UN GROUPE D'INDIVIDUS DÉCRITS PAR PLUSIEURS VARIABLES QUANTITATIVES

PRÉDIRE UNE DES VARIABLES À PARTIR DES AUTRES

Mots-clés : Régression ; Corrélation semi-partielle ; corrélation partielle ; Variable prédictrice ; Variable prédite ; Rapport de corrélation ; R^2 ; Multicolinéarité ; Nuage bivarié ;

Ce document a été réalisé avec la version 7 (version bêta) de SES-Pegase.

On pourra se reporter à l'analyse du dossier CLIMAT pour un exemple de prédiction d'une variable à partir d'une seule autre variable.

TYPE DES DONNÉES ANALYSÉES

Nous présenterons l'analyse d'un dossier particulier, le dossier CLIMAT. Mais cette analyse s'applique à toutes données de la forme suivante (cf. Tableau 1) :

Tableau 1 : Structure du tableau des données individuelles

	X1	X2	X3	X4	(. . .)
i1					
i2					
i3					
i4					
i5					
i6					
(. . .)					

On a un ensemble d'*individus* (i1, i2…). Ces individus peuvent être des personnes, des pays, des enfants, des animaux, des voitures, des familles…
Sur ces individus, ont été recueillies plusieurs variables quantitatives (X1, X2…).
Ces variables peuvent être des notes, des temps, des poids, des QI… Elles ne sont pas nécessairement dans la même unité. C'est-à-dire que l'on peut avoir, parmi ces variables, à la fois une note, un poids, un temps en seconde et un temps en heures, etc.

QUESTIONS

On se demande s'il est possible de prédire les valeurs d'une des variables (variable à prédire ou VD) en fonction des valeurs des autres variables qui ont alors le statut de variables prédictrices (VP) ou variables indépendantes (VI).

La méthode appropriée est la régression multiple. Il existe de nombreuses variantes de cette méthode. Nous présentons ici la méthode la plus classique, la régression linéaire multiple. Elle consiste à rechercher une relation de type linéaire entre la VD et les variables prédictrices. A partir de l'équation de régression obtenue on peut répondre aux questions suivantes :

1. Peut-on trouver une équation simple permettant de prédire, au moins en partie, les valeurs de la VD connaissant les valeurs des autres variables ?
2. Quelle est la qualité globale de la prédiction ?
3. Quel est le poids de chaque prédictrice dans la prédiction de la VD ?
4. Quelle prédictrice a le poids le plus important ?
5. Quelle valeur de la VD prédire pour un individu dont on ne connait que les valeurs observées sur les prédictrices ?

UN EXEMPLE : LE DOSSIER FAMILY[1]

Présentation des données

Lors d'une étude réalisée aux USA, 35 familles ayant une fille aînée (ou fille unique) en "ninth grade" (= classe de Troisième) ont été choisies au hasard.
- Le père a répondu à un questionnaire sur ses intérêts pour le sport (échelle numérique de 0 à 50) : PERE.
- La mère a répondu au même questionnaire : MERE.
- Le professeur d'éducation physique de chacune des filles a noté (de 0 à 20) ses performances sportives : NOTE.
- La fille a également répondu au même questionnaire que son père et sa mère : FILLE.

Tableau 2: Les données FAMILY (partie)

INDIVIDUS	PERE	MERE	NOTE	FILLE
f01	23	25	8	24
f02	32	30	13	30
f03	25	25	15	25
f04	36	31	10	35
f05	37	36	9	39
f06	31	33	10	30
f07	28	22	8	27
f08	26	18	5	14
(...)				
(...)				
f28	23	26	5	11
f29	29	24	8	27
f30	38	30	9	24
f31	41	20	15	40
f32	39	31	7	32
f33	27	29	6	10
f34	48	43	14	37
f35	33	30	7	19

Questions

On se demande dans quelle mesure l'intérêt du père (PÈRE) pour le sport, l'intérêt de la mère pour le sport (MERE), et la note de la fille en sport (NOTE) permettent de prédire l'intérêt de la fille pour le sport (FILLE).

1. Quelle est la qualité globale de la prédiction de l'intérêt de la fille pour le sport (FILLE) à partir des trois autres variables (PÈRE, MERE, NOTE) ?
2. Dans la prédiction de l'intérêt de la fille pour le sport (FILLE) quel est le poids, de chacune des autres variables (PÈRE, MERE, NOTE) ?
3. Dans la prédiction de l'intérêt de la fille pour le sport (FILLE), laquelle des autres variables (PÈRE, MERE, NOTE) a le poids le plus important ?
4. Quel intérêt pour le sport peut-on prédire pour une fille dont on ne connait que les valeurs observées sur les trois autres variables (PÈRE, MERE, NOTE) ?

Type et statut des variables

Les *individus* sont en fait ici des familles (comprenant chacune plusieurs personnes).
Les quatre variables sont numériques mais ne sont pas toutes sur la même échelle (échelle de 0 à 50 pour trois d'entre elles et note sur 20 pour l'autre).

La variable FILLE est la variable à prédire ou variable dépendante (VD).
Les trois autres variables (PÈRE, MERE, NOTE) ont le statut de variables prédictrices (VP) ou variables indépendantes (VI).

[1] Source: Hays, W.L. (1994) - Statistics, Fort Worth: Harcourt Brace College Publishers (5ème édition), p.671-672.

