Analyse de données quantitatives longitudinale

Introduction à la régression

Jean-François Bickel
Université de Fribourg

Année académique 2009-2010 (SA09)

Survol

Préambule à SPSS

Introduction

La régression simple

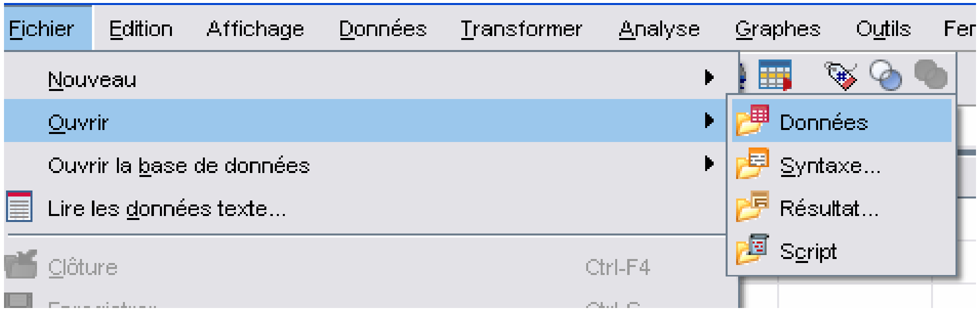
Exemple 1 avec SPSS

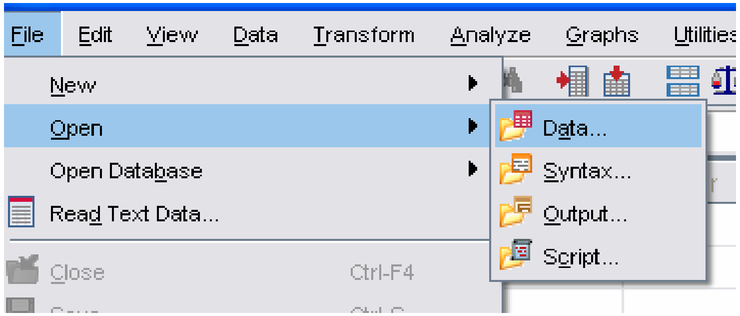
La régression multiple

Les variables indépendantes catégorielles

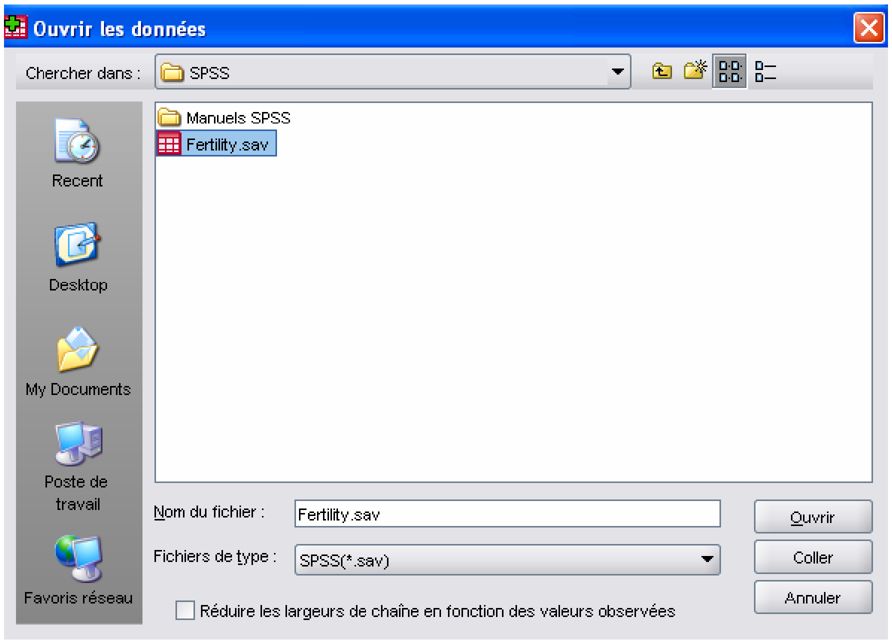
Exemple 2 avec SPSS

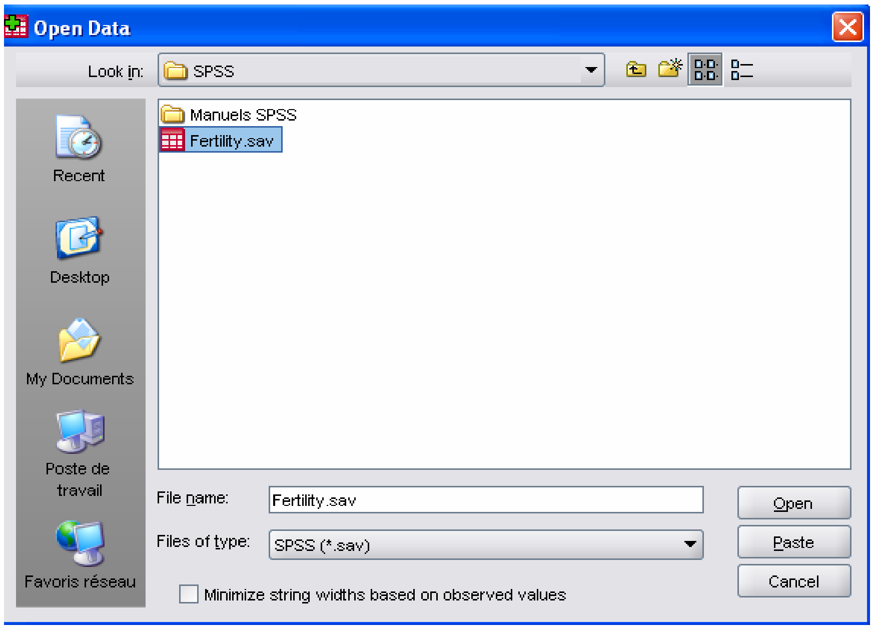
Ouvrir le fichier de données





Ouvrir le fichier de données





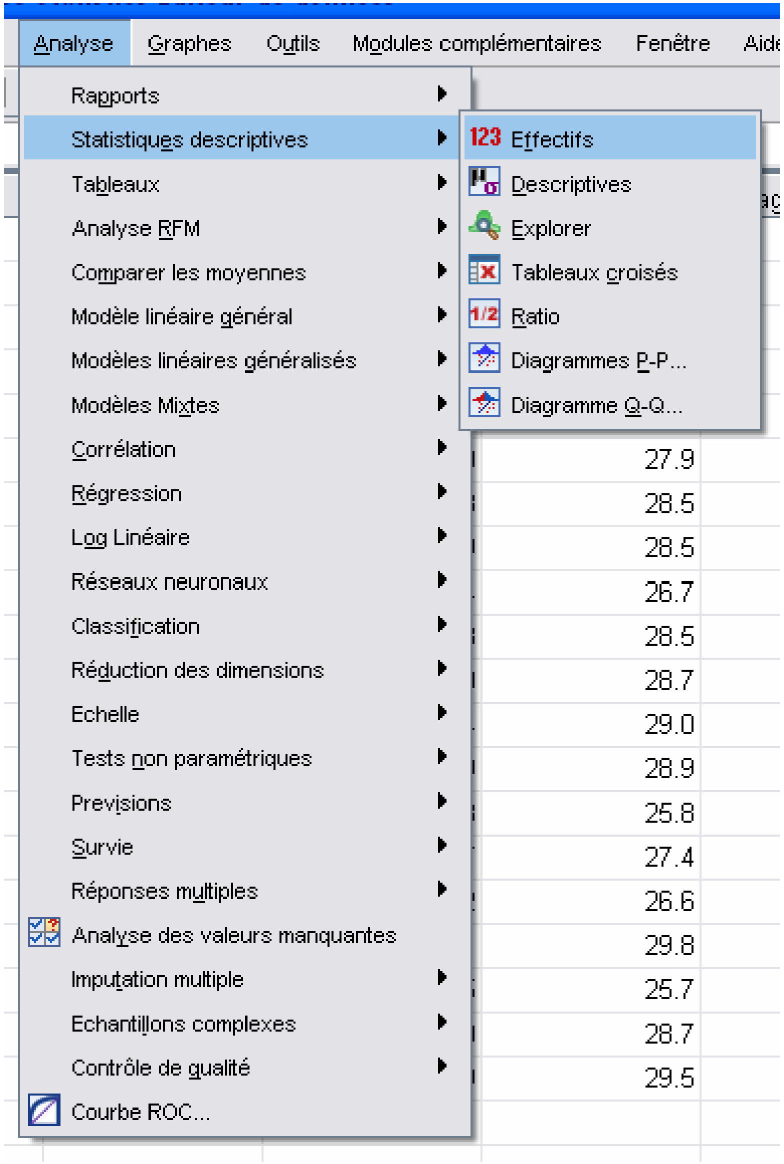
Le fichier de données SPSS

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Modules complémentaires Fenêtre Aide

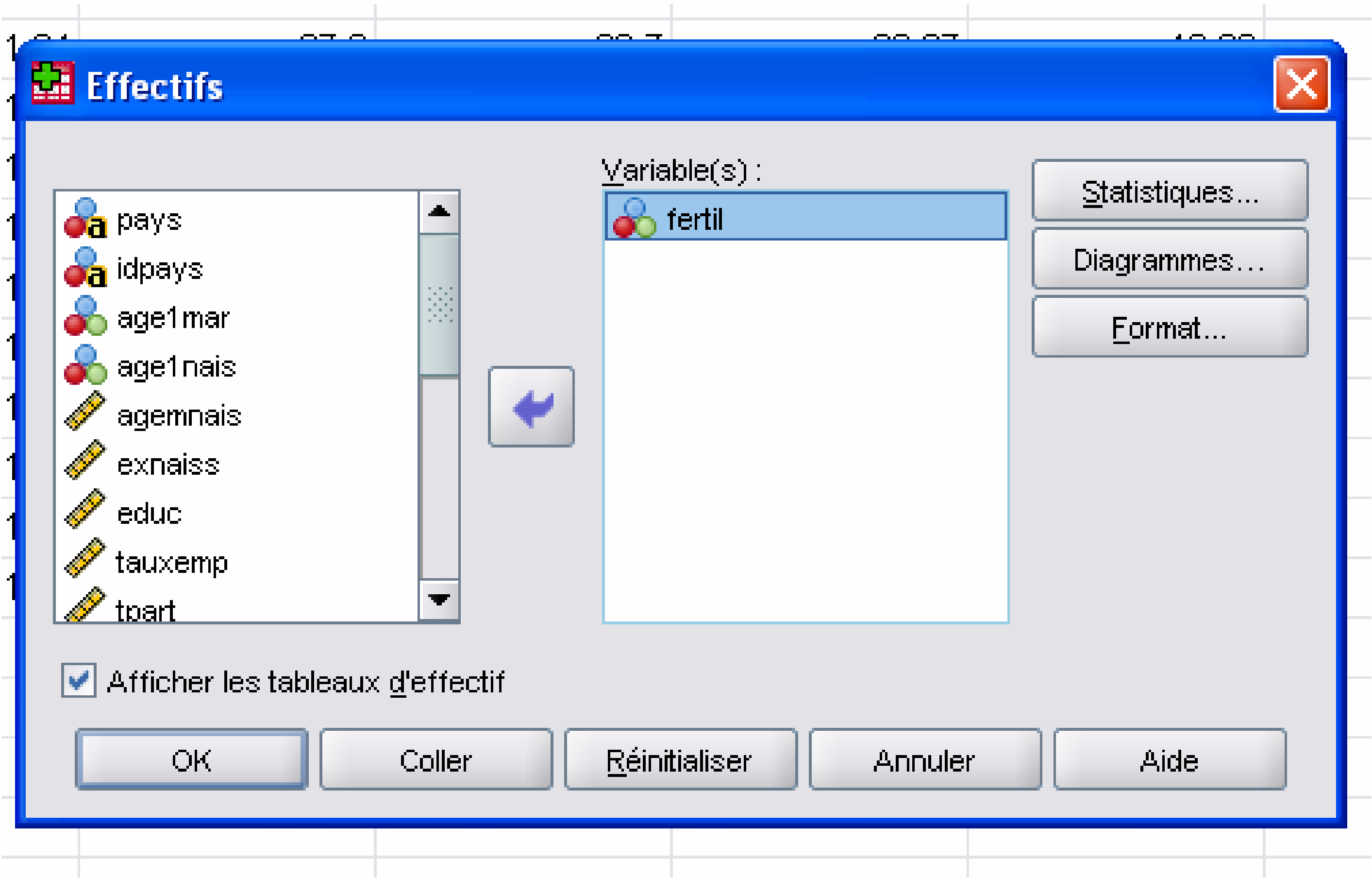
