

Valeurs manquantes, estimations manquées?

Nicolas Corneau-Tremblay

août 2016

Ce document présente (brièvement) sous quelles conditions l'estimation de modèles de régression à partir d'un échantillon contenant des valeurs manquantes pose et ne pose pas problème.

1 Le cas général

Soit un modèle de régression

$$y = g(X) + u$$

où $g(\cdot)$ est une forme générale définissant la relation entre y et X . L'espérance de y conditionnelle à X peut être écrite telle que

$$E(y|X) = g(X) + E(u|X)$$

Sous l'hypothèse que tous les éléments de X sont orthogonaux au terme d'erreur, c'est-à-dire $E(u|X) = 0$, l'estimation de

$$E(y|X) = \hat{g}(X)$$

peut être faite sans biais à l'aide d'un estimateur adéquat¹.

Soit à présent un échantillon contenant des valeurs manquantes pour certaines observations. Soit également s , une variable indicatrice prenant la valeur 0 si l'une des variables $\{y, X\}$ d'un individu est non-observée et 1 autrement. La sélection des valeurs manquantes peut être due à un processus aléatoire affectant soit y ou X (*missing completely at random, MCAR*). Puisque ce processus survient de façon aléatoire, il n'est lié à aucune caractéristique particulière des individus. L'estimation de $g(\cdot)$ peut donc être faite sans risque de biais, puisque $E(u|X) = 0$ est toujours respectée.

La sélection peut aussi être due à un processus non-aléatoire qui est fonction des autres variables considérées dans le modèle de régression. Si la sélection est fonction des variables indépendantes contenues dans X (*missing at random, MAR*), il est alors possible d'écrire la fonction suivante

$$s = h(X)$$

où s , le fait pour un individu d'avoir une valeur manquante, suit une fonction $h(\cdot)$ qui dépend des variables indépendantes X . Dans ce cas,

$$E(u|X, s) = E(u|X, h(X)) = E(u|X)$$

Puisque la variable s dépend uniquement des variables indépendantes X à travers sa fonction $h(\cdot)$, l'espérance de u conditionnellement à X n'est pas affectée par s , le processus de sélection. Ceci s'explique par le fait que lorsque u est considéré en maintenant X fixe, s ne contient aucune variation, puisque lui même ne varie qu'en fonction de X . Le problème de régression demeure alors

$$E(y|X, s) = E(y|X) = g(X) + E(u|X)$$

¹D'autres hypothèses sont également nécessaires pour effectuer cette estimation sans biais, mais leur importance est secondaire dans la présente discussion.

et peut être estimé sans biais, toujours sous l'hypothèse que $E(u|X) = 0$.

Un problème dans l'estimation de la régression survient lorsque s est fonction de y (*not missing at random, NMAR*), par exemple lorsque

$$s = h(y, X)$$

Dans ce cas alors

$$E(u|X, s) = E(u|X, h(y, X)) \neq E(u|X)$$

puisque cette fois s peut varier avec y et ainsi faire varier u . Il est —à noter que cela survient même lorsque u est considéré en maintenant X fixe et lorsque tous les éléments de X sont orthogonaux à u . Le modèle de régression souffre alors d'endogénéité et son estimation risque d'être biaisée, puisque

$$E(y|X, s) = g(X) + E(u|X, s)$$

où

$$E(u|X, s) \neq 0$$

2 Le cas linéaire

Cette section investigate plus particulièrement le problème de sélection des variables manquantes pour le modèle de régression linéaire. Les résultats de la section précédente y sont développés pour ce cas particulier. Soit un modèle linéaire

$$y = X\beta + u \tag{1}$$

où $u \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$ et où l'hypothèse $E(u|X) = 0$ est respectée. Le cas linéaire où $g(X) = X\beta$ est un cas particulier de la forme générale présentée dans la section précédente. Soit également $s = h(\cdot)$ le processus de sélection des valeurs manquantes. Par exemple, soit

$$s = \mathbb{1}\{X < c\}$$

une fonction indicatrice prenant la valeur 1 si la valeur de X est en dessous d'une constante c , et 0 si elle est égale ou supérieure à c . Si $s = 0$, alors l'individu possède une valeur manquante en X , ce qui est un cas de sélection des valeurs manquantes sur les variables indépendantes (*MAR*). Qu'advient-il alors du modèle à estimer? Si l'on prend $E(y|X, s)$, l'équation (1) devient

$$E(y|X, s) = E(y|X, X < c) = X\beta + E(u|X, X < c)$$

Rappelons qu'un biais survient si $E(u|X, X < c) \neq 0$. Conséquemment au développement suivant

$$\begin{aligned} E(u|X, X < c) &= E[E(u|X = x)] \quad \forall \quad X < c \quad \text{par la loi des espérances itérées} \\ &= E[0] \quad \text{puisque } E(u|X = x) = 0 \text{ sous l'hypothèse } E(u|X) = 0 \\ &= 0 \end{aligned}$$

il est possible de voir que le problème de biais ne se pose pas dans un modèle linéaire lorsque la sélection des valeurs manquantes est faite sur les variables indépendantes.