ANALYSER LES VARIABLES UNE À UNE

Avant de mettre en relation les variables, il est recommandé de commencer par analyser chacune des variables (distribution, tendance centrale, dispersion), indépendamment des autres. Pour une présentation détaillée de l'analyse d'une seule variable quantitative, on pourra se reporter à l'exemple présenté par ailleurs : l'analyse du dossier NOTEBAC.

```
Lancer SESAnalyse
Menu Fichier – Ouvrir un dossier SES (*.SES)
Sélectionner le dossier FAMILY.SES
Menu Nouvelle analyse
Sélectionner PÈRE, MERE, NOTE comme Prédictrices (VI) et FILLE comme Observée (VD)
```

Analyser les variables prédictrices

Forme des distributions

```
SES-Pegase
Menu Analyses VI
Sélectionner les VI une à une pour obtenir les histogrammes
```

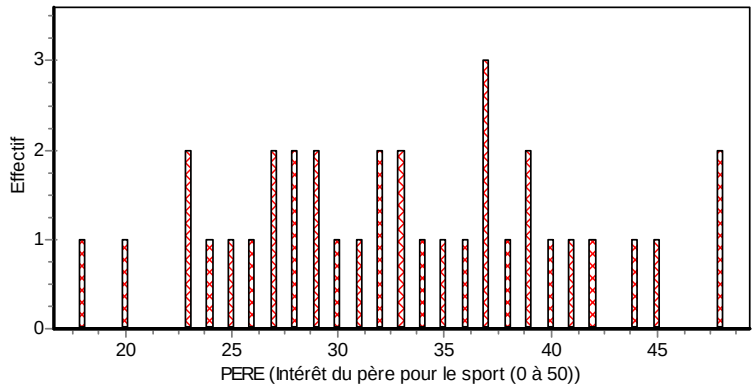

Figure 1 : Distribution des valeurs de PERE

Tendance centrale des valeurs

```
SES-Pegase
Menu Analyses VI – Tendance centrale ? – Indices de tendance centrale
```

Tableau 3 : Indices de tendance centrale des prédictrices

	PERE	MERE	NOTE
Moy	33.1	29.5	10.7
Med	17.5	29.5	10.5

Dispersion des valeurs

```
SES-Pegase
Menu Analyses VI – Dispersion ? – Minimum et maximum
```

Tableau 4 : Minimum, maximum et écart-type des prédictrices

	PERE	MERE	NOTE
Min	18	14	4
Max	48	43	19
Ety	7.6	6.6	3.8

Analyser la VD

Forme de la distribution

```
SES-Pegase
Menu Analyse VI
Sélectionner les VI une à une pour obtenir les histogrammes
```

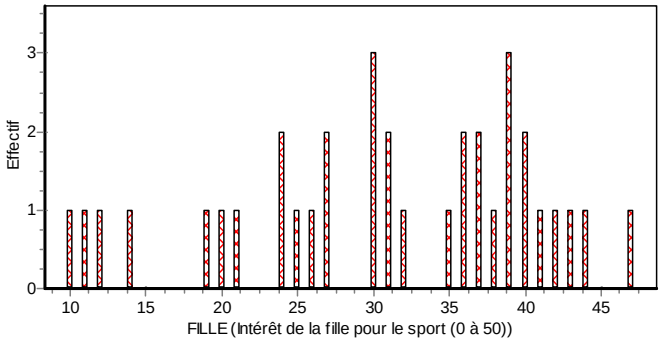

Figure 2 : Distribution des valeurs de FILLE

On constate une dissymétrie du côté des valeurs faibles (peu de filles avec un intérêt pour le sport entre 10 et 15)

Tendance centrale des valeurs

```
SES-Pegase
Menu Analyses VD – Tendance centrale ? – Indices de tendance centrale
```

Tableau 5: Indices de tendance centrale de la VD

	FILLE
Moy	30.8
Med	30.5

Dispersion des valeurs

```
SES-Pegase
Menu Analyses VD – Dispersion ? – Minimum et maximum
```

Tableau 6: Minimum, maximum et écart-type de FILLE

	FILLE
Min	10
Max	47
Ety	9.8

ANALYSER LES LIAISONS BIVARIÉES ENTRE PRÉDICTRICES

Pour une présentation détaillée de l'analyse de la liaison entre deux variables quantitatives, on pourra se reporter à l'exemple présenté par ailleurs : l'analyse du dossier INTELLIGENCE.

```
SES-Pegase
Analyse - Prédire la première variable – Analyser les liaisons entre VI
Pour visualiser les différents graphiques, utiliser la boite de dialogue, à droite du graphique.
```

Nuages bivariés entre prédictrices

```
SES-Pegase
Menu Analyses VI – Forme des liaisons entre couples de modalités – Nuages bivariés
```

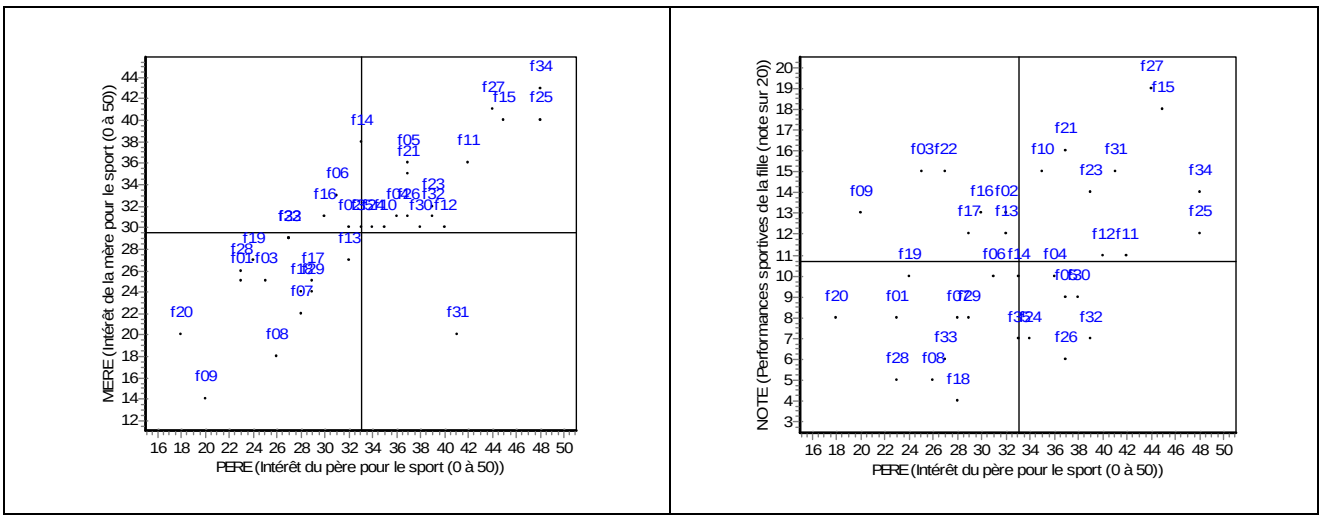

Figure 3 : Nuages bivariés entre prédictrices

Il apparaît sur ces deux graphiques que la liaison entre PERE et MERE (intérêt du père et de la mère pour le sport) est plus forte que celle entre intérêt du père pour le sport (PERE) et les performances physiques de la fille (NOTE).

