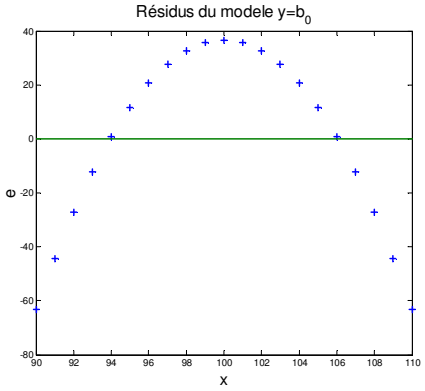
Sélection de variables

Cas 1: $Y=(X-100)^2$

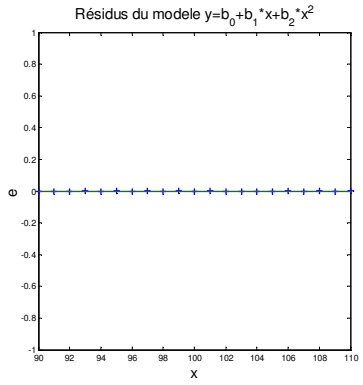
a) Sélection avant et « stepwise »:

La variable la plus corrélée est x^2, mais SCRm=16.3 et SCE=22416. L'ajout de la variable x^2 n'est donc pas significatif puisque (16.3/1)/(22416/19)=.0138, ce qui est très inférieur à une $F_{1,19,.05}$=4.38. Donc, aucune variable n'est sélectionnée mais les résidus sont inacceptables car ils montrent clairement qu'une tendance quadratique a été oubliée dans le modèle.



Supposons que l'on force le modèle $Y=b_0+b_1X+b_2X^2+e$.

On obtient les résidus suivants :



b) Élimination arrière : aucune variable ne peut être éliminée -> on a le bon modèle.

c) Statistique C_p : avec x et x^2 : C_p=3
avec x, $C_p=1.1*10^{28}$
avec x^2, $C_p=1.1*10^{28}$

Clairement seul le modèle avec x et x^2 est adéquat, donc on a le bon modèle.

2e cas : $Y=X^3+300*x+N(0,1)$

a) Sélection avant :

Étape	Variable	F ajout	F table (.05)
1	x^2	8080	4.38
2	x	6.5	4.41
3	x^3	58000	4.45

Les 3 ajouts sont significatifs. Le R^2 vaut 1. Toutefois, l'examen des coefficients et de leurs écarts-types révèle que le coefficient associé à x^2 et à la constante n'est pas significativement différent de zéro :

Variable	b	Écart-type
x	300.2	0.34
x^2	-.029	0.04
x^3	1.00	0.001
cte	-0.44	0.764

En reprenant la régression sans la constante et sans le x^2, on obtient à nouveau un R^2 de 1, et les coefficients de x et x^3 sont fortement significatifs.

b) Élimination arrière :

SCE avec cte, x, x^2 x^3 : 18.23
SCE avec cte, x, x^3 : 18.80
SCE avec x, x^3 : 18.81

On peut éliminer x^2 car $\frac{(18.80-18.23)/1}{18.23/17} = 0.53 < F_{1,17,.05} = 4.45$.

On peut éliminer la constante, car : $\frac{(18.81-18.80)/1}{18.80/18} = 0.01 < F_{1,18,.05} = 4.41$

c) Stepwise : on inclut x^2, x, x^3 puis on élimine x^2 et on élimine la constante.

d) Statistique C_p :

Modèle	C_p	Modèle	C_p
cte, x, x^2, x^3	4	x, x^2, x^3	3
cte, x, x^2	$5.8*10^5$	x, x^2	$9.2*10^5$
cte, x,x^3	2.5	x,x^3	1.2
cte, x^2, x^3	$7.8*10^5$	x^2, x^3	$8*10^5$
cte, x^1	$1.9*10^7$	x^1	$3.4*10^7$
cte, x^2	$7.9*10^5$	x^2	$3.9*10^6$
cte, x^3	$1.1*10^7$	x^3	$3.9*10^7$

Les seuls modèles acceptables sont (cte, x, x^2 et x^3), (cte, x, x^3), (x,x^2,x^3) et (x,x^3), celui-ci étant le meilleur de tous au sens de la statistique C_p. Si l'on examine le modèle cte, x, x^2 x^3, on constate que la constante et le coefficient pour b^2 ne sont pas significatifs puisque l'intervalle de confiance inclut le 0 au niveau α=0.05. (b_0/s_{b0}=-0.57; b_1/s_{b1}=885; b_2/s_{b2}=-0.7; b_3/s_{b3}=763).

Note : Les coefficients trouvés avec x et x^3 (modèle sans constante) sont b_1=299.9 et b_3=1.00, soit presque exactement le modèle théorique. Le R^2 vaut presque 1.

Les résidus ne montrent pas d'anomalies particulières :