Soit à présent

$$s = \mathbb{1}\{y < c\}$$

Dans ce cas, la sélection s'effectue sur la variable dépendante (*NMAR*). L'équation (1) devient cette fois

$$E(y|X, s) = E(y|X, y < c) = X\beta + E(u|X, y < c)$$

Encore une fois, un biais survient si $E(u|X, y < c) \neq 0$. Dans ce cas

$$\begin{aligned} E(u|X, y < c) &= E(u|X, X\beta + u < c) \\ &= E(u|X\beta + u < c) \\ &= E(u|u < c - X\beta) \\ &= \sigma E\left(\frac{u}{\sigma} \mid \frac{u}{\sigma} < \frac{c - X\beta}{\sigma}\right) \\ &= \sigma \left[\frac{\phi(c - X\beta)}{\Phi(c - X\beta)} \right] \text{ sous l'hypothèse } u \stackrel{i.i.d.}{\sim} N(0, \sigma^2) \end{aligned}$$

Alors $E(y|X, s)$ devient

$$\begin{aligned} E(y|X, s) &= X\beta + E(u|X, y < c) \\ &= X\beta + \sigma \left[\frac{\phi(c - X\beta)}{\Phi(c - X\beta)} \right] \end{aligned}$$

où $\phi(\cdot)$ et $\Phi(\cdot)$ sont respectivement la fonction de densité et la fonction de répartition de la loi normale centrée réduite. Il est possible de voir que $E(u|X, y < c) \neq 0$, et donc que l'estimation d'un modèle linéaire qui ne prend pas en compte la sélection sur y mène à des paramètres biaisés.

3 Simulation

Il est possible d'explorer ces preuves théoriques à l'aide d'une simulation.

3.1 Régressions linéaires simples

Soit le modèle de régression

$$y = \alpha + \beta_1 \textit{exposition} + \beta_2 x + \beta_3 \textit{interaction} + u \quad (2)$$

où

$$\begin{aligned} \textit{exposition} &\sim N(0, 1) \\ x &\sim N(0, 1) \\ \textit{interaction} &= \textit{exposition} * x \\ u &\stackrel{i.i.d.}{\sim} N(0, 1) \end{aligned}$$

et

$$\alpha = \beta_1 = \beta_2 = \beta_3 = 1$$

Les variables *exposition*, *x*, *interaction* et *y* ont été créées par simulation². En estimant le modèle de régression présenté en (2) à partir de données simulées, Les résultats suivant sont obtenus

²L'ensemble du code et de l'output *Stata* utilisé dans cette section et la suivante est disponible en Appendix du présent document.

Régression : modèle de base

Variable	Coefficient	(Std. Err.)
expo	1.019**	(0.033)
x	1.015**	(0.032)
inter	1.024**	(0.033)
Intercept	0.959**	(0.032)
<hr/>		
N	1000	
R ²	0.738	
F (3,996)	936.143	
<hr/>		
Significance levels :	† : 10%	* : 5% ** : 1%

Les résultats sont cohérents, puisque chaque paramètre est relativement près de sa vraie valeur.

Le premier type de valeur manquante abordé précédemment est celui dû à un processus aléatoire. Ainsi, en retirant de façon aléatoire le 1/4 de l'échantillon simulé, l'estimation des paramètres devient

Régression : sélection aléatoire (MCAR)

Variable	Coefficient	(Std. Err.)
expo	1.059**	(0.037)
x	1.041**	(0.038)
inter	1.031**	(0.040)
Intercept	0.971**	(0.036)
<hr/>		
N	750	
R ²	0.739	
F (3,746)	703.928	
<hr/>		
Significance levels : † : 10% * : 5% ** : 1%		

Comme anticipé, les résultats sont encore une fois cohérents.

Le second type de valeurs manquantes est lorsqu'une sélection sur les variables indépendantes survient. Dans le cas présent, le 1/4 de l'échantillon ayant les plus grandes valeurs pour x a été retiré. Les résultats sont

Régression : sélection sur x (MAR)

Variable	Coefficient	(Std. Err.)
expo	1.019**	(0.042)
x	0.972**	(0.051)
inter	1.028**	(0.053)
Intercept	0.932**	(0.042)
<hr/>		
N	750	
R ²	0.576	
F (3,746)	337.975	
<hr/>		
Significance levels : † : 10% * : 5% ** : 1%		

Malgré une sélection forte sur une variable indépendante et la perte de beaucoup d'observations, les estimations ne semblent pas être affectées. Il est à noter que ce résultat est vérifié même si la variable x est corrélée à la variable *exposition*.