Corrélations (simples) entre prédictrices

```
SES-Pegase
Menu Analyses par VI – Liaisons linéaires ? Sens des liaisons/VI.
```

Tableau 7: Sens des liaisons linéaires entre prédictrices

r	PERE	MERE	NOTE
PERE	+	+	+
MERE	+	+	+
NOTE	+	+	+

Les liaisons linéaires entre prédictrices sont toutes positives.

```
SES-Pegase
Menu Analyses par VI – Liaisons linéaires ? Force des liaisons/VI - Corrélations.
```

Tableau 8: Force des liaisons linéaires entre prédictrices

r	PERE	MERE	NOTE
PERE	1.00	0.77	0.42
MERE	0.77	1.00	0.38
NOTE	0.42	0.38	1.00

Les valeurs indiquées dans le tableau sont celles du coefficient de corrélation linéaire de Bravais-Pearson, noté *r.*
Ce tableau confirme ce qui était visible sur les nuages bivariés :
- toutes les corrélations entre prédictrices sont positives,
- la corrélation entre PERE et MERE (r = +.77) est la plus forte,
- les deux autres corrélations sont toutefois également fortes (supérieures à .40) ou modérées (comprises entre .20 et .40).

Rechercher les colinéarités entre prédictrices

Avant de commencer l'analyse de régression où on cherche à prédire la prédictrice à partir des VD, on vérifie que l'une des prédictrices n'est pas une combinaison linéaire des autres prédictrices *(Multicolinéarité*) auquel cas il faudra la retirer de l'analyse. Cette précaution est particulièrement nécessaire pour les analyses inférentielles (tests, intervalles de confiance, etc.)
Réf. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). p.98-100, 419-425.

```
SES-Pegase
Menu Analyses VI - Liaisons Variable * Autres variables – R², R et Tolérance.
```

Tableau 9: Indices de multicolinéarité entre prédictrices

	PERE	MERE	NOTE
R²	61.7%	60.0%	18.7%
VIF	2.61	2.5	1.23

Le Tableau 9 indique, pour chaque VP, la force de sa liaison linéaire (colinéarité) avec les autres VP. Ainsi, dans la colonne PÈRE, on a les indices de la colinéarité entre PÈRE d'une part, MERE et NOTE d'autre part. On retiendra le R^2 (61.7%) qui indique la part de la variance de PÈRE dont rend compte MERE et NOTE réunis.
En cas de colinéarité parfaite on aurait $R^2 = 100\%$. Généralement, on considère qu'il y a un problème de colinéarité si R^2 est supérieur à 90% (Cohen et al. 2003).

On ne décèle pas ici de problème de multicolinéarité pour aucune des trois prédictrices ($R^2 < 90\%$).

ANALYSER LES LIAISONS ENTRE CHAQUE PRÉDICTRICE ET LA VARIABLE À PRÉDIRE

Avant de mettre en œuvre la régression il est utile d'analyser les liaisons entre chaque prédictrice et la VD. Une première manière de procéder est de regarder les nuages bivariés pour analyser la nature de la liaison (de type linéaire ou autre). Du point de vue des indicateurs numériques, le plus simple est de calculer le coefficient de corrélation de Bravais-Pearson. Mais dans ce contexte de la régression multiple, on s'intéressera plutôt aux corrélations semi-partielles. En effet la corrélation entre une VD et une prédictrice dépend des corrélations entre cette prédictrice et les autres. Les corrélations semi-partielles indiquent, pour chaque prédictrice, la corrélation entre la VD et la part propre de cette prédictrice (une fois enlevés ses liens avec les autres prédictrices).

Forme des liaisons ?

```
SES-Pegase
Menu Analyse - Prédire la première variable – Analyser les poids de chaque VI – Nuages bivariés entre VI et VD.
Utiliser la boite de dialogue, à droite des graphiques, pour sélectionner la VI à mettre en relation avec la VD.
```

