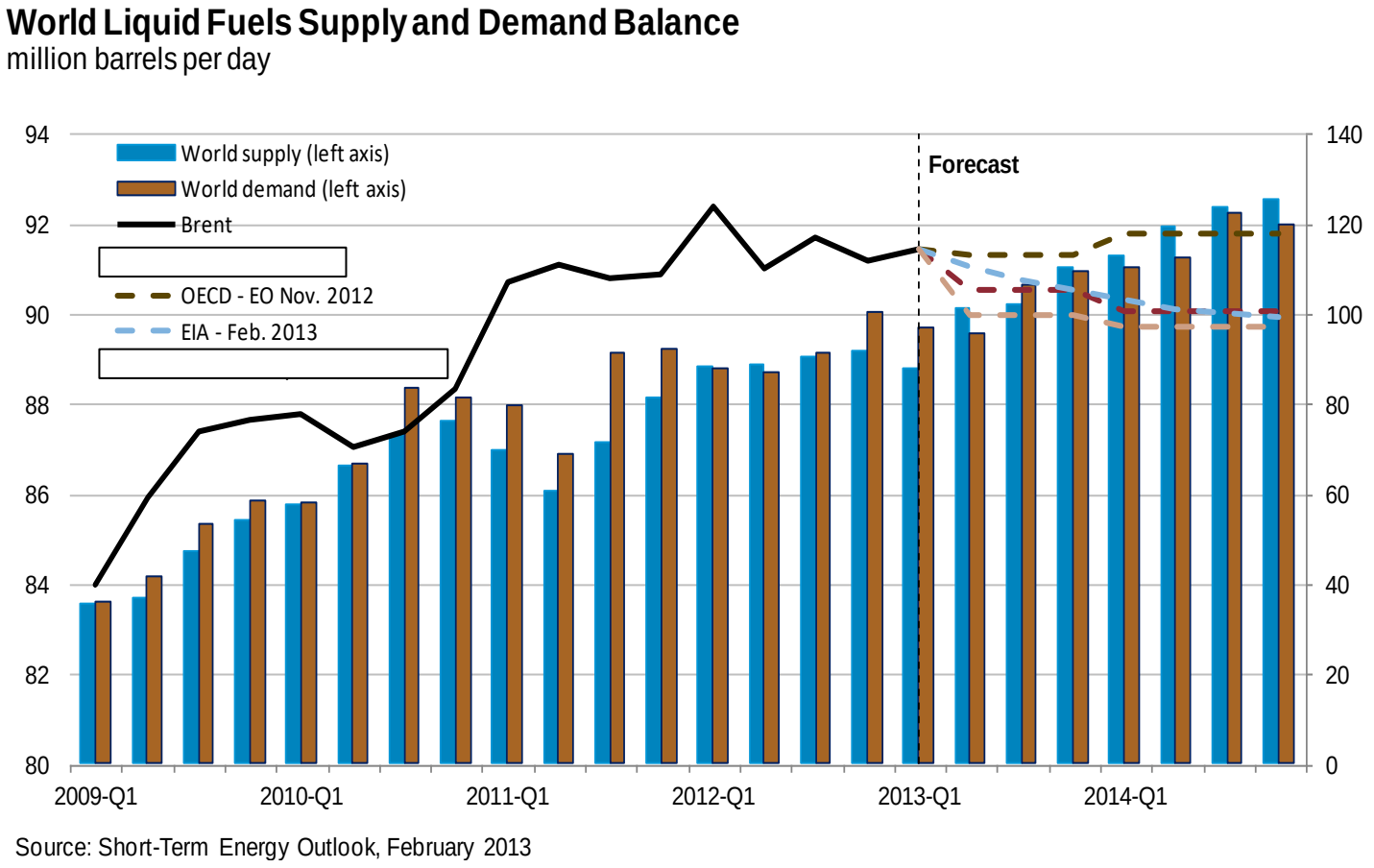
M1 - Cours de Modélisation Appliquée

Modèle de régression linéaire multivarié

Laurent Ferrara

Février 2019

Exemple: Consommation mondiale du pétrole



Exemple: Consommation mondiale du pétrole

Expliquer la conso mondiale de pétrole C(t) (en Δlogs) par :