Enfin, le dernier type de valeur manquante est celui où une sélection est faite sur la variable dépendante.

Dans le cas présent, le 1/4 de l'échantillon ayant les plus grandes valeurs pour y a été retiré. Les résultats obtenus sont les suivants

Régression : sélection sur y (NMAR)

Variable	Coefficient	(Std. Err.)
expo	0.746**	(0.040)
x	0.731**	(0.041)
inter	0.834**	(0.041)
Intercept	0.589**	(0.038)
<hr/>		
N	750	
R ²	0.442	
F (3,746)	196.821	
<hr/>		
Significance levels :	† : 10%	* : 5% ** : 1%

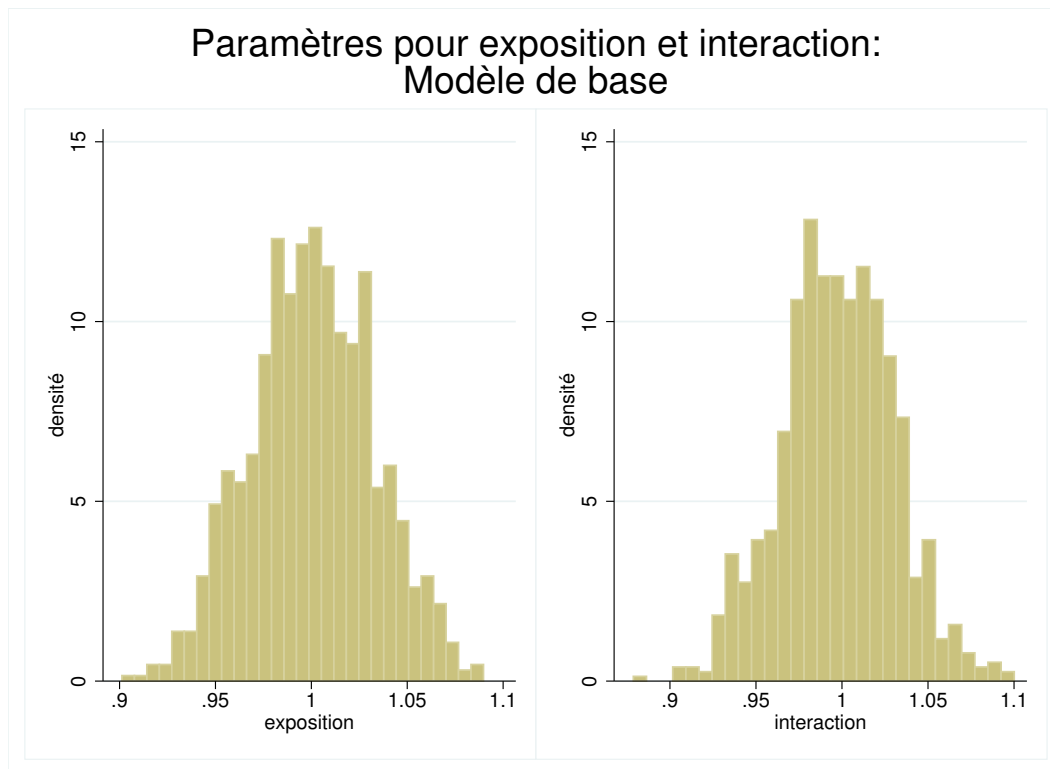
Dans ce dernier exemple, on voit bien que la sélection non-aléatoire sur y cause de lourds problèmes de biais. Les paramètres dévient fortement de leur vraie valeur. Ces résultats sont cohérents avec les présentations théoriques faites dans les sections précédentes.

3.2 Monte Carlo

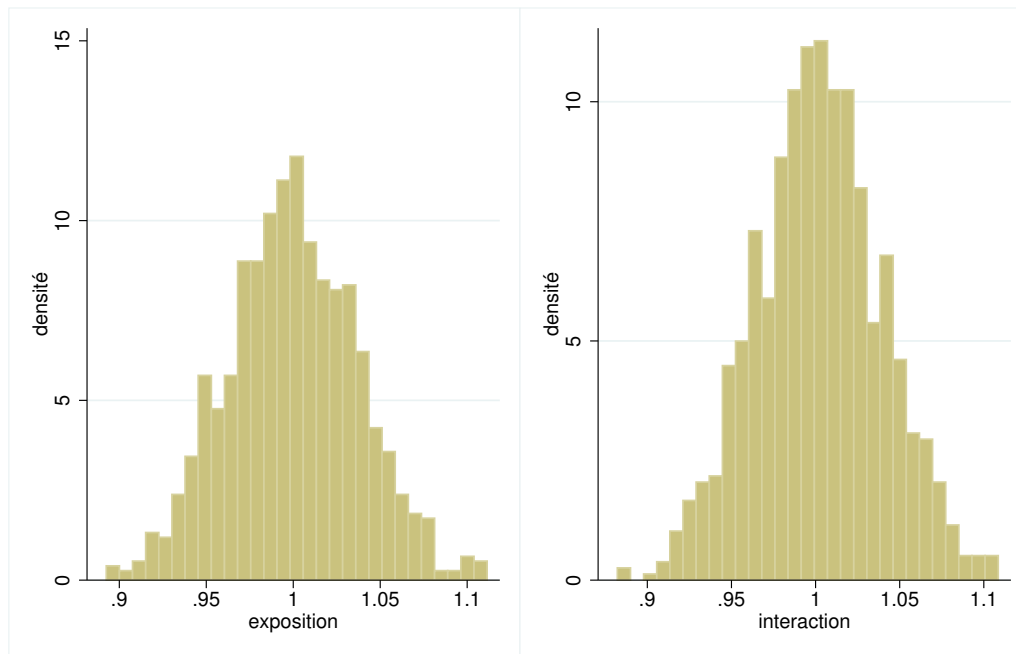
Soit le même modèle de régression qu'en (2)