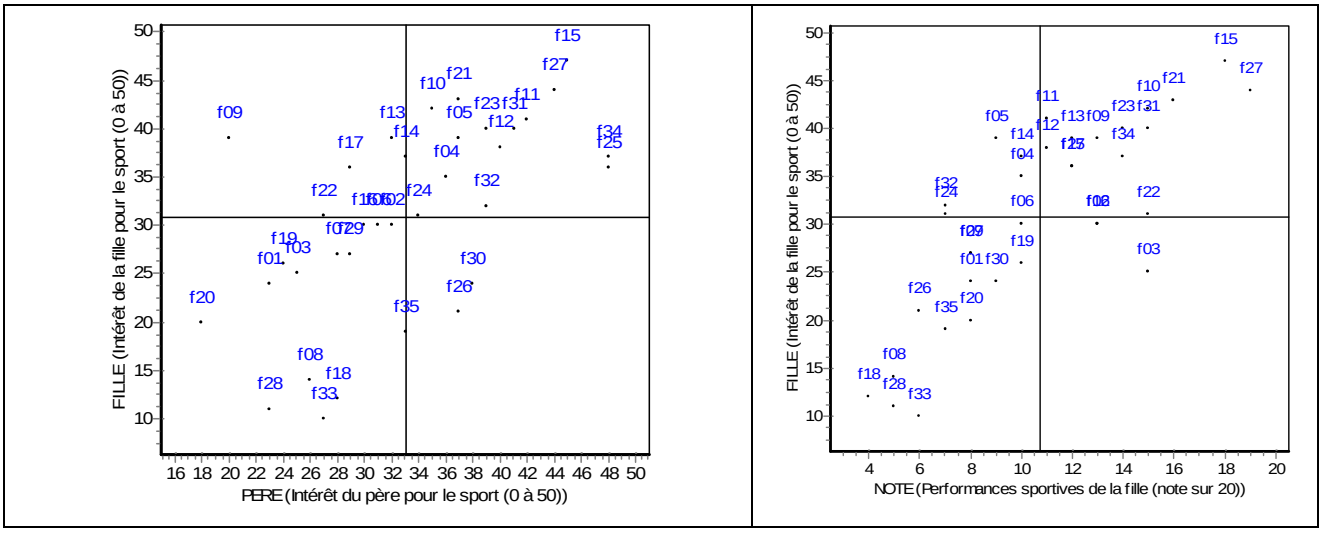

Figure 4 : Nuages bivariés entre variables prédictrices et variable à prédire (FILLE)

Comme pour l'analyse des relations entre prédictrices, l'analyse des nuages bivariés entre VD et prédictrices permet de voir si les liaisons sont de type linéaire, s'il existe des valeurs atypiques, s'il existe des groupes d'individus, etc.
cf. le document « Des individus et deux variables quantitatives ».

Concernant la liaison entre PERE et FILLE, la liaison apparaît globalement positive. Toutefois, il semble qu'il y ait :
- un point relativement atypique. Il s'agit de la famille f09 où le PERE a un faible intérêt pour le sport (environ 20) alors que la fille a un intérêt élevé pour le sport (environ 38),
- deux autres points atypiques (f25 et f34 en haut à droite) où le père a un intérêt très élevé pour le sport (environ 48) alors que la fille présente un intérêt plus faible (environ 35).
- deux groupes hétérogènes (cf. le groupe composé de f28, f33, f08, f18… f30, en bas à droite, séparé des autres).

Sens des liaisons linéaires ?

```
SES-Pegase
Menu Liaisons VI VD - Liaisons linéaires VI*VD ? – Sens des liaisons linéaires
```

Tableau 10: Sens des liaisons linéaires entre prédictrices et VD

r	FILLE
PERE	+
MERE	+
NOTE	+

Les liaisons linéaires entre le VD et les prédictrices sont toutes positives.

Force des liaisons linéaires ?

```
SES-Pegase
Menu Liaisons VI VD - Liaisons linéaires VI*VD ? – Force des liaisons linéaires
```

Tableau 11: Force des liaisons linéaires (coefficients de corrélation de Bravais-Pearson) entre les prédictrices et la VD

r	FILLE
PERE	+.61
MERE	+.47
NOTE	+.79

Les liaisons linéaires entre la VD (FILLE) et chacune des prédictrices (PERE, MERE, NOTE) apparaissent fortes (> .40).

ÉQUATION DE RÉGRESSION LINÉAIRE ?

L'objet central de la régression linéaire est l'équation de régression. On commencera donc par s'intéresser à cette équation. Cette équation permet de faire, pour chaque individu de l'échantillon, des prédictions (estimations) sur les valeurs de la VD, connaissant les valeurs des prédictrices. On peut enfin, pour un individu extérieur à l'échantillon, utiliser l'équation de régression pour estimer son score sur la VD, ne connaissant que ses scores sur les prédictrices.

```
Menu Analyse - Prédire la première variable – Analyser les prédictions – Équation de régression sur variables brutes
```

On obtient : FILLE_pred = -0,52 +0,52 x PERE -0,15 x MERE + 1,72 x NOTE

Cette équation permet de prédire, pour chaque unité (ici chaque famille) l'intérêt de la fille pour le sport (FILLE_pred) connaissant l'intérêt du père pour le sport (PERE), l'intérêt de la mère pour le sport (MERE) ainsi que ses performances physiques (NOTE).
Ainsi pour la famille f1 (où PERE = 23, MERE = 25, NOTE = 8) cette équation permet de prédire, pour la fille, un intérêt pour le sport égal à 22 (plus précisément 21.53) sur 50 :

FILLE_pred = -0,52 +0,52 x 23 -0,15 x 25 +1,72 x 8 = 21.5 (si on fait le calcul sur ces valeurs arrondies)

EXISTENCE D'UN LIEN GLOBAL ENTRE PRÉDICTRICES ET VD DANS L'ÉCHANTILLON ?

Cette première question, sur le lien global entre variables prédictrices et variable à prédire peut se décomposer en deux sous-questions :

1. Existe-t-il, ou non, un lien entre ces variables ? Ou, dis autrement, peut-on prédire, au moins en partie les valeurs individuelles de l'intérêt des filles pour le sport (FILLE), connaissant les valeurs sur les trois autres variables (PÈRE, MERE, NOTE) ?
2. Si un tel lien existe, quelle est la force de ce lien ? Est-il faible, ou au contraire fort ?

L'équation obtenue précédemment montre l'existence d'une liaison entre l'ensemble des variables prédictrices (VP) et la VD dans l'échantillon.

Il est toujours possible, sauf rare exception, de trouver une équation linéaire permettant de prédire, pour un échantillon, les valeurs de la variable à prédire, connaissant les valeurs des prédictrices. La question qui se pose alors est de savoir s'il existe un lien dans la population d'où provient cet échantillon. Le test *F* vise à répondre à cette question.

EXISTENCE D'UN LIEN GLOBAL ENTRE PRÉDICTRICES ET VD DANS LA POPULATION ?

```
SES-Pegase
Menu Analyse
```

Tableau 12: Test F de l'ANOVA

	F	ddl1	ddl2	p
FILLE	26.92	3	31	<.0001

Le test *F* permet d'évaluer si, dans la population, il existe au moins une prédictrice qui prédit une part non nulle de la VD. L'hypothèse nulle (H0) est : « Dans la population, toutes les prédictrices ont un poids nul ».
On peut voir également ce test comme le test d'une autre hypothèse nulle, équivalente à la précédente : « Dans la population, $R^2 = 0$ ».