1 : pays Allemagne

	pays	idpays	fertil	age1mar	age1nais	agemnais
1	Allemagne	DE	1.32	27.6	29.1	29.56
2	Autriche	AT	1.40	27.0	27.2	29.17
3	Belgique	BE	1.72	26.7	27.4	.
4	Danemark	DK	1.83	29.7	28.4	30.29
5	Espagne	ES	1.38	27.9	29.3	30.88
6	Finlande	FI	1.84	28.0	27.9	29.96
7	France	FR	2.00	27.8	28.5	29.72
8	Grèce	GR	1.39	26.9	28.5	29.87
9	Hongrie	HU	1.34	25.4	26.7	28.70
10	Irlande	IE	1.93	27.3	28.5	30.66
11	Italie	IT	1.34	27.0	28.7	30.87
12	Luxembourg	LU	1.65	27.4	29.0	29.94
13	Pays-Bas	NL	1.70	27.9	28.9	30.58
14	Pologne	PL	1.27	24.3	25.8	28.34
15	Portugal	PT	1.35	25.7	27.4	29.48
16	Rép. tchèque	CZ	1.33	25.2	26.6	28.91
17	Royaume-Uni	UK	1.84	27.1	29.8	29.17
18	Slovaquie	SK	1.24	24.5	25.7	27.94
19	Suède	SE	1.85	30.0	28.7	30.53
20	Suisse	CH	1.43	27.9	29.5	30.64

