Exercice 6 : Étude d'un modèle de régression multiple