Lorsque *p* est supérieur à .05 (test non significatif) on ne peut pas conclure à l'existence d'une liaison linéaire entre les prédictrices et la VD dans la population.
Lorsque *p* est inférieur à .05 (test significatif) on peut conclure à l'existence d'une liaison linéaire entre les prédictrices et la VD dans la population ; au moins une des VI a un poids dans la prédiction de FILLE.

Ce test *F* n'est qu'une première étape de l'analyse du poids des prédictrices dans la prédiction. En effet, ce test n'indique pas, lorsqu'il est significatif comme ici, si chacune des prédictrices a un poids dans la prédiction et si ce poids est important.

FORCE DU LIEN GLOBAL ENTRE PRÉDICTRICES ET VD DANS L'ÉCHANTILLON ?

Un indice brut : Les résidus

Sauf cas particulier, exceptionnel, les prédictions obtenues par l'équation de régression ne correspondent pas aux valeurs effectivement observées. On peut alors calculer, pour chaque individu, l'écart entre la valeur observée et la valeur prédite par l'équation de régression. On parle d'erreurs (de prédiction) ou encore de résidus pour nommer ces écarts.

```
SES-Pégase
Menu Analyse - Prédire la première variable – Prédictions – Observés, Prédits et résidus
Menu Liaisons VI VD – Régression linéaire – Protocole obs_pred_res
```

Le tableau ainsi obtenu permet de voir, famille par famille, les écarts entre les valeurs observées et les valeurs prédites par la régression.
Ainsi pour la famille f01, l'intérêt (observé) de la fille pour le sport est de 24 alors que l'équation lui prédit un intérêt de 21.5. L'écart entre la valeur observée (24) et la valeur prédite (21.5) est de 2.5 (2.47 plus précisément). C'est cette quantité (2.5) que l'on désigne par résidu ou erreur (de prédiction).

```
SES-Pégase
Placer le curseur de la souris sur une case du tableau pour afficher une valeur plus précise
(par exemple, pour f01, 21.52519 pour la valeur prédite.
```

Un premier moyen d'évaluer la qualité de la prédiction (qualité de l'ajustement) de la VD par les prédictrices est donc d'analyser les résidus. Des résidus élevés traduisent une prédiction de mauvaise qualité ; des résidus faibles traduisent une prédiction de bonne qualité.

Tableau 13 : Valeurs observées, prédites et résiduelles pour la prédiction de FILLE

FAMILLE	FILLE	FILLE_pred	FILLE_res
f01	24	21.53	2.47
f02	30	34.06	-4.06
f03	25	34.59	-9.59
f04	35	30.84	4.16
f05	39	28.91	10.09
f06	30	27.95	2.05
f07	27	24.56	2.44
f08	14	18.96	-4.96
f09	39	30.16	8.84
f31	40	43.64	-3.64
f32	32	27.25	4.75
f33	10	19.59	-9.59
f34	37	42.2	-5.20
f35	19	24.28	-5.28

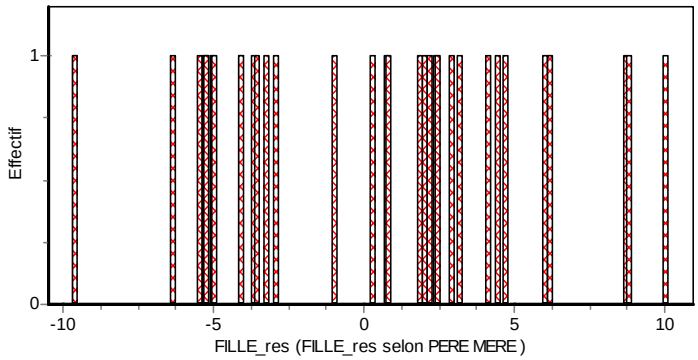

Figure 5 : Distribution des résidus de la régression

Les résidus ou erreurs de prédiction varient de 0 à 10 points (en valeur absolue). Toutefois, la grande majorité de ces résidus est inférieure à 5.[2]

SES-Pegase

Tableau 14 : Dispersion des résidus de la régression (écart-type et écart-type corrigé)

Ety	5.18
EtyC	5.50

En moyenne, l'erreur de prédiction est proche de 5 points (*Ety* = 5.18).

Un indice calibré : le coefficient de détermination, R^2

SES-Pégase
Menu Analyse - Prédire la première variable – Évaluer la qualité de la prédiction – Coefficient de détermination R^2.
Cliquer sur le bouton Inférer

Tableau 15 : Valeurs observées des coefficients R^2, R et R^2 ajusté

	R^2	R	R^2_ajusté
FILLE	72.30%	0.85	69.60%

On trouve R^2 = 72%. La combinaison linéaire de l'ensemble des prédictrices rend compte, globalement, de 72% de la variance de FILLE (R^2=72%).
Cette valeur apparaît importante (> 16%).

Lorsque le nombre de prédictrices est important, on peut préférer le R^2 ajusté qui compense la tendance de R^2 à augmenter en fonction du nombre de variables prédictrices.

[2] *Propriété : la moyenne des résidus est égale à 0.*

FORCE DU LIEN GLOBAL ENTRE PRÉDICTRICES ET VD DANS LA POPULATION ?

Tableau 16 : Valeur observée et intervalle de confiance du coefficient R^2

	R^2	Erreur-type	IC_inf	IC_sup
FILLE	72.30%	0.07	58%	86%

IC 95% sur R^2 = [58% ; 86%]. Il semble que, dans la population, R^2 est compris entre 58% et 86% (au niveau de confiance 95%).

ATTENTION: Le calcul de l'*IC* sur R^2 par SES-Pegase est une bonne approximation seulement pour *ddl* > 60 (ici, *ddl* = 31).
Référence : Cohen, Cohen, Weist, Aiken (2003), p.88.
Pour une meilleure estimation de l'intervalle de confiance sur R^2, on utilisera SES-Colibri. On trouve alors *IC* 95% = [49% ; 80%]

```
SES-Pegase
Menu Outils – SES-Colibri
Saisir F = 26.92, ddl1 = 3, ddl2 = 31 ; N_total= 35
Calculer
```

On obtient *IC* 95% = [49% ; 80%]. On peut conclure à un R^2 parent important (l'intervalle ne comprend que des valeurs supérieures à 16%).