$$y = \alpha + \beta_1 \text{exposition} + \beta_2 x + \beta_3 \text{interaction} + u$$

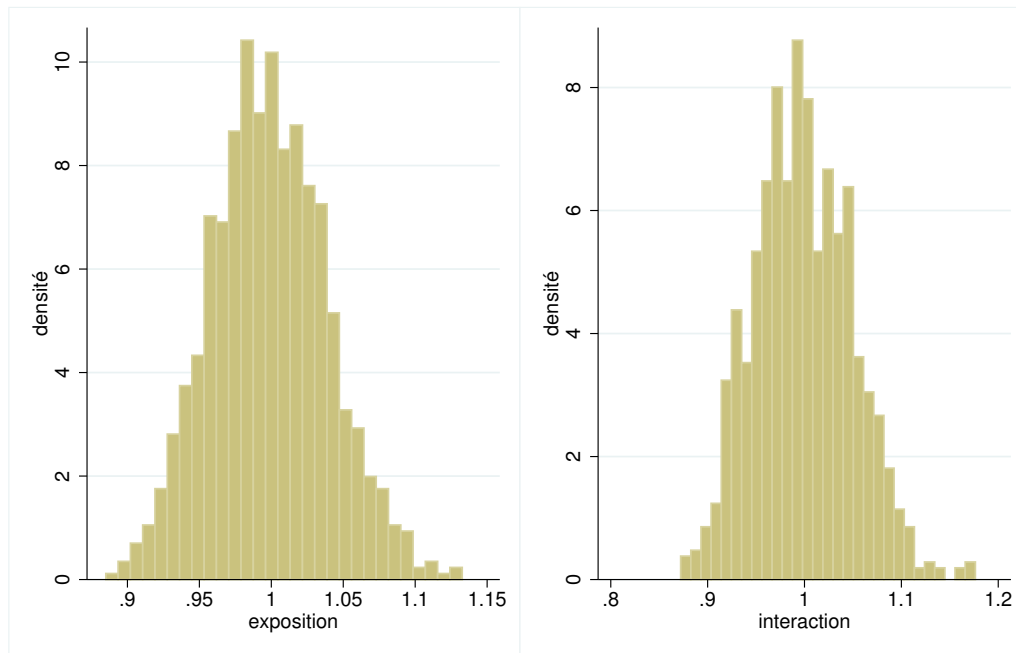
Cette sous-section présente les densités estimées des paramètres de régression pour *exposition* et *interaction* pour 1000 itérations. Elles ont été estimées pour chacun des types de sélections de la sous-section précédente.



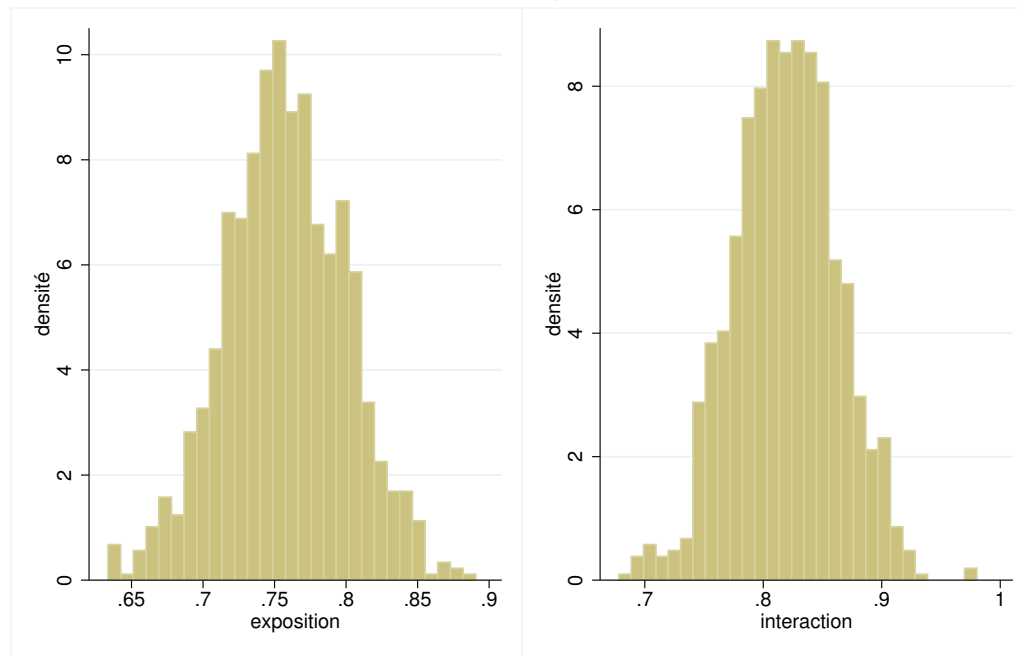
Paramètres pour exposition et interaction:
Sélection aléatoire (MCAR)



