

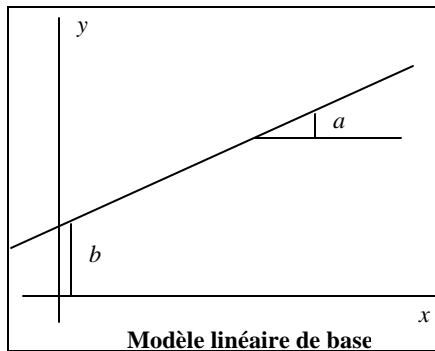
NOTE SUR L'ANALYSE DE RÉGRESSION (présentation simplifiée)

1. MODÈLE

Le modèle de base de l'analyse de régression linéaire est le suivant :

$$y = b + a_1x_1 + a_2x_2 + \dots + a_px_p + e$$

où y est la variable dépendante (ou à expliquer).
 $x_1 \dots x_p$ sont les variables indépendantes (explicatives).
 $a_1 \dots a_p$ sont les coefficients angulaires des variables indépendantes.
 b est la constante de régression.
 e est un résidu.



Fondamentalement, ce modèle pose donc l'hypothèse d'une relation linéaire de cause à effet entre la variable dépendante et les variables indépendantes.

2. RÉSULTATS

Excel fournit les résultats suivant lorsque l'on réalise une régression multiple (sortie du Data Analysis ToolPack) :

	y	x1	x2			
	100	20	34			
	200	21	37			
	300	25	26			
	400	24	28			
Regression Statistics						
Multiple R	0.88207712					
R Square	0.778060046					
Adjusted R Square	0.334180139					
Standard Error	105.3422882					
Observations	4					
Analysis of Variance						
	df	Sum of Squares	Mean Square	F	Significance F	
Regression	2	38903.00231	19451.50115	1.7528616	0.471105035	
Residual	1	11096.99769	11096.99769			
Total	3	50000				
	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	-1559.46882	2304.736401	-0.67663652	0.54713621	-30843.7959	27724.8583
x1	66.62817552	63.53295841	1.04871829	0.37134396	-740.631143	873.887494
x2	9.930715935	29.51877007	0.336420383	0.75871167	-365.139213	385.000645

3. INTERPRÉTATION

1. R Square :

Le R carré (ou R^2) est une mesure de la précision de l'ajustement de la droite de régression. Il s'agit du rapport entre la variation de la variable dépendante (mesurée par sa variance) expliquée par le modèle de régression et sa variation totale (mesurée par sa variance).

Exemple : $R^2 = 35\%$ signifie que 35% des variations de la variable dépendante sont expliqués par le modèle de régression et que 65% restent par conséquent inexpliqués.

La racine carrée du coefficient R^2 (également appelé coefficient de détermination) donne le coefficient de corrélation.

2. Adjusted R Square :

Le R^2 ajusté est utilisé en cas de régression multiple. Il s'interprète de la même manière que le R^2 mais tient compte de l'augmentation du nombre de variables explicatives.

3. F (ou test de Fisher) :

Le test de Fisher mesure le rapport entre la variance de la variable dépendante expliquée et non-expliquée par le modèle de régression. L'hypothèse que ce test tente d'évaluer (hypothèse nulle) est que le rapport entre la variance expliquée par le modèle est (approximativement) égale à la variance qui reste inexpliquée (auquel cas $F \sim 1$).

Exemple : $F = 1.75$ signifie que la part de la variance de la variable dépendante expliquée par le modèle est 1.75 fois plus importante que la part de la variance de la variable dépendante qui reste inexpliquée.

4. Significance F :

Le test de Fisher permet donc de tester l'hypothèse (appelée hypothèse nulle) selon laquelle la variance expliquée est égale à la variance inexpliquée (cas où $F \sim 1$). "Significance F" donne probabilité d'observer, si l'hypothèse nulle est vérifiée, un F supérieur ou égal au F calculé.

Exemple : "Significance F" = 0.47 signifie qu'il y a 47 chances sur 100 que l'on observe, sur un échantillon donné, un F supérieur ou égal au F calculé sachant que l'hypothèse nulle ($F \sim 1$) est vraie.

5. Coefficients

Les coefficients Intercept, x1, x2 donnent les valeurs des coefficients estimés \hat{b} , \hat{a}_1 et \hat{a}_2, \dots du modèle repris ci-dessus.

6. t Statistic (test de Student)

La statistique t permet de tester l'hypothèse (nulle) selon laquelle la valeur des coefficients de régression ne sont pas significativement différents de 0 (en d'autres termes, qu'il existe bien une relation entre la variable dépendante et la variable indépendante en question). La valeur que doit atteindre le test de Student pour que l'on puisse rejeter l'hypothèse nulle dépend du nombre d'observations et du niveau de confiance recherché (de 90% à 99% en général). En pratique, la valeur critique oscille le plus souvent autour de 2.

7. P-Value

P-Value donne la probabilité que le coefficient ait une valeur nulle compte tenu de la valeur de la statistique t.

Exemple : P-Value = 0.37 signifie qu'il y a 37 chances sur 100 pour que la vraie valeur du coefficient en question soit nulle.

8. Lower et Upper x%

Ces deux valeurs donnent les bornes de l'intervalle de confiance dans lequel se situe la vraie valeur du coefficient en question compte tenu du niveau de confiance auquel on travaille.

Exemple : Lower 95% = -740 et Upper 95% = 873 signifie qu'il y a 95% de chance que la vraie valeur du coefficient en question se situe entre -740 et 873.

4. REMARQUES

- les modèles de régression reposent sur une relation de causalité. Il est important, dans le cadre d'une utilisation en finance, qu'ils ne soient pas construits à la suite d'une simple observation de régularités statistiques inexplicables (*data mining*) mais qu'ils se fondent sur une justification théorique de la relation entre les variables choisies.
- le modèle étant déterminé, il peut être utilisé à des fins de prévisions. Il suffit pour cela d'y introduire la valeur attendue des variables indépendantes et d'observer la valeur correspondante de la variable dépendante.
- la fonction "Forecast" d'Excel repose bien sur un modèle de régression linéaire. La variable explicative est le temps.
- les modèles de régression linéaire ne permettent, comme leur nom l'indique, que de saisir des relations linéaires. Un R^2 très faible ne permet donc que de conclure à l'absence de relation linéaire entre les variables concernées (et non à l'absence de relation tout court). Le problème classique de parité (ou odd/even) illustre de manière cruelle ce propos.
- le modèle de régression linéaire repose sur un certain nombre d'hypothèses. On rappellera l'absence de collinéarités trop fortes entre les variables indépendantes (ce que l'on vérifiera à un premier niveau au moyen d'une matrice de corrélation), des résidus de moyenne nulle, de variance constante et non auto-corrélés (ensemble de conditions qui pourront être testées par un plot des résidus, au niveau duquel on vérifiera plus particulièrement l'absence de comportements systématiques dans les résidus, et par l'utilisation du test de Durbin Watson).