Rédiger le compte rendu de l'analyse

Il semble que,
dans la population d'où est extrait cet échantillon,
une combinaison linéaire de trois variables, l'intérêt du père pour le sport (PÈRE), l'intérêt de la mère pour le sport (MERE) et la note en sport (NOTE),
permet de prédire une partie de la variance des intérêts des filles pour le sport (F (3 ; 31) = 26.92, p < .001).
La part de variance prédite apparaît forte (*IC* 95% = [49% ; 80%] ne comprend que des valeurs > 16%).

SENS DES LIAISONS ENTRE CHAQUE PRÉDICTRICE ET LA VD DANS L'ÉCHANTILLON ?

Pour analyser le poids de chacune des prédictrices dans la prédiction de la VD, on analyse les coefficients de régression partiels (coefficients b de la régression) qui indiquent le sens et la force de la liaison entre la prédictrice et la VD.

```
Menu Analyse - Prédire la première variable – Analyser les poids de chaque VI – Existence de liaisons (partielles) entre VI et VD.
Cliquer sur le bouton Inférer (cf. ci-dessus).
```

Tableau 17 : Coefficients b de la régression et tests t de Student sur les prédictrices

	PERE	MERE	NOTE
b	0.52	-0.15	1.72

Le fait que les coefficients soient tous différents de 0 indique que toutes les prédictrices (PERE, MERE, NOTE) ont un poids dans la prédiction de FILLE (Intérêt de la fille pour le sport).

Par ailleurs, le signe de ces coefficients b indique que, les autres variables étant maintenues constantes :
- si PERE augmente, FILLE tend également à augmenter (b > 0),
- si MERE augmente, FILLE tend à diminuer (b < 0),
- si NOTE augmente, FILLE tend également à augmenter (b > 0).

SENS DES LIAISONS ENTRE CHAQUE PRÉDICTRICE ET LA VD DANS LA POPULATION ?

Tableau 18 : Coefficients b de la régression et tests t de Student sur les prédictrices

	PERE	MERE	NOTE
b	0.52	-0.15	1.72
Erreur-type	0.2	0.22	0.27
t	2.64	-0.65	6.29
ddl	31	31	31
p	0.013	0.5189	<.0001

Chaque test *t* permet de conclure sur le poids de chaque prédictrice dans la population.
Lorsque le test est significatif, on peut en conclure que le poids de la prédictrice dans la population est de même sens que dans l'échantillon.
Si le test *t* est non significatif, on ne peut pas conclure sur le poids de la prédictrice dans la population.

N.B. : Un test *t* significatif ne signifie pas que l'effet de la prédictrice est important. Les analyses qui suivent permettent de se prononcer sur l'importance du poids de chaque prédictrice.

FORCE DES LIAISONS ENTRE CHAQUE PRÉDICTRICE ET LA VD, DANS L'ÉCHANTILLON

Des indices bruts : les coefficients de régression

Menu Analyse - Prédire la première variable – Analyser les poids de chaque VI – Ampleur des liaisons entre VI et VD
Cliquer sur le bouton **Inférer**

Tableau 19 : Coefficients b de la régression

Variables	PERE	MERE	NOTE
b	0.52	-0.15	1.72

Dans cet échantillon, les autres variables étant maintenues constantes, en moyenne :
- si PERE augmente de 1, FILLE tend à augmenter de 0.52 (cf. b),
- si MERE augmente, FILLE tend à diminuer de -0.15 (cf. b),
- si NOTE augmente de 1, FILLE tend à augmenter de 1.72 (cf. b).

Si on considère, dans ce contexte particulier (cette échelle est notée de 0 à 50) qu'un coefficient b est important lorsqu'il est supérieur à 0.50, on conclut ici à un effet important de PERE et de NOTE.

Des indices calibrés : corrélations et pourcentage de variance

Corrélations semi-partielles entre prédictrices et VD

SES-Pegase
Menu Analyse - Prédire la première variable – Analyser les poids de chaque VI – Corrélations semi-partielles entre VI et VD

Tableau 20 : Corrélations simples (r) et corrélations semi-partielles (sr) entre les prédictrices et la variable à prédire (FILLE)

	PERE	MERE	NOTE
r	+.61	+.47	+.79
sr	+.25	+.06	+.59

Chaque corrélation semi-partielle (sr) indique la force de la liaison entre la VD et la part de la prédictrice indépendante des autres prédictrices, c'est-à-dire une fois enlevée l'influence linéaire des autres prédictrices. Elles peuvent être comparées aux corrélations simples (r) qui étaient toutes fortes.
Lorsque l'on enlève de chaque prédictrice ce qu'elle a de commun avec les autres prédictrices :
- seule la liaison (semi-partielle) entre NOTE et FILLE reste forte (> .40).
- la liaison (semi-partielle) entre PERE et FILLE est modérée (comprise entre .20 et .40),
- la liaison (semi-partielle) entre MERE et FILLE apparaît alors faible (< .20).

Remarque : Ces corrélations semi-partielles sont celles observées sur l'échantillon. Il faudrait prolonger cette analyse par le calcul d'intervalles de confiance afin d'estimer les valeurs des corrélations semi-partielles dans la population.

Pourcentage de variance prédite par chaque prédictrice

Tableau 21 : Pourcentage de variance brut (r^2) et semi-partielles (sr^2) entre la VD et chaque prédictrice

Variables	PERE	MERE	NOTE
r^2	36.9%	21.7%	63.0%
sr^2	6.2%	0.4%	35.4%

Dans ce Tableau 21, r^2 est le carré de la corrélation simple (r) et sr^2 est égal au carré de la corrélation semi-partielle (sr).
Ces indices sont donc équivalents aux indices du Tableau 20 mais ils présentent l'avantage de s'exprimer comme des pourcentages de variance. L'indice sr^2 est la part propre à chaque prédictrice (les autres prédictrices sont maintenues constantes) dans la prédiction de la variance de la VD. Les valeurs de cet indice sont à comparer à celles de l'indice r^2 qui étaient toutes fortes (> 16%).
Lorsque l'on enlève de chaque prédictrice ce qu'elle a de commun avec les autres prédictrices, on constate que :
- la part de variance due à NOTE reste forte (> 16%),
- celle due à PÈRE est, seulement, modérée (entre 4% et 16%),
- celle due à MERE apparaît faible (< 4%).