Paramètres pour exposition et interaction:
Sélection sur x (MAR)



Paramètres pour exposition et interaction: Sélection sur y (NMAR)



Cet exemple apporte une autre preuve que seul le processus de sélection sur y cause des problèmes systématiques de biais. En effet, comme le montrent les histogrammes, seul ce processus fait inmanquablement dévier les paramètres estimés de leur vraie valeur.

A Variable omise : le cas linéaire

Un problème similaire (mais distinct) à celui des valeurs manquantes est celui des variables omises. Le cas linéaire est abordé dans cette section, notamment les conditions nécessaires pour que l'omission d'une variable cause un problème de biais d'estimation dans les paramètres.

Soit, par exemple, le modèle linéaire

$$y = \alpha + \beta_1 X + \beta_2 s + u \quad (3)$$

où cette fois s n'est plus le processus de sélection des valeurs manquantes mais une simple variable. L'estimation de l'équation en (3) peut être faite par moindres carrés ordinaires, menant aux équations

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^N (X_i - \bar{X}) y_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ \hat{\beta}_2 &= \frac{\sum_{i=1}^N (s_i - \bar{s}) y_i}{\sum_{i=1}^N (s_i - \bar{s})^2} \\ \hat{\alpha} &= \bar{y} - (\hat{\beta}_1 \bar{X} + \hat{\beta}_2 \bar{s}) \end{aligned}$$

Sous l'hypothèse que les variables $\{X, s\}$ sont exogènes, l'estimation se fait sans biais. Cependant, si au lieu de l'équation (3), l'équation

$$y = \alpha + \beta_1 X + \tilde{u} \quad (4)$$

est estimée, où $\tilde{u} = \beta_2 s + u$, alors l'omission de la variable s dans l'estimation peut engendrer un biais dans les autres paramètres. Pour bien voir cela, il est possible d'écrire

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^N (X_i - \bar{X}) y_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^N (X_i - \bar{X}) (\alpha + \beta_1 X_i + \beta_2 s_i + u_i)}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ &= \beta_1 + \beta_2 \frac{\sum_{i=1}^N (X_i - \bar{X}) s_i}{\sum_{i=1}^N (X_i - \bar{X})^2} + \frac{\sum_{i=1}^N (X_i - \bar{X}) u_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \end{aligned}$$

Sous l'hypothèse que $\sum_{i=1}^N (X_i - \bar{X}) u_i = \text{cov}(X, u) = 0$, c'est-à-dire que X est bien exogène, il reste

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^N (X_i - \bar{X}) s_i}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

On voit alors que l'estimation de $\hat{\beta}_1$ sera biaisée et déviara de β_1 , sa vraie valeur, si

$$\beta_2 \frac{\sum_{i=1}^N (X_i - \bar{X}) s_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \neq 0$$

Pour qu'il y est effectivement un problème de biais, deux conditions doivent être remplies.

Tout d'abord, il faut que l'expression $\frac{\sum_{i=1}^N (X_i - \bar{X}) s_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \neq 0$. Cela signifie que $\text{cov}(X, s) \neq 0$, c'est-à-dire que la variable omise s est corrélée à X , la variable observée.

Ensuite, il faut que $\beta_2 \neq 0$. Cette seconde condition implique que la variable s est corrélée à la variable dépendante y .

Si ces deux conditions, à savoir que la variable omise s est corrélée à la fois à X et à y , ne sont *pas* satisfaites, alors la variable omise ne pose pas de risque de biais.