Statistique univariée



Statistique univariée



Statistique univariée

Statistique univariée

Statistiques

fertil

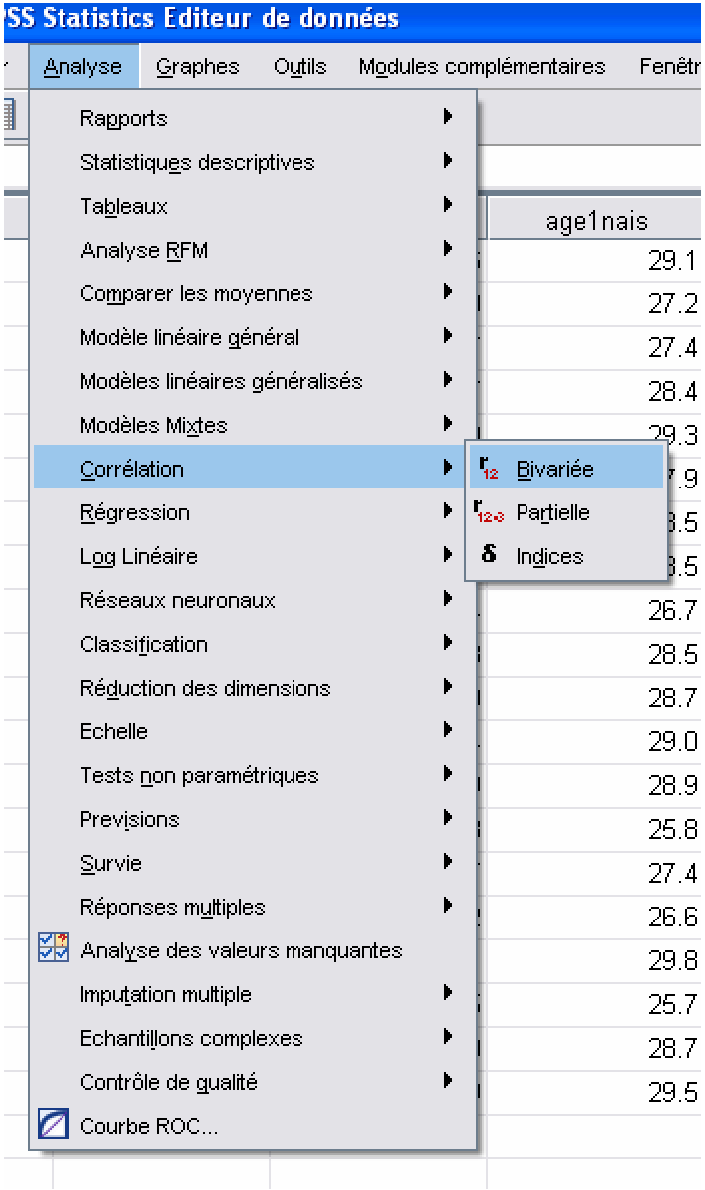
N	Valide	20
	Manquante	0
Moyenne		1.5575
Médiane		1.4150
Ecart-type		.25538
Minimum		1.24
Maximum		2.00
Centiles	25	1.3400
	50	1.4150
	75	1.8375

Statistique univariée

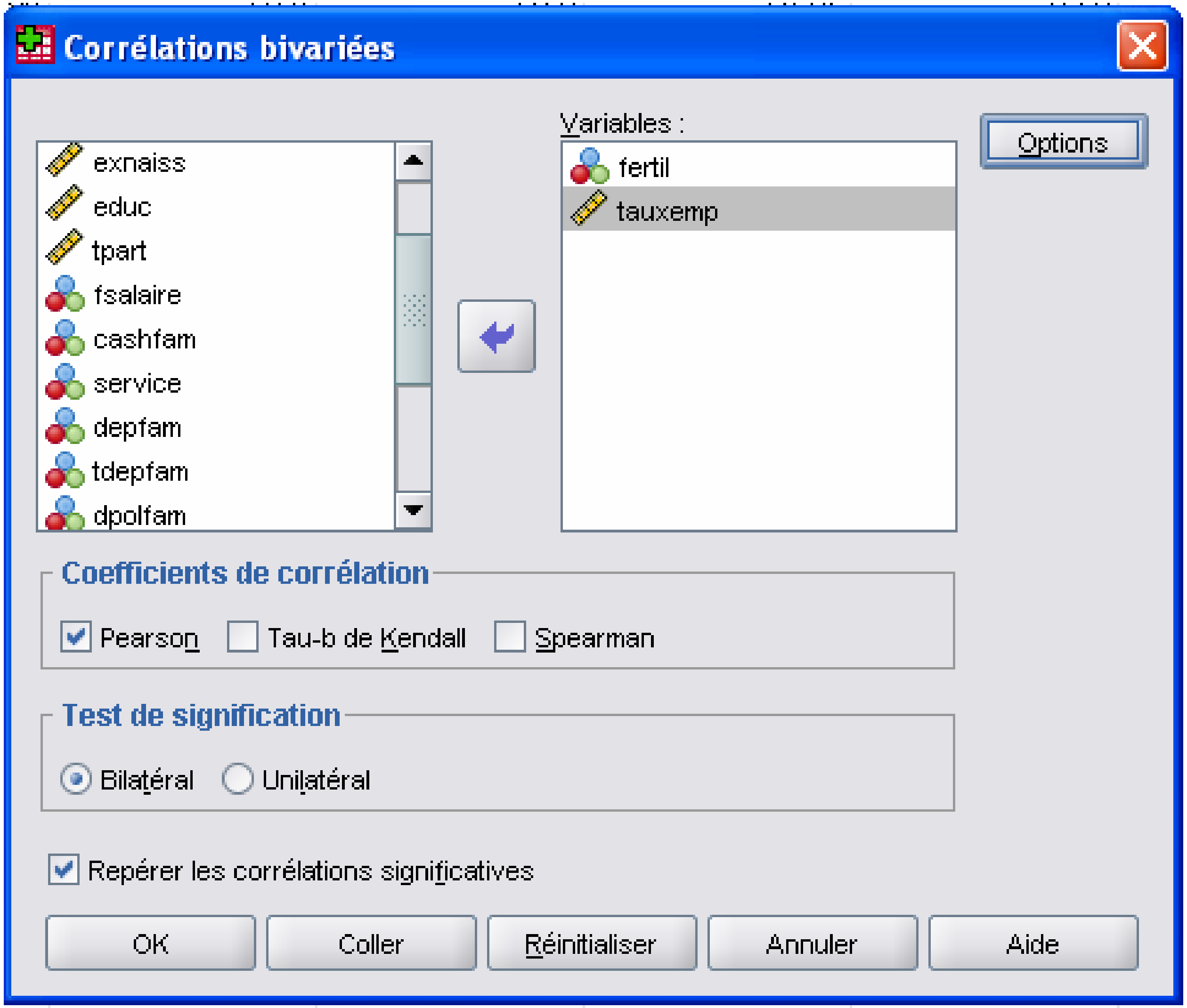
fertil

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	1.24	1	5.0	5.0	5.0
	1.27	1	5.0	5.0	10.0
	1.32	1	5.0	5.0	15.0
	1.33	1	5.0	5.0	20.0
	1.34	2	10.0	10.0	30.0
	1.35	1	5.0	5.0	35.0
	1.38	1	5.0	5.0	40.0
	1.39	1	5.0	5.0	45.0
	1.40	1	5.0	5.0	50.0
	1.43	1	5.0	5.0	55.0
	1.65	1	5.0	5.0	60.0
	1.70	1	5.0	5.0	65.0
	1.72	1	5.0	5.0	70.0
	1.83	1	5.0	5.0	75.0
	1.84	2	10.0	10.0	85.0
	1.85	1	5.0	5.0	90.0
	1.93	1	5.0	5.0	95.0
	2.00	1	5.0	5.0	100.0
	Total	20	100.0	100.0	

Analyse bivariée - corrélation



Analyse bivariée - corrélation



Analyse bivariée - corrélation

Corrélations

		fertil	tauxemp
fertil	Corrélation de Pearson	1	.526*
	Sig. (bilatérale)		.017
	N	20	20
tauxemp	Corrélation de Pearson	.526*	1
	Sig. (bilatérale)	.017	
	N	20	20

*. La corrélation est significative au niveau 0.05 (bilatéral).