P(t): Prix du pétrole (en Δlogs)

PIB(t) : Demande de pétrole (en Δlogs)

à l'aide du modèle suivant :

$$C(t) = b_0 + b_1 P(t) + b_2 PIB(t) + \varepsilon(t)$$

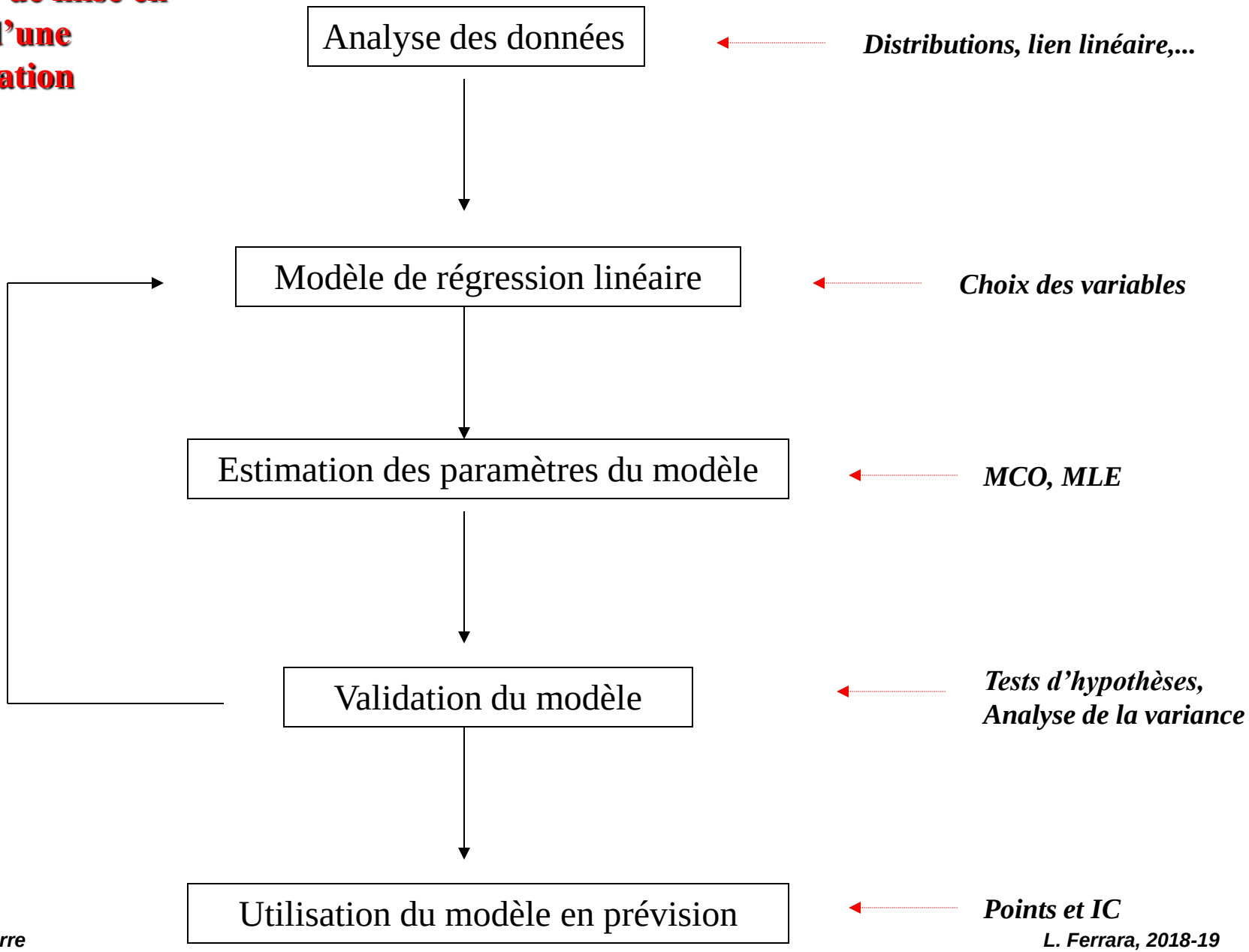
b_0

b_1 = élasticité-prix

b_2 = élasticité-revenu

$$C_t = \underset{(0.43)}{1,34} - \underset{(0.06)}{0,07} P_t + \underset{(0.02)}{2,23} PIB_t,$$

Schéma de mise en œuvre d'une modélisation



Modèle linéaire multivarié

Soit p+1 variables continues Y et X^1, …, X^p, . On observe les unités expérimentales : $(y_i, x_i^1, ..., x_i^p)$ pour i = 1, …, *n*.