On retrouve, nécessairement, des conclusions équivalentes aux conclusions précédentes formulées à partir des corrélations.

Remarque : Ces parts de variance sont celles observées sur l'échantillon. Il faudrait prolonger cette analyse par le calcul d'intervalles de confiance afin d'estimer les valeurs de ces pourcentages dans la population.

FORCE DES LIAISONS ENTRE CHAQUE PRÉDICTRICE ET LA VD, DANS LA POPULATION

Faute de disposer des éléments nécessaires pour faire de l'inférence sur les autres indices, on s'intéresse ici uniquement aux coefficients de régression.

Tableau 22 : Coefficients b observés et Intervalles de confiance (95%) sur ces coefficients b

Variables	PERE	MERE	NOTE
b	0.52	-0.15	1.72
Erreur-type	0.20	0.22	0.27
IC_inf	0.12	-0.60	1.16
IC_sup	0.92	0.31	2.27

La première ligne rappelle les valeurs des coefficients de régression partiels (b) observés dans l'échantillon.
Les deux dernières lignes de ce tableau indiquent, pour chaque prédictrice, les limites inférieures et supérieures des intervalles de confiance (95%) sur la valeur de ces coefficients dans la population parente :

Ainsi, pour PERE, IC 95% = [+0.12 ; +0.92]. Dans la population, le coefficient de régression est compris entre 0.12 et 0.92 (au seuil p = .05). Cet intervalle confirme que le coefficient de régression est positif dans la population (l'intervalle ne comprend que des valeurs positives). Le test t nous avait déjà indiqué ce résultat.

Selon le critère défini précédemment (poids fort de la prédictrice si le coefficient de régression est supérieur à 0.50) :
- on ne peut pas conclure à un effet important de PERE (l'intervalle comprend des valeurs inférieures à 0.50)
- on ne peut pas conclure à un effet important de MERE (l'intervalle comprend des valeurs inférieures à 0.50)
- on peut conclure à un effet important de NOTE (l'intervalle ne comprend que des valeurs supérieures à 0.50)

Rédiger le compte rendu de l'analyse

Il semble que,
dans la population d'où provient cet échantillon,
les autres variables étant maintenues constantes :
- en moyenne, comme dans l'échantillon, si PERE augmente, FILLE tend également à augmenter (t[31]=2.64, p=.013 <.05, Significatif).
- on ne peut pas se prononcer sur le sens du lien entre MERE et FILLE (t[31] = -0.65, p = .52 >.05, Non Significatif),

- en moyenne, comme dans l'échantillon, si NOTE augmente, FILLE tend également à augmenter (t[31]=6,29, p = <.0001 <.05, Significatif).

Si l'on considère qu'une prédictrice a un poids fort dès lors que son coefficient de régression est supérieur à 0.50 :
- on ne peut pas conclure à un effet important de PERE (IC95% sur b = [+0.12 ; +0.92] comprend des valeurs inférieures à 0.50),
- on ne peut pas conclure à un effet faible de MERE (*IC*95% sur b = [-0.60 ; +0.31] comprend des valeurs inférieures et des valeurs supérieures à 0.50),
- on peut conclure à un effet important de NOTE (*IC*95% sur b = [+1.16 ; +2.27] ne comprend que des valeurs supérieures à 0.50).

COMPARER LES POIDS DES PRÉDICTRICES

On se demande quelle est, parmi l'ensemble des variables prédictrices, la meilleure prédictrice, une fois enlevé l'effet des autres variables.
Si les variables sont toutes sur la même échelle, on peut se prononcer en comparant les coefficients b (cf. Tableau 19).
Dans le cas contraire (le plus courant) comme ici, lorsque les variables ne sont pas sur la même unité, on comparera les coefficients Beta

Ici les variables ne sont pas sur la même échelle : on a des scores sur une échelle d'intérêt pour le sport, de 0 à 50, et des performances physiques notées de 0 à 20. On ne peut pas comparer ces coefficients. En effet leur valeur dépend du codage choisi pour chaque variable. Ainsi, les coefficients auraient été différents si la performance physique de la fille avait été notée sur 10 ou sur 40 au lieu de 20.

Coefficients Beta

Les coefficients Beta sont les coefficients de régression obtenus après avoir centré et réduit l'ensemble des variables (prédictrices et VD). Les variables étant alors toutes sur la même échelle (moyenne= 0 et écart-type= 1) les coefficients de régression peuvent être comparés.

```
Menu Analyse - Prédire la première variable – Comparer le poids des VI – Coefficients Beta
```

Tableau 23 : Coefficients de régression sur les variables centrées-réduites (scores Z)

Variables	PERE	MERE	NOTE
Beta	0.40	-0.10	0.66
Erreur-type	0.15	0.15	0.1
IC_inf	0.09	-0.4	0.45
IC_sup	0.71	0.21	0.87

L'équation de régression sur ces variables centrées et réduites (scores Z) est donc :

FILLE_Zpred = +0,40 x PERE_Z -0,10 x MERE_Z + 0,66 × NOTE_Z

Pour cet échantillon, c'est la variable NOTE qui a le poids le plus fort (Beta = 0.66).