Introduction à la régression

La régression est une méthode statistique qui traite les valeurs d'une variable dépendante y comme étant fonction des valeurs prises par une ou plusieurs variables indépendantes x

$$y = f(x_k)$$

pour k=1,2,3… variables indépendantes

Introduction à la régression

De manière équivalente:

La distribution de y est conditionnelle: elle dépend des valeurs prises par x_1, x_2... x_k

Introduction à la régression

Dans le cas le plus général et le plus simple, la fonction est traitée comme étant *linéaire*

La relation entre les « x » et y peut être représentée par une équation ayant la forme

$y = a + b_1x_1 + b_2x_2 + ... + b_kx_k$

Introduction à la régression

Pour n'importe quelle combinaison de valeurs des variables x_2, x_3, …, x_k, la relation entre la variable x_1 et la variable y peut être représentée par une ligne droite

Et la même chose pour la relation entre x_2 et y, x_3 et y, …, x_k et y

La régression simple

La régression simple (=bivariée) comprend une seule variable indépendante

L’équation de régression a donc la forme:

$y = a + bx$

La régression simple

Soit la question de la relation entre fécondité et taux d'emploi

Elle peut être exprimée sous la forme d'une équation de régression

$$y_{(\text{fécondité})} = a + b_{(\text{taux d'emploi})}$$

La régression simple

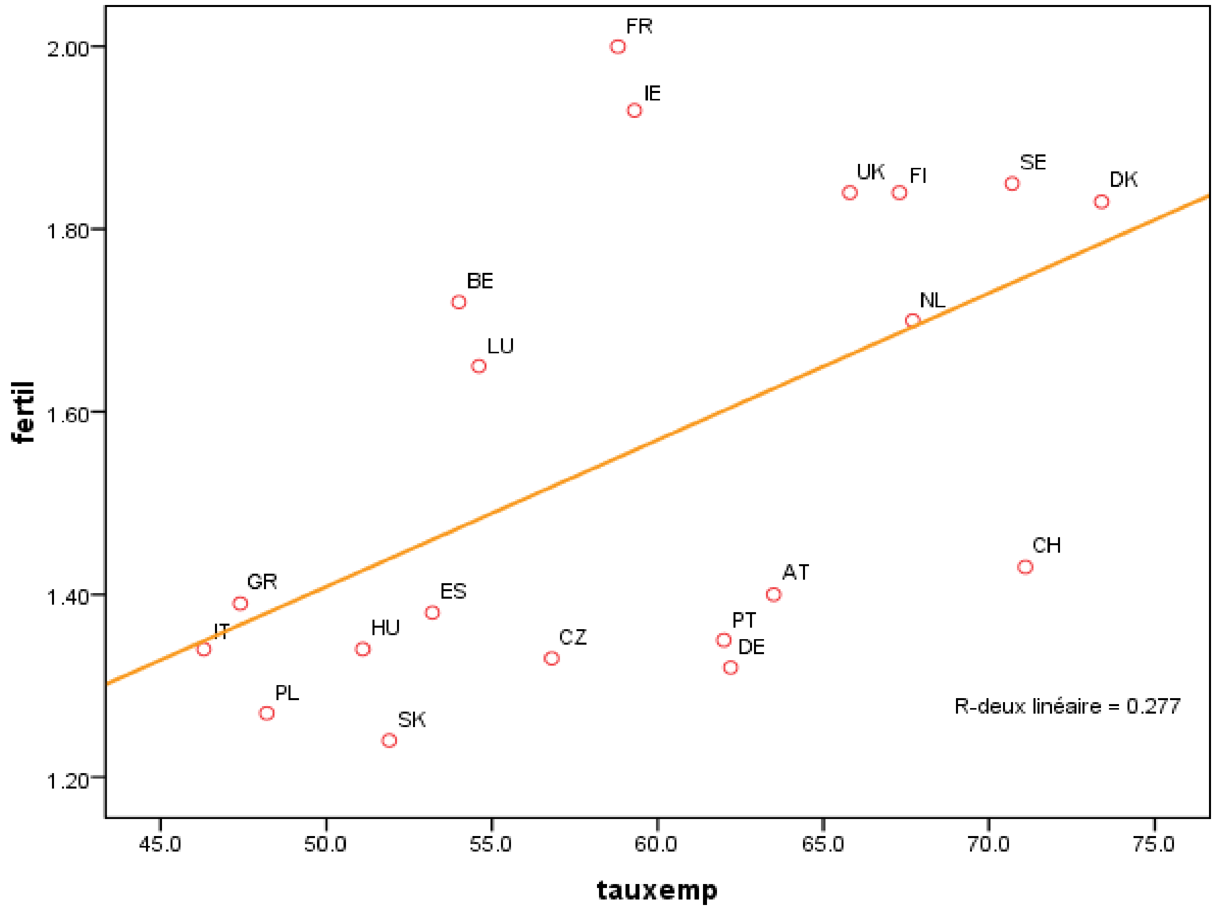
Cette équation a donc deux coefficients

a qui est une constante et qui s'interprète comme étant la valeur de y lorsque $x = 0$

b qui exprime la relation entre y et x et qui s'interprète comme étant le changement moyen de y lorsque x augmente d'une unité

- La variable x est de type quantitatif

La régression simple



La régression simple

L’objectif est de déterminer les coefficients de l’équation pour la droite de régression qui permette la meilleure approximation possible des données observées

Méthode ordinaire des moindres carrés

Objectif: minimisation des erreurs de prédiction (voir ci-dessous)

La régression simple

Droite de régression:

$y = a + bx$

avec pente

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

et constante

$$a = \bar{y} - b\,\bar{x}$$

La régression simple

Où:

r est la corrélation entre y et x

s_y est l'écart type de y

s_x est l'écart type de x

$\bar{y}$ et $\bar{x}$ sont les moyennes des deux variables

La régression simple

Dans la réalité, les valeurs prédites de *y* (=$\hat{y}$) à l'aide de l'équation de régression sont différentes des valeurs observées pour *y*

On parle à ce propos d'*erreur* de prédiction

$\text{Erreur}_i = \text{valeur observée}_i - \text{valeur prédite}_i$

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

- *i* réfère à la i$^{\text{ème}}$ observation (cas, individu)

La régression simple

En cela, les deux équations ci-dessous sont équivalentes

$\hat{y} = a + bx$

$y = a + bx + erreur$

La régression simple

Sur cette base, on peut calculer la somme des erreurs de prédictions pour l'ensemble des observations

Les erreurs sont au préalable élevées au carré pour éviter qu'elles s'annulent

La somme des erreurs équivaut donc à:

$$\Sigma (y_i - \hat{y}_i)^2$$

La régression simple

Cette somme représente la variation de y autour de la droite de régression

On qualifie cette variation de *résiduelle*, car c’est la part de la variation de y qui « reste » après qu’on ait pris en compte celle qui dépend des valeurs prises par x

La régression simple

On parle donc de la « somme des carrés des erreurs résiduelles » (en abrégé RSS) ou plus simplement de « résidus »

$$RSS = \Sigma \text{ résidus}_i^2 = \Sigma\, (y_i - \hat{y}_i)^2$$

La régression simple

La variation totale de *y* autour de sa moyenne, c'est-à-dire en ignorant la droite de régression, est également calculée comme étant une somme des carrés des erreurs (en abrégé TSS)

Soit:

TSS = $\Sigma (y_i - \bar{y})^2$

La régression simple

La différence entre les deux mesures de variation est la quantité de variation qui est expliquée par la régression de *y* sur *x*

variation expliquée	=	variation totale	-	variation résiduelle

La régression simple

En abrégé:

RegSS = TSS – RSS

$$= \Sigma\, (y_i - \bar{y})^2 - \Sigma\, (y_i - \hat{y}_i)^2$$

$$= \Sigma\, (\hat{y}_i - \bar{y})^2$$

La régression simple

La corrélation élevée au carré (r^2) exprime la variation expliquée comme une fraction de la variation totale de *y*

$$r^2 = \frac{\text{variation expliquée}}{\text{variation totale}} = \frac{\text{TSS - RSS}}{\text{TSS}}$$

La régression simple

On parle ainsi de r^2 comme d'une mesure de « réduction proportionnelle de l'erreur »

On l'utilise comme indicateur de la qualité globale (« *fit* ») de l'équation de régression, c'est-à-dire du degré auquel elle permet de prédire y