Le modèle linéaire s'écrit sous forme matricielle:

$$Y = X\,b + \varepsilon$$

avec

$$Y = (y_1, ..., y_n)^t \quad (n \times 1)$$

$$b = (b_0, b_1, ..., b_p)^t \quad (p+1 \times 1)$$

$$\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^t \quad (n \times 1)$$

et :

$$X = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^k & \dots & x_1^p \\ \dots & & & & & \\ 1 & x_i^1 & \dots & x_i^k & \dots & x_i^p \\ \dots & & & & & \\ 1 & x_n^1 & \dots & x_n^k & \dots & x_n^p \end{pmatrix} \quad (n \times (p+1))$$

Hypothèses du modèle linéaire :

- **H1** : E(Y) fonction linéaire des X^1, …, X^p .
- **H2** : Les erreurs, ε_i, sont indépendantes entre elles
- **H3** : $E(\varepsilon_i) = 0$, les erreurs sont d'espérance nulle
  (en moyenne le modèle est bien spécifié)

- **H4** : $E(\varepsilon^2_i) = \sigma^2$, les erreurs sont de variance égale pour toute valeur de X

  (hypothèse d ’homoscédasticité)

- **H5** : $E(X_i\ \varepsilon_i) = 0$, les erreurs,sont indépendantes des valeurs de X

- **H6** : Hypothèse de Normalité

  Les erreurs, ε_i, sont identiquement distribuées selon la loi Normale.

Hypothèses supplémentaires structurelles

- **H7** : Absence de colinéarité entre les X^1, …, X^p .
- **H8** : *(X'X) / n* tend vers une matrice finie non singulière lorsque n tend vers l 'infini
- **H9** : n > p+1

Estimation des paramètres

- Objectif : estimer le vecteur b

- Par les MCO, on minimise la forme quadratique :

$$Q(b) = \sum_{i=1}^{n} \varepsilon_i^2 = (Y - bX)^t (Y - bX)$$

$$\rightarrow \quad \frac{\partial Q(b)}{\partial b} = 2X^t Y + 2X^t Xb = 0$$

Et :

$$\hat{b} = (X^t X)^{-1} X^t Y$$

Solution réalisable si la matrice carrée X^tX est inversible !!!

$\rightarrow$ des hypothèses sont nécessaires

En cas de colinéarité parfaite entre 2 variables explicatives, cette matrice est singulière et la méthode des MCO est défaillante.

Le modèle estimé s'écrit donc :

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i^1 + \ldots + \hat{b}_p x_i^p$$

Soit :

$$\hat{Y} = X\hat{b} = X(X^t X)^{-1} X^t Y$$

ie:

$$\hat{Y} = HY$$

L’erreur de prévision (ou résidu) est donnée par :

$$e_i = y_i - \hat{y}_i$$

Soit :