B Code *Stata*

```
name: <unnamed>
log: /Users/nicot/Dropbox (CEDIA)/ULaval/Travail/SLIM/Selection/select
> code.smcl
log type: smcl
opened on: 11 Aug 2016, 11:16:41

1 .
2 . //Simulation de regressions
3 . clear all

4 . cd "/Users/nicot/Dropbox (CEDIA)/ULaval/Travail/SLIM/Selection"
   /Users/nicot/Dropbox (CEDIA)/ULaval/Travail/SLIM/Selection

5 .
6 . set obs 1000
   number of observations (_N) was 0, now 1,000

7 . set seed 123

8 .
9 . *creation du modele
10 . gen expo=rnormal(0,1)

11 . gen x=rnormal(0,1)

12 . gen inter=expo*x

13 .
14 . gen error=rnormal(0,1)

15 .
16 . gen y=1+expo+x+inter+error

17 .
18 . sum y expo x inter
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	1,000	.9978072	1.987555	-6.321197	10.84289
expo	1,000	.0204975	.9719571	-4.604603	2.807839
x	1,000	.0202697	.9961863	-3.542322	3.40045
inter	1,000	-.002354	.9786722	-8.675727	5.957363

```

19 .
20 . regress y expo x inter

```

Source	SS	df	MS	Number of obs	=	1,000
Model	2913.24857	3	971.082857	F(3, 996)	=	936.14
Residual	1033.17408	996	1.03732337	Prob > F	=	0.0000
				R-squared	=	0.7382
				Adj R-squared	=	0.7374
Total	3946.42265	999	3.95037303	Root MSE	=	1.0185

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
expo	1.018943	.0331618	30.73	0.000	.953868	1.084018
x	1.015131	.032414	31.32	0.000	.9515233	1.078738
inter	1.023784	.0330023	31.02	0.000	.9590223	1.088546
_cons	.958755	.0322214	29.76	0.000	.8955253	1.021985

```

21 .
22 . *outtex, detail level legend title(R\'{e}gression : mod\'{e}le de base) key(
    > modele)
23 .
24 . *selection aleatoire
25 . gen toto = runiform()

26 . sort toto

27 . generate random = _n <= 750

28 .
29 . bysort random: summarize x

```

```

-> random = 0

```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	250	.0091655	1.104196	-3.023799	3.40045

```

-> random = 1

```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	750	.0239711	.9582475	-3.542322	3.139379

```

30 .
31 . regress y expo x inter if random==1

```

Source	SS	df	MS	Number of obs	=	750
Model	2098.52164	3	699.507212	F(3, 746)	=	703.93
Residual	741.315286	746	.993720222	Prob > F	=	0.0000
				R-squared	=	0.7390
				Adj R-squared	=	0.7379
Total	2839.83692	749	3.79150457	Root MSE	=	.99686

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
expo	1.059464	.037394	28.33	0.000	.9860545	1.132874
x	1.040988	.0380342	27.37	0.000	.9663211	1.115655
inter	1.031223	.0402996	25.59	0.000	.9521084	1.110337
_cons	.9707738	.0364255	26.65	0.000	.8992651	1.042283

```

32 .
33 . *outtex, detail level legend title(R\'{e}gression : s\'{e}lection al\'{e}ato
> ire (MCAR))
34 .
35 .
36 . *selection sur x, avec x et expo correles
37 . sort x

38 . gen nl=_n

39 . gen selectx=nl <= 750

40 .
41 . bysort selectx: summarize x

```

```

-> selectx = 0

```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	250	1.274757	.5667745	.6579035	3.40045

```

-> selectx = 1

```

Variable	Obs	Mean	Std. Dev.	Min	Max
x	750	-.3978927	.7186514	-3.542322	.657464

```

42 .
43 . regress y expo x inter if selectx==1

```

Source	SS	df	MS	Number of obs	=	750
Model	1006.48905	3	335.496351	F(3, 746)	=	337.98
Residual	740.528516	746	.99266557	Prob > F	=	0.0000
				R-squared	=	0.5761
				Adj R-squared	=	0.5744
Total	1747.01757	749	2.33246672	Root MSE	=	.99633

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
expo	1.019353	.0424256	24.03	0.000	.9360654	1.102641
x	.9721352	.0507898	19.14	0.000	.8724273	1.071843
inter	1.027505	.0529332	19.41	0.000	.9235897	1.131421
_cons	.9320326	.0416651	22.37	0.000	.8502378	1.013827

```

44 .
45 . *outtex, detail level legend title(R\'{e}gression : s\'{e}lection sur x (MAR
    > ))
46 .
47 . *selection sur y
48 . sort y
49 . gen n2=_n
50 . gen selecty=n2 <= 750
51 .
52 . bysort selecty: summarize y

```

```

-> selecty = 0

```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	250	3.634055	1.46297	2.097084	10.84289

```

-> selecty = 1

```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	750	.1190579	1.210155	-6.321197	2.069365