La régression simple

Une autre mesure d'intérêt est celle de l'écart type des résidus (des erreurs de prédictions)

$$s_r = \sqrt{\Sigma\, résidus_i^2 / (n-2)} = \sqrt{\Sigma\, (y_i - \hat{y}_i)^2 / (n-2)}$$

La régression simple

Mesure la proximité avec laquelle la droite de régression « colle » aux points observés

Plus précisément, s_r est la distance moyenne des observations par rapport à la droite de régression

Autrement dit, « l'erreur typique » commise par l'équation de régression en prédisant *y* à partir de *x*

La régression simple

De plus, si les résidus sont approximativement distribués selon une courbe normale, 2/3 de ces résidus sont dans un éventail de $\pm$ 2 écarts types (autour de $\hat{y}$) et 95% d’entre eux sont dans un éventail de $\pm$ 4 écarts types

La régression simple

Nota Bene:

Dans SPSS, on parle de « erreur standard de l'estimation »

C'est dommage car le terme « erreur standard » est normalement utilisé pour parler de la distribution d'une statistique particulière

Exemple 1 avec SPSS

Exemple

Analyse de l’indice de fertilité comme étant fonction du taux d’emploi des femmes

- •Variable pour indice de fertilité: <fertil>
- •Variable pour taux d’emploi: <tauxemp