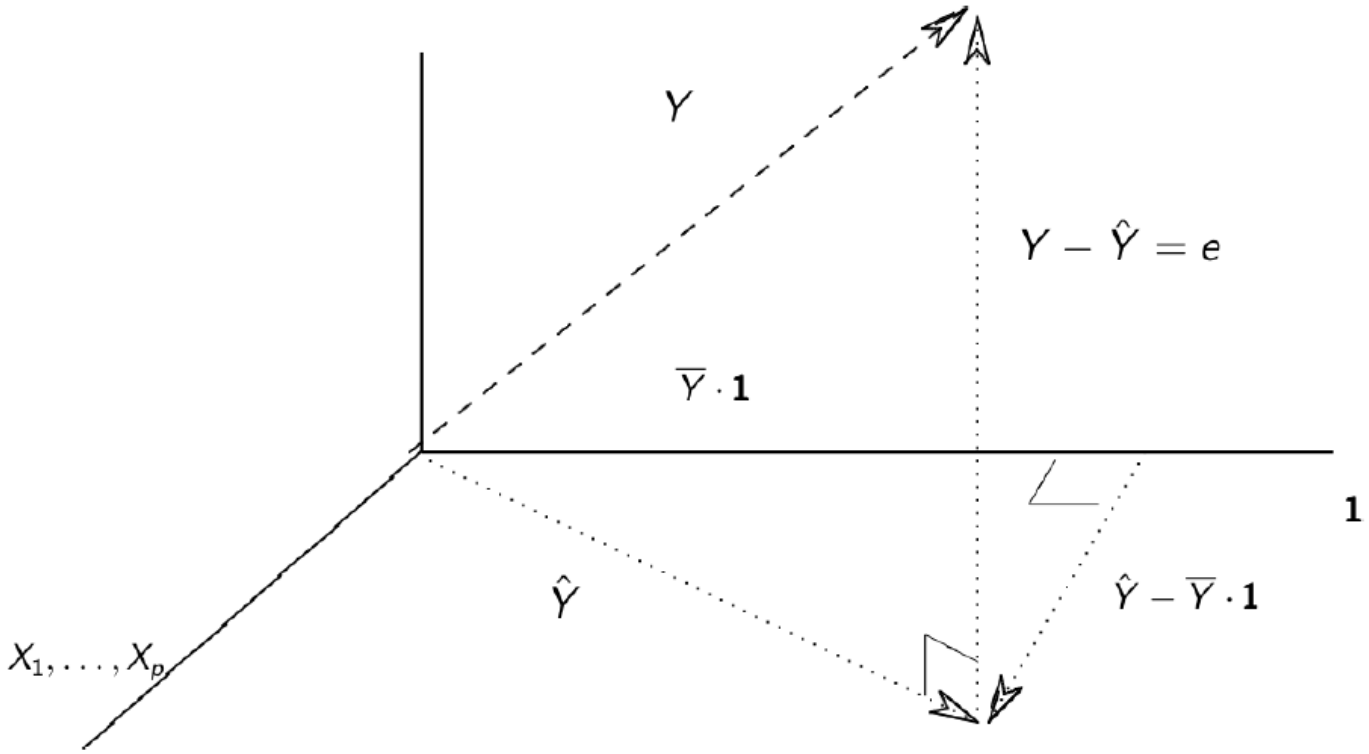
$$e = (I - H)Y$$

<u>Remarques :</u>

R1 : Il faut distinguer l’erreur inobservable du modèle (ε) et le résidu (e) qui lui est estimé

R2: En termes géométriques, le vecteur (e) est la projection orthogonale sur le sous-espace vectoriel Vect(X)

Interprétation géométrique



Propriétés des estimateurs

- L ’estimateur $\hat{b}$ est le meilleur estimateur non-biaisé de b au sens où sa variance est la plus faible possible et
- On mq :

$$V(\hat{b}) = \sigma_{\varepsilon}^{2}(X^{t}X)^{-1}$$

- Un ESB de la variance résiduelle est donné par :

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$$

Propriétés des estimateurs

- Sous l'hypothèse de Normalité, l' EMV coïncide avec le l'estimateur MCO mais est un estimateur efficace;

  ie: sa matrice des variances-covariances atteint la borne de Cramer –Rao

- L'estimateur de la variance résiduelle suit une loi :

$$\hat{\sigma}_{\varepsilon}^{2} \approx \sigma^{2} \frac{Chi^{2}(n-p-1)}{n-p-1}$$

Validation: Somme des carrés

- SST = Sum of Squares Total

$$SST = \left\| Y - \bar{y}1 \right\|^2 = Y^t Y - n\bar{y}^2$$

- SSR = Sum of Squares Regression

$$SSR = \left\| \hat{Y} - \bar{y}1 \right\|^2 = \hat{Y}^t \hat{Y} - n\bar{y}^2 = b^t X^t Y - n\bar{y}^2$$

- SSE = Sum of Squared Errors

$$SSE = \left\| Y - \hat{Y} \right\|^2$$

=> SST = SSR+SSE

Validation: Coefficient de détermination

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

- Le coefficient de détermination est la part de variation de Y expliquée par le modèle, ie : il doit être le plus proche de 1

- Remarque:

  l'ajout de variables explicatives augmente automatiquement ce coefficient de détermination

Validation: Coefficient de détermination ajusté

$$R^2_{Adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- On pondère par le nombre de paramètres à estimer.