Rédiger le compte rendu de l'analyse

On constate que,
pour cet échantillon de 35 familles,
si l'on procède à une régression sur les variables centrées-réduites (scores Z),
c'est la note en sport (NOTE) qui est la meilleure prédictrice de l'intérêt de la fille pour le sport (FILLE)
(Beta = 0.66, cf. Tableau 23)
devant l'intérêt du père pour le sport (Beta = 0.40) et l'intérêt de la mère pour le sport (Beta = -0.10)

< insérer Tableau 23 >

RECHERCHER LE MEILLEUR SOUS-ENSEMBLE DE PRÉDICTRICES

Lorsque l'on dispose de plusieurs prédictrices supposées prédire la VD et dont on connaît le pouvoir prédictif (ici R^2 = 72%), il peut être intéressant de rechercher s'il existe un sous-ensemble plus petit de prédictrices (ici deux prédictrices au lieu de trois) qui rendrait suffisamment compte à lui seul des variations des valeurs de la VD chez les individus (FILLE).

```
SES-Pégase
Menu Analyses VI VD – Régression linéaire – Meilleur sous-ensemble de prédictrices.
```

Tableau 24: Parts de variance prédites (R^2) de FILLE par les différents couples de prédictrices

VI	VI	**VD**	**R^2**
PERE	MERE	FILLE	36.9%
PERE	NOTE	FILLE	71.9%
MERE	NOTE	FILLE	66.1%

Ce tableau indique les trois combinaisons possibles de deux prédictrices parmi les trois (cf. les deux premières colonnes). La dernière colonne indique le R^2 correspondant à chacune de ces paires.

C'est la paire (PÈRE, NOTE) qui est la meilleure paire de variables pour prédire l'intérêt de la fille pour le sport (R^2 = 71.9%).

Rédiger le compte rendu de l'analyse

On constate,
pour cet échantillon de 35 familles,
que le meilleur couple de prédictrices de l'intérêt de la fille pour le sport (FILLE)
est constitué de la note en sport (NOTE) et de l'intérêt du père pour le sport (PÈRE).
Ces deux variables prédisent une part importante de la variance de FILLE (R^2 = 71.9% > 16%).[3]

FAIRE DES PRÉDICTIONS EXTÉRIEURES À L'ÉCHANTILLON

Il est possible de calculer, à partir de l'équation de régression, une prédiction pour un individu extérieur à l'échantillon. Ainsi, supposons que l'on ne connaisse pas, pour une fille, son intérêt pour le sport, mais que l'on connaisse l'intérêt de son père pour le sport (30), l'intérêt de sa mère pour le sport (40) et enfin sa note en sport (15).

```
SES-Pégase
Menu Analyse - Prédire la première variable – Prédictions – Prédiction individuelle.
Saisir les trois valeurs Pour PÈRE, MERE et NOTE dans les cases respectives
Cliquer sur le bouton Calcul.
```

La prédiction (35) s'affiche dans la case de droite (« FILLE_pred »).

Il est souhaitable de calculer alors un intervalle de confiance sur la prévision

```
SES-Pegase
Non disponible dans la version actuelle de SES-Pegase).
```

[3] REMARQUE : L'ajout de la troisième variable fait très peu augmenter la part de variance prédite (cf. Tableau 15, R^2 = 72.3 %).

ANNEXE

Calcul du R^2

Les deux tableaux suivants affichent les statistiques (sommes de carrés et variances) qui sont à la base du calcul du coefficient de détermination R^2.

> SES-Pegase
> Menu **Analyse - Prédire la première variable – Évaluer la qualité de la prédiction – Décomposition des sommes de carrés (SC)**

Tableau 25 : Décomposition de la somme des carrés de la régression

SC_pred	2443
SC_res	937
SC_totale	3380

On constate la propriété de décomposition des sommes de carrés : la somme des carrés totale (3380) est égale à la somme des carrés prédite par la régression (2443) + la somme des carrés résiduelle (937.5).
Le coefficient de détermination R^2 permet d'évaluer la part de SC_pred par rapport à la SC_totale.
R^2 = SC_pred / SC_totale = 2443 / 3380 = 0.72 (72%).

De la même manière qu'avec les sommes de carrés, R^2 peut être calculé à partir des variances prédites et résiduelles :

> Menu **Analyse - Prédire la première variable – Analyser le poids global des VI – Décomposition des variances**

Tableau 26 : Décomposition de la somme des carrés de la régression

V_pred	69.8
V_res	26.8
V_total	96.6

On constate la propriété de décomposition des variances : la variance totale (96.6) est égale à la variance prédite (69.8) + la variance résiduelle (26.8).
R^2 = V_pred / V_totale = 69.8 / 96.6 = 0.72 (72%).

Corrélations simples (r), semi-partielles (sr) et partielles (pr) entre VP et VD

> SES-Pégase
> Menu **Analyse - Prédire la première variable – Analyser les poids de chaque VI – Corrélations partielles entre VI et VD**

Variables	PERE	MERE	NOTE
r	+.61	+.47	+.79
sr	+.25	+.06	+.59
pr	+.43	-.12	+.75

Chaque corrélation semi-partielle (sr) indique le signe et la force de la liaison entre la VD et chaque VP débarrassée de ce qu'elle a en commun avec les autres VP.
Chaque corrélation partielle (pr) indique le signe et la force de la liaison entre la VD et chaque VP, toutes deux « débarrassées » de ce qu'elles ont en commun avec les autres prédictrices.

Part de variance (r^2), semi-partielles (sr^2) et partielles (pr^2) entre VP et VD

Tableau 27 : Pourcentage de variance brut (r^2), semi-partielles (sr^2) et partielles (pr^2) entre la VD et chaque prédictrice

Variables	PERE	MERE	NOTE
r^2	36.9%	21.7%	63.0%
sr^2	6.2%	0.4%	35.4%
pr^2	18.3%	1.4%	56.1%

Chaque pourcentage de variance semi-partiel (sr^2) indique le pourcentage de variance de la VD, prédit par la VP débarrassée de ce qu'elle a en commun avec les autres VP.
Chaque pourcentage de variance partiel (pr^2) indique le pourcentage de variance de la VD, prédit par la VP, toutes deux « débarrassées » de ce qu'elles ont en commun avec les autres prédictrices.

RÉFÉRENCES

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Third Edition)*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates.

LISTE DES TABLEAUX

LISTE DES FIGURES

SOMMAIRE