```

53 .
54 . regress y expo x inter if selecty==1

```

Source	SS	df	MS	Number of obs	=	750
Model	484.617296	3	161.539099	F(3, 746)	=	196.82
Residual	612.274102	746	.820742764	Prob > F	=	0.0000
				R-squared	=	0.4418
				Adj R-squared	=	0.4396
Total	1096.8914	749	1.4644745	Root MSE	=	.90595

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
expo	.7459536	.0402281	18.54	0.000	.6669798	.8249275
x	.731424	.0413839	17.67	0.000	.6501812	.8126669
inter	.8337301	.0405708	20.55	0.000	.7540836	.9133766
_cons	.5888732	.0384818	15.30	0.000	.5133277	.6644187

```

55 .
56 . *outtex, detail level legend title(R\'{e}gression : s\'{e}lection sur y (NMA
    > R))
57 .
58 .
59 . //Simulation MC
60 . *modele
61 . clear all

62 . capture program drop model

63 . program define model, rclass
    1.
64 . drop _all
    2. set obs 1000
    3.
65 . generate expo = rnormal(0,1)
    4. generate x = rnormal(0,1)
    5. generate inter=expo*x
    6. generate error = rnormal(0,1)
    7. generate y = 1 + expo + x + inter + error
    8.

```

```

66 . regress y expo x inter
    9.
67 . return scalar bcons = _b[_cons]
    10. return scalar bexpo = _b[expo]
    11. return scalar bx = _b[x]
    12. return scalar binter = _b[inter]
    13. end

68 .
69 . simulate b_cons=r(bcons) b_expo=r(bexpo) b_x=r(bx) b_inter=r(binter), reps(1
    > 000) nodots: model

      command:  model
      b_cons:   r(bcons)
      b_expo:   r(bexpo)
      b_x:      r(bx)
      b_inter:  r(binter)

70 .
71 . program drop _all

72 . sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
b_cons	1,000	.9986508	.0313321	.8839754	1.103671
b_expo	1,000	1.000525	.03217	.9012343	1.089759
b_x	1,000	1.001356	.0322228	.9023424	1.096507
b_inter	1,000	.9975404	.0327809	.8787383	1.100114

```

73 . hist b_expo, xtitle(exposition) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.90123433, width=.00650086)

74 . graph save hist_b_expo, replace
    (file hist_b_expo.gph saved)

```

```

75 . hist b_inter, xtitle(interaction) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.87873834, width=.00763365)

76 . graph save hist_b_inter, replace
    (file hist_b_inter.gph saved)

77 . graph combine hist_b_expo.gph hist_b_inter.gph, title("Paramètres pour expos
    > ition et interaction:" "Modèle de base", color(black)) graphregion(fcolor(wh
    > ite))

78 .
79 . gr export "model.eps", as(eps) preview(off) replace
    (file model.eps written in EPS format)

80 .
81 .
82 . *selection aleatoire
83 . clear all

84 . capture program drop random

85 . program define random, rclass
    1.
86 . drop _all
    2. set obs 1000
    3.
87 . generate expo = rnormal(0,1)
    4. generate x = rnormal(0,1)
    5. generate inter=expo*x
    6. generate error = rnormal(0,1)
    7. generate y = 1 + expo + x + inter + error
    8.
88 . gen toto = runiform()
    9. sort toto
    10. generate random = _n <= 750
    11.
89 . regress y expo x inter if random==1
    12.

```

```

90 . return scalar bcons = _b[_cons]
    13. return scalar bexpo = _b[expo]
    14. return scalar bx = _b[x]
    15. return scalar binter = _b[inter]
    16. end

91 .
92 . simulate b_cons=r(bcons) b_expo=r(bexpo) b_x=r(bx) b_inter=r(binter), reps(1
    > 000) nodots: random

        command: random
        b_cons:  r(bcons)
        b_expo:  r(bexpo)
        b_x:     r(bx)
        b_inter: r(binter)

93 .
94 . program drop _all

95 . sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
b_cons	1,000	.9988118	.0363242	.8671843	1.113319
b_expo	1,000	.9996793	.0371919	.8924965	1.111483
b_x	1,000	.9998527	.0353161	.8851036	1.148267
b_inter	1,000	1.001782	.037343	.8821909	1.108563

```

96 . hist b_expo, xtitle(exposition) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.89249647, width=.00755125)

97 . graph save hist_b_expo, replace
    (file hist_b_expo.gph saved)

98 . hist b_inter, xtitle(interaction) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.88219094, width=.00780593)

```



```

99 . graph save hist_b_inter, replace
    (file hist_b_inter.gph saved)

100 . graph combine hist_b_expo.gph hist_b_inter.gph, title("Paramètres pour expos
    > ition et interaction:" "Sélection aléatoire (MCAR)", color(black)) graphregi
    > on(fcolor(white))

101 .
102 . gr export "random.eps", as(eps) preview(off) replace
    (file random.eps written in EPS format)

103 .
104 .
105 . *selection sur x
106 . clear all

107 . capture program drop selectx

108 . program define selectx, rclass
    1.
109 . drop _all
    2. set obs 1000
    3.
110 . generate expo = rnormal(0,1)
    4. generate x = rnormal(0,1)
    5. generate inter=expo*x
    6. generate error = rnormal(0,1)
    7. generate y = 1 + expo + x + inter + error
    8.
111 . sort x
    9. gen n1=_n
    10. gen selectx=n1 <= 750
    11.
112 . regress y expo x inter if selectx==1
    12.
113 . return scalar bcons = _b[_cons]
    13. return scalar bexpo = _b[expo]
    14. return scalar bx = _b[x]
    15. return scalar binter = _b[inter]
    16. end

```

```

114 .
115 . simulate b_cons=r(bcons) b_expo=r(bexpo) b_x=r(bx) b_inter=r(binter), reps(1
    > 000) nodots: selectx

```

```

    command: selectx
    b_cons:  r(bcons)
    b_expo:  r(bexpo)
    b_x:     r(bx)
    b_inter: r(binter)

```

```

116 .
117 . program drop _all
118 . sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
b_cons	1,000	1.002956	.0427962	.8598431	1.115789
b_expo	1,000	.9985082	.0402189	.8850054	1.132673
b_x	1,000	1.002135	.0504367	.8331351	1.155716
b_inter	1,000	.9988019	.0504382	.8721379	1.1764

```

119 . hist b_expo, xtitle(exposition) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.88500541, width=.00854027)

120 . graph save hist_b_expo, replace
    (file hist_b_expo.gph saved)

121 . hist b_inter, xtitle(interaction) ytitle(densité) graphregion(fcolor(white))
    (bin=29, start=.8721379, width=.01049181)

122 . graph save hist_b_inter, replace
    (file hist_b_inter.gph saved)

123 . graph combine hist_b_expo.gph hist_b_inter.gph, title("Paramètres pour expos
    > ition et interaction:" "Sélection sur x (MAR)", color(black)) graphregion(fc
    > olor(white))

```

```

124 .
125 . gr export "selectx.eps", as(eps) preview(off) replace
    (file selectx.eps written in EPS format)

126 .
127 .
128 . *selection sur y
129 . clear all

130 . capture program drop selecty

131 . program define selecty, rclass
    1.
132 . drop _all
    2. set obs 1000
    3.
133 . generate expo = rnormal(0,1)
    4. generate x = rnormal(0,1)
    5. generate inter=expo*x
    6. generate error = rnormal(0,1)
    7. generate y = 1 + expo + x + inter + error
    8.
134 . sort y
    9. gen n2=_n
    10. gen selecty=n2 <= 750
    11.
135 . regress y expo x inter if selecty==1
    12.
136 . return scalar bcons = _b[_cons]
    13. return scalar bexpo = _b[expo]
    14. return scalar bx = _b[x]
    15. return scalar binter = _b[inter]
    16. end

137 .
138 . simulate b_cons=r(bcons) b_expo=r(bexpo) b_x=r(bx) b_inter=r(binter), reps(1
    > 000) nodots: selecty

        command: selecty
        b_cons:  r(bcons)
        b_expo:  r(bexpo)
        b_x:     r(bx)
        b_inter: r(binter)

```

```
139 .
140 . program drop _all
```

```
141 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
b_cons	1,000	.6501032	.0436375	.4756556	.7768875
b_expo	1,000	.7575017	.0423007	.6335891	.8906894
b_x	1,000	.7607323	.0415359	.5884346	.8960094
b_inter	1,000	.8190863	.0436601	.6781988	.9803056

```
142 . hist b_expo, xtitle(exposition) ytitle(densité) graphregion(fcolor(white))
(bin=29, start=.63358909, width=.00886553)
```

```
143 . graph save hist_b_expo, replace
(file hist_b_expo.gph saved)
```

```
144 . hist b_inter, xtitle(interaction) ytitle(densité) graphregion(fcolor(white))
(bin=29, start=.67819875, width=.01041748)
```

```
145 . graph save hist_b_inter, replace
(file hist_b_inter.gph saved)
```

```
146 . graph combine hist_b_expo.gph hist_b_inter.gph, title("Paramètres pour expos
> ition et interaction:" "Sélection sur y (NMAR)", color(black)) graphregion(f
> color(white))
```

```
147 .
```

```
148 . gr export "selecty.eps", as(eps) preview(off) replace
(file selecty.eps written in EPS format)
```

```
149 .
```

```
150 . log close
      name: <unnamed>
      log: /Users/nicot/Dropbox (CEDIA)/ULaval/Travail/SLIM/Selection/select
> code.smcl
      log type: smcl
      closed on: 11 Aug 2016, 11:17:53
```