Validation: Tests sur les paramètres

- On montre que la statistique T suit une loi de Student à (n-p-1) ddl:

$$T = \frac{\hat{b}_j - b_j}{\sigma_{b_j}}$$

Racine du jième terme diagonal de la matrice de variance-cov des paramètres estimés

- On utilise T pour tester H0: $b_j = 0$

- Un intervalle de confiance à (1-α) est donné par:

$$b_j \pm t_{\alpha/2,(n-p-1)}\sigma_{b_j}$$

Validation: Tests du modèle global

- On peut tester globalement l'hypothèse nulle:

$$H0: \quad b_1 = b_2 = ...b_p = 0$$

- On utilise la statistique:

$$F = \frac{SSR / p}{SSE /(n - p - 1)}$$

qui suit une loi de Fischer à (p, n-p-1) ddl

Validation: Tests d'un modèle réduit

- On peut tester l'hypothèse nulle d'un modèle réduit à q<p variables explicatives:

$$H0: \quad b_{q+1} = b_{q+2} = ... = b_p = 0$$

- Sous H0, on utilise la statistique:

$$F = \frac{(SSE_q - SSE_p)/q}{SSE_p/(n-p-1)}$$

qui suit une loi de Fischer à (q, n-p-1) ddl.

- L'ajout des (p-q) variables explicatives est justifié si $(SSE_q - SSE_p)$ est « suffisamment grand ».

Prévision

- Soit une nouvelle observation: $X_0 = (x_0^1, ..., x_0^p)^t$
- Prédicteur : $\hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0^1 + ... + \hat{b}_p x_0^p$
- IC pour Y:

$$\hat{y}_0 \pm t_{\alpha/2,(n-p-1)} \sigma_\varepsilon (1 + v_0^t (X^t X)^{-1} v_0)^{1/2}$$

- IC pour E(Y):

$$\hat{y}_0 \pm t_{\alpha/2,(n-p-1)} \sigma_\varepsilon (v_0^t (X^t X)^{-1} v_0)^{1/2}$$

$$v_0 = (1, x_0^1, ..., x_0^p)$$

Extensions

Effet croisé:

$$y_i = b_0 + b_1 x_i^1 + b_2 x_i^2 + \gamma x_i^1 x_i^2 + \varepsilon_i$$

Effet non-linéaire:

$$y_i = b_0 + b_1 z_i + b_2 z_i^2 + b_3 z_i^3 + \varepsilon_i$$

Exemple: IMF Working Paper,

« Walking Hand in Hand: Fiscal Policy and Growth in Advanced Economies » by Cotarelli and Jaramillo (2012)

Problème de politique économique:
La consolidation fiscale et budgétaire dans les pays avancés après la récession 2008-09 pèse sur la croissance de court terme mais semble nécessaire pour favoriser la croissance à long terme via une baisse de la dette publique et une baisse des taux longs souverains (spreads = écarts de taux).