L’équation de régression a donc la forme

$\hat{y}_{\text{(fertil)}} = \quad a + b_{1}\text{(tauxemp)}$

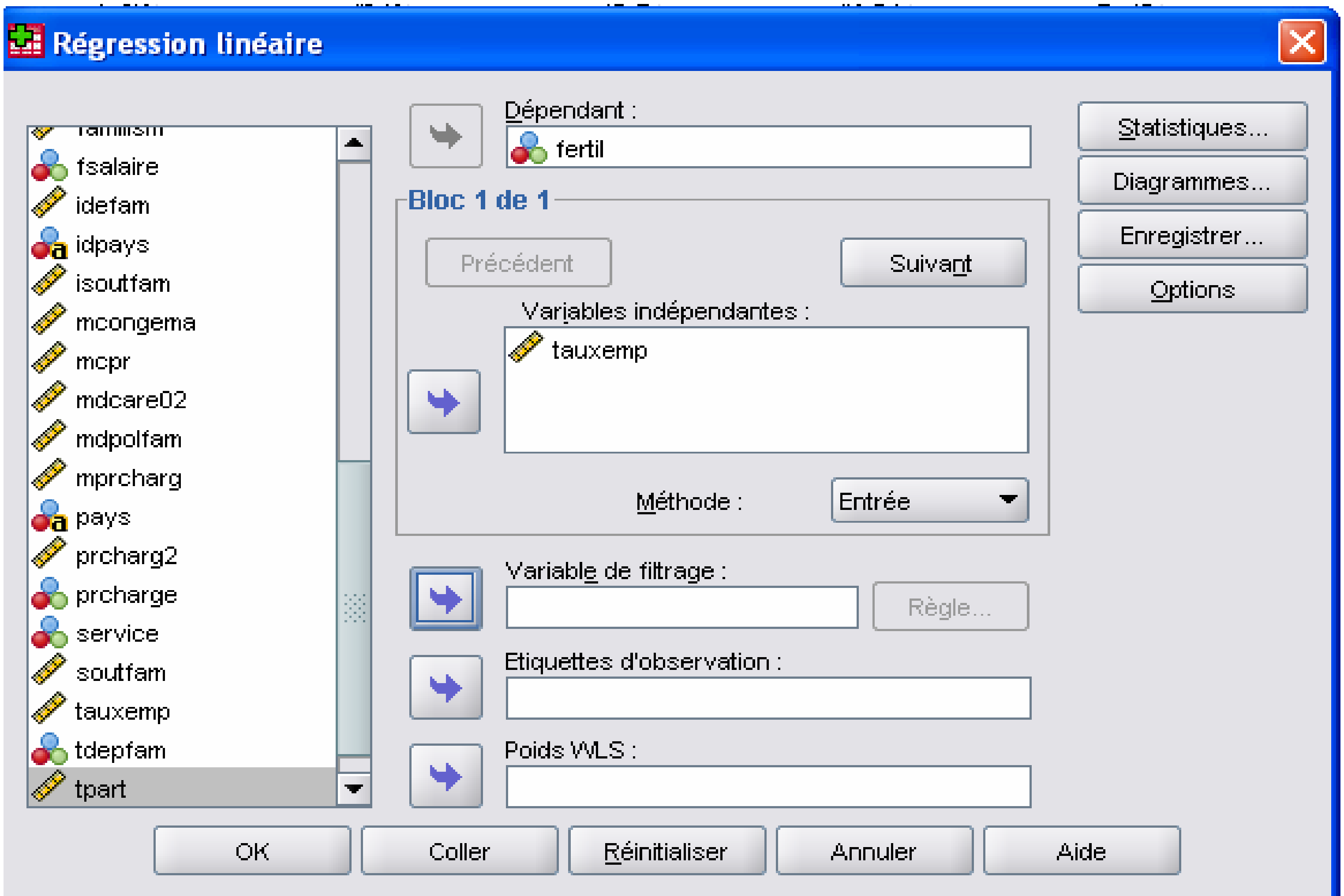
Exemple 1 avec SPSS

Instructions SPSS

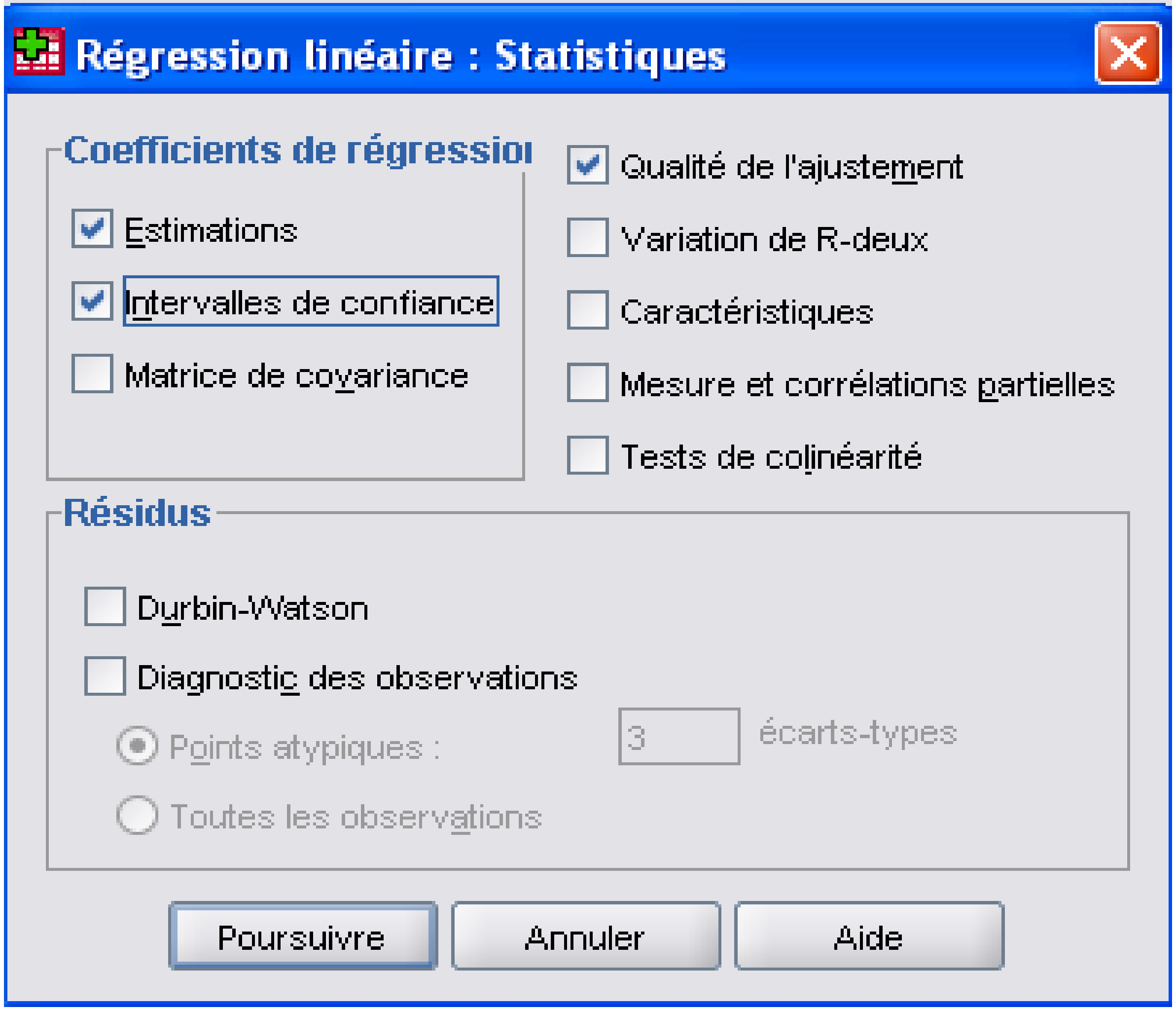
Par l'interface graphique

Menu *Analyse* -> *Régression* -> *Linéaire*

Exemple 1 avec SPSS



Exemple 1 avec SPSS



Exemple 1 avec SPSS

Par la syntaxe

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT fertil
  /METHOD=ENTER tauxemp.
```

Exemple 1 avec SPSS

Syntaxe simplifiée

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS DEFAULTS CI
  /DEPENDENT fertil
  /METHOD=ENTER tauxemp.
```

Exemple 1 avec SPSS

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.526[a]	.277	.236	.22316

a. Valeurs prédites : (constantes), tauxemp

Exemple 1 avec SPSS

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	.343	1	.343	6.883	.017[a]
	Résidu	.896	18	.050		
	Total	1.239	19			

a. Valeurs prédites : (constantes), tauxemp

b. Variable dépendante : fertil

Exemple 1 avec SPSS

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	.605	.366		1.652	.116	-.164	1.375
	tauxemp	.016	.006	.526	2.624	.017	.003	.029

a. Variable dépendante : fertil

La régression multiple

La régression multiple est une *extension* de la régression simple

Le réel pouvoir heuristique de l'analyse de régression réside en effet dans sa capacité à prendre en compte de manière simultanée plusieurs variables indépendantes

La régression multiple

D’une part, cela permet de réduire la grandeur des résidus et donc de rendre compte d’une plus grande part de la variation de la variable dépendante y

Objectif: prédire y de manière plus précise

La régression multiple

D’autre part et surtout, il est possible d’évaluer plus correctement « l’effet » de *x* sur *y*

- Si *x* est lié à une autre variable indépendante *z*, qui exerce elle-même un effet sur *y*, la prise en compte de *z* permet (a) soit de ne pas attribuer faussement à *x* un effet qui est dû en réalité à *z*; (b) soit de spécifier l’effet de *x* sur *y*, cet effet « passant » par *z*

La régression multiple

On étend donc ce que l'on vient de voir dans le cas de la régression simple à *k* variables indépendantes x_1, x_2, ..., x_k

L'équation de régression linéaire prend alors la forme

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \textit{résidus}$$

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

La régression multiple

La procédure est identique et vise à déterminer les valeurs de a, b_1, b_2, ..., b_k qui minimise la sommes des résidus (Σrésidus^2)

Les coefficients de cette équation ont l'interprétation suivante

a est la valeur prédite de y quand les variables x_1, x_2, ..., x_k valent toutes 0

La régression multiple

b_1 est la changement moyen de y qui est associé à l'accroissement d'une unité de x_1, toutes les autres variables indépendantes (càd., $x_2, ..., x_k$) étant maintenues constantes

$b_2, ..., b_k$ s'interprètent de la même manière

- Cette interprétation pour x_1 (respectivement $x_2, ..., x_k$) vaut quand x_1 (ou $x_2, ..., x_k$) est une variable de type quantitatif

Variables indépendantes catégorielles

Des variables catégorielles peuvent également être introduites dans une équation de régression au titre de variables indépendantes

Ce cours traite de la manière de faire la plus courante

La *procédure* comprend *2 étapes* successives

Variables indépendantes catégorielles

Etape 1:

On transforme *chaque catégorie*, *sauf une*, de la variable en une nouvelle variable « *dummy* » (binaire) codée

1 pour toutes les observations appartenant à la catégorie en question

0 pour toutes les autres observations

Variables indépendantes catégorielles

En procédant ainsi pour toutes les *k* catégories de la variable *sauf une*, on obtient *k-1* variables dummies

La catégorie restante est considérée comme *catégorie de référence*

Variables indépendantes catégorielles

Exemple: types de familialisme

Selon la construction adoptée, la variable comprend *k=4* catégories

familialisme optionnel / ... explicite / défamilialisme / familialisme implicite

Elle est recodée en *k-1=3* catégories

Variables indépendantes catégorielles

On aura donc 3 variable dummies telles que, par exemple

Dummy 1: *1*= familialisme explicite, *0*=autres types de familialisme

Dummy 2: *1*=défamilialisme, *0*=autres types de familialisme

Variables indépendantes catégorielles

Dummy 3: *1*=familialisme implicite, *0*=autres types de familialisme

Dans cet exemple, la *catégorie de référence* est donc *familialisme optionnel*

- N.B. Les observations appartenant à la catégorie de référence sont codées *0* sur toutes les variables dummies

Variables indépendantes catégorielles

Le choix de la catégorie de référence est arbitraire

Un choix fréquent est celui de la catégorie majoritaire

- Permet la comparaison de situations « minoritaires » à celle majoritaire ou « normale »

Un intérêt substantiel peut motiver le choix

Variables indépendantes catégorielles

Règle pratique 1

Créez *vous-mêmes* les variables dummies

Puis introduisez-les *vous-mêmes* dans l’équation de régression (selon la procédure expliquée ci-dessous)

Variables indépendantes catégorielles

Règle pratique 2:

Donnez des noms aux variables dummies tels qu'ils désignent de manière synthétique la catégorie codée *1*

- Permet de se souvenir aisément de la catégorie cible et du sens de la comparaison avec la catégorie servant de référence

Variables indépendantes catégorielles

Exemples:

Nommer *explicit* la variable dummy codée *1* pour le familialisme explicite et *0* pour les autres types de familialisme

Et ainsi de suite

Variables indépendantes catégorielles

Etape 2:

Introduire dans l'équation de régression les *k-1* variables dummies « exprimant » la variable originelle

- Les observations appartenant à la catégorie de référence sont aussi dans l'équation: ce sont celles qui sont codées *0* sur toutes les dummies

Variables indépendantes catégorielles

En reprenant notre exemple:

Introduire *simultanément* les trois variables dummies « exprimant » le type de familialisme

Variables indépendantes catégorielles

Le coefficient *b* de la variable dummy exprime la différence entre la valeur moyenne de *y* pour la catégorie cible et la valeur moyenne de *y* pour la catégorie de référence

Mode de lecture: *y* pour catégorie cible *par rapport à y* pour catégorie de référence

Exemple 2 avec SPSS

Exemple

Analyse de l’indice de fertilité comme étant fonction du taux d’emploi des femmes et du type de familialisme