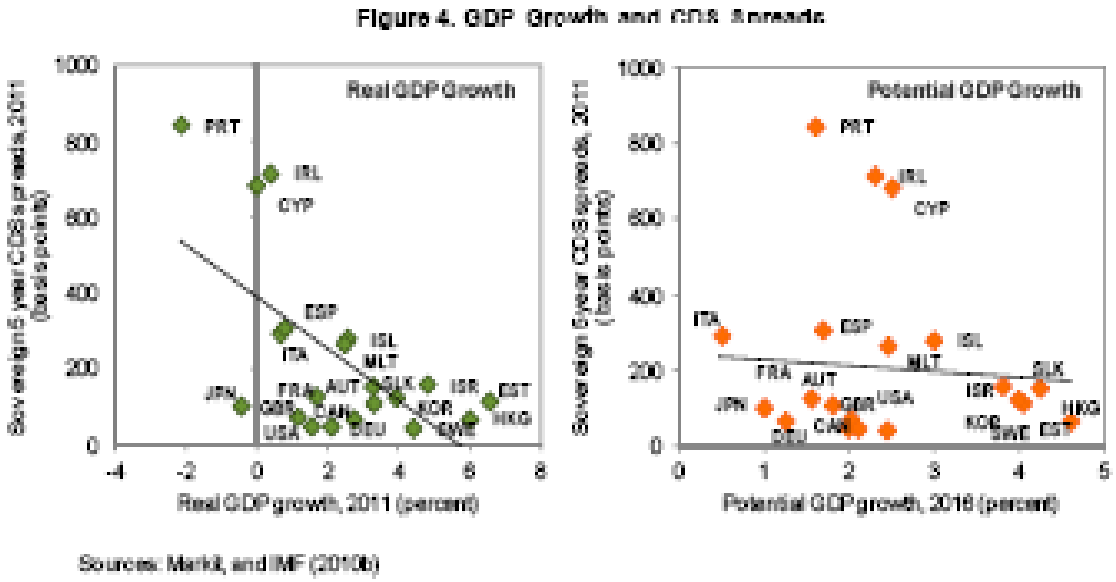
Equation de relation entre dette / taux longs / croissance :

$$d_t - d_{t-1} = \left(\frac{r_t - \pi_t - g_t}{1 + g_t} \right) d_{t-1} - p_t$$

But du modèle linéaire: Rechercher les déterminants des spreads

Exemple: IMF Working Paper,

« Walking Hand in Hand: Fiscal Policy and Growth in Advanced Economies » by Cotarelli and Jaramillo (2012)



Sources: Markit, and IMF (2010b)

Exemple: IMF Working Paper,

« Walking Hand in Hand: Fiscal Policy and Growth in Advanced Economies » by Cotarelli and Jaramillo (2012)

Table A1. Determinants of CDS Spreads in Advanced Economies, Cross Section Analysis 2011

	(1)	(2)	(3)	(4)	(5)
Gross debt to GDP 2011	0.0124**	0.0126**	0.0120***	0.0120***	0.0122***
	(2.735)	(2.773)	(2.848)	(2.844)	(2.931)
Primary balance to GDP 2011 for Euro Area	-0.172***	-0.177***	-0.182***	-0.187***	-0.194***
	(-3.581)	(-3.661)	(-3.578)	(-3.602)	(-3.800)
Real GDP growth 2011	-0.210**	-0.241***	-0.242***	-0.239***	-0.230***
	(-2.208)	(-3.502)	(-3.631)	(-3.699)	(-3.860)
Real GDP growth squared	0.0359**	0.0348**	0.0342**	0.0327**	0.0323**
	(2.786)	(2.495)	(2.554)	(2.577)	(2.545)
Debt held by a country's central bank or by foreign central banks to GDP	-0.0261	-0.0248	-0.0205	-0.0222	-0.0272*
	(-1.157)	(-1.078)	(-1.185)	(-1.348)	(-1.842)
Inflation rate 2011	0.263**	0.255**	0.241***	0.244***	0.222***
	(2.659)	(2.566)	(2.935)	(3.113)	(2.907)
NPV of health spending to GDP 2010	-0.00165	-0.00200	-0.00229	-0.00196	
	(-0.508)	(-0.668)	(-0.868)	(-0.766)	
NPV of pension spending to GDP 2010	0.00174	0.00124	0.00111		
	(0.422)	(0.307)	(0.265)		
Primary balance to GDP 2014 for Euro Area	-0.0290	-0.0208			
	(-0.444)	(-0.307)			
Potential output growth, average 2011-2016	-0.103				
	(-0.520)				
Constant	3.529***	3.485***	3.551***	3.576***	3.490***
	(6.312)	(6.133)	(6.805)	(7.030)	(7.495)
Observations	31	31	31	31	31
R-squared	0.769	0.765	0.764	0.763	0.760

Robust t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1