- Variable pour indice de fertilité: <fertil>
- Variable pour taux d’emploi: <tauxemp>
- Variable pour types de familialisme: <typfam>

Exemple 2 avec SPSS

Etape 1: création des variables dummies

Mot-clé: *recode* (« *recoder* »)

```
recode typfam (1=1) (2,3,4=0) into optional.
recode typfam (2=1) (1,3,4=0) into explicit.
recode typfam (3=1) (1,2,4=0) into defami.
recode typfam (4=1) (1,2,3=0) into implicit.
exe.
```

Exemple 2 avec SPSS

Etape 2: Analyse de régression proprement dite

Une procédure pas à pas est adoptée

Exemple 2 avec SPSS

1) Une première équation de régression ne contient que le taux d'emploi des femmes comme variable indépendante

2) Puis une seconde équation de régression contient le taux d'emploi des femmes et le type de familialisme comme variables indépendantes

Exemple 2 avec SPSS

L'équation de régression *1* a donc la forme

$\hat{y}_{\text{(fertil)}} = a + b_{1\text{(tauxemp)}}$

L'équation de régression *2* a donc la forme

$\hat{y}_{\text{(fertil)}} = a + b_{1\text{(tauxemp)}} + b_{2\text{(explicit)}} + b_{3\text{(defami)}} + b_{4\text{(implicit)}}$

Exemple 2 avec SPSS

Ces deux équations de régression successives sont enchâssées (*nested*) l'une dans l'autre:

a) Elles portent sur les *mêmes observations*

b) La seconde équation est plus complète que la première (qualifiée de *réduite*) car elle utilise davantage d'information (le type de familialisme

Exemple 2 avec SPSS

Ces deux équations peuvent être calculées successivement par SPSS dans une seule et même analyse

Ceci permet de demander à SPSS de comparer entre elles les deux équations et de déterminer si celle qui est plus complète permet effectivement de mieux prédire y

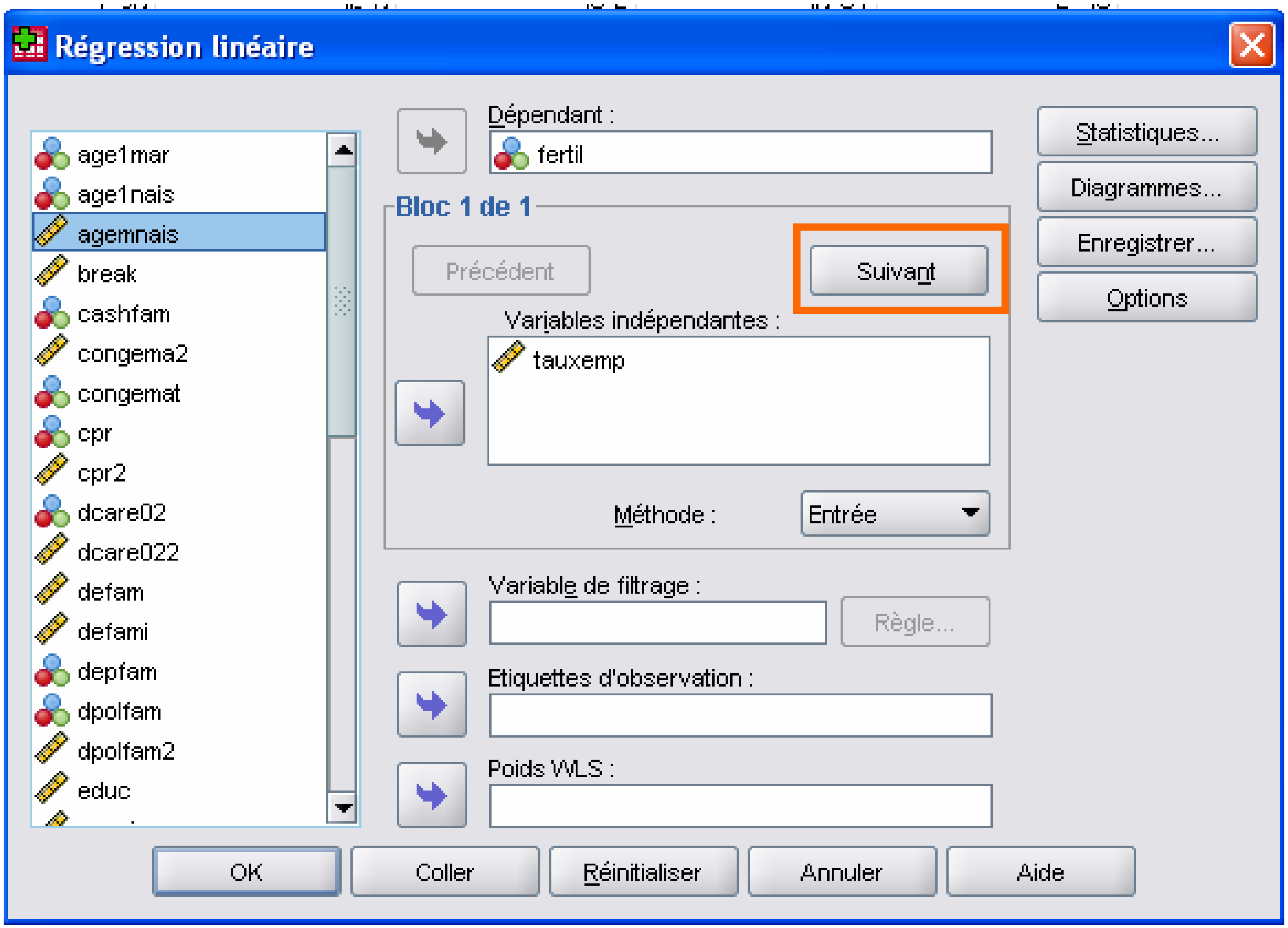
Exemple 2 avec SPSS

Instructions SPSS

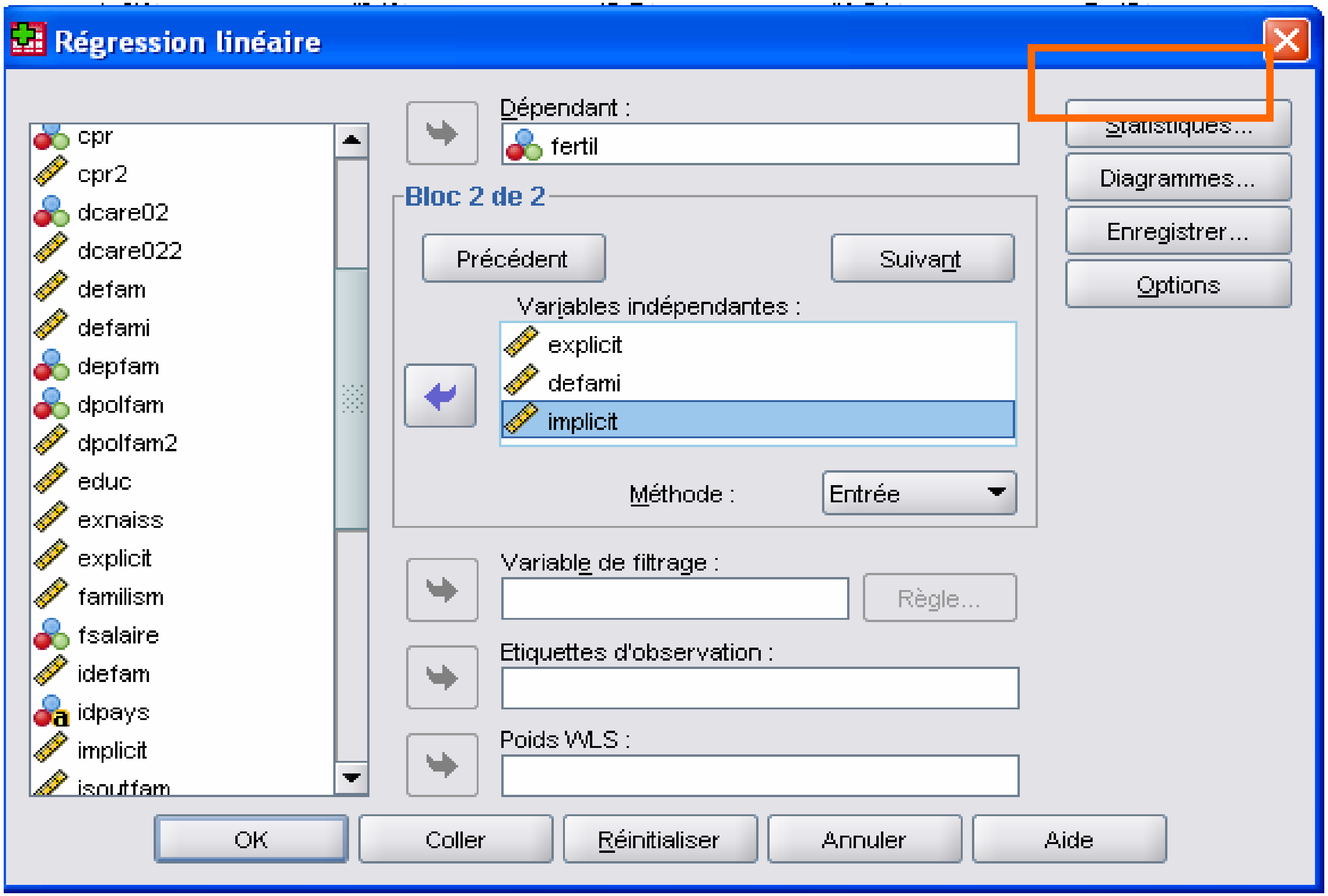
Par l’interface graphique

Menu *Analyse* -> *Régression* -> *Linéaire*

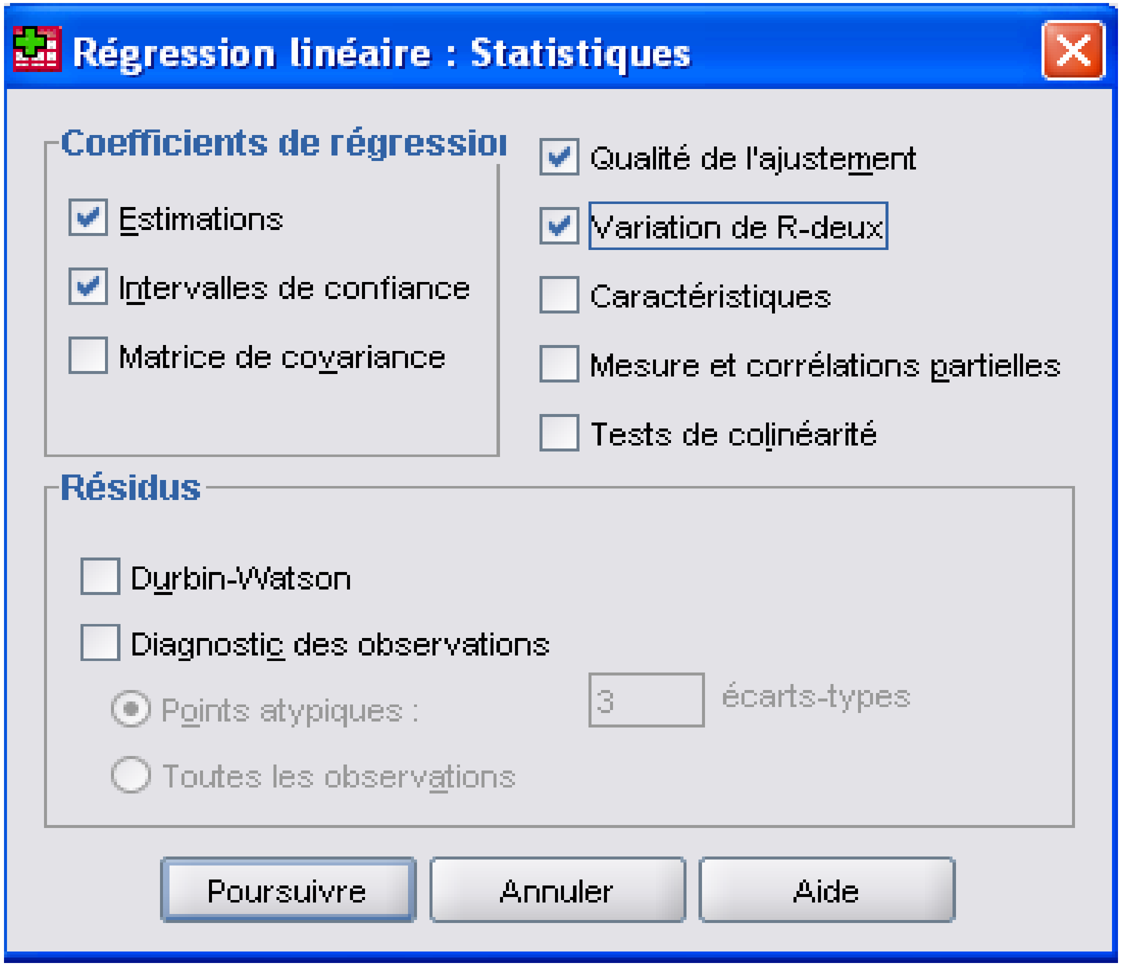
Exemple 2 avec SPSS



Exemple 2 avec SPSS



Exemple 2 avec SPSS



Exemple 2 avec SPSS

Par la syntaxe

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT fertil
  /METHOD=ENTER tauxemp
  /METHOD=ENTER explicit defami implicit.
```

Exemple 2 avec SPSS

Syntaxe simplifiée

```
REGRESSION
  /STATISTICS DEFAULTS CI CHANGE
  /DEPENDENT fertil
  /METHOD=ENTER tauxemp
  /METHOD=ENTER explicit defami implicit.
```

Exemple 2 avec SPSS

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.526[a]	.277	.236	.22316
2	.756[b]	.572	.458	.18806

Modèle	Changement dans les statistiques				
	Variation de R-deux	Variation de F	ddl 1	ddl 2	Modification de F signification
1	.277	6.883	1	18	.017
2	.295	3.448	3	15	.044

Exemple 2 avec SPSS

ANOVA[c]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	.343	1	.343	6.883	.017[a]
	Résidu	.896	18	.050		
	Total	1.239	19			
2	Régression	.709	4	.177	5.009	.009[b]
	Résidu	.531	15	.035		
	Total	1.239	19			

a. Valeurs prédites : (constantes), tauxemp

b. Valeurs prédites : (constantes), tauxemp, defami, explicit, implicit

c. Variable dépendante : fertil

Exemple 2 avec SPSS

Coefficients[a]

Modèle		Coefficients non standardisés B	Coefficients non standardisés Erreur standard	Coefficients standardisés Bêta
1	(constante)	.605	.366	
	tauxemp	.016	.006	.526
2	(constante)	1.479	.434	
	tauxemp	.006	.006	.194
	explicit	-.480	.151	-.772
	defami	-.240	.134	-.418
	implicit	-.335	.141	-.641

Modèle	t	Signification	Intervalle de confiance à 95% de B Borne inférieure	Intervalle de confiance à 95% de B Borne supérieure
1	1.652	.116	-.164	1.375
	2.624	.017	.003	.029
2	3.411	.004	.555	2.403
	.948	.358	-.007	.019
	-3.189	.006	-.802	-.159
	-1.799	.092	-.525	.044
	-2.382	.031	-.634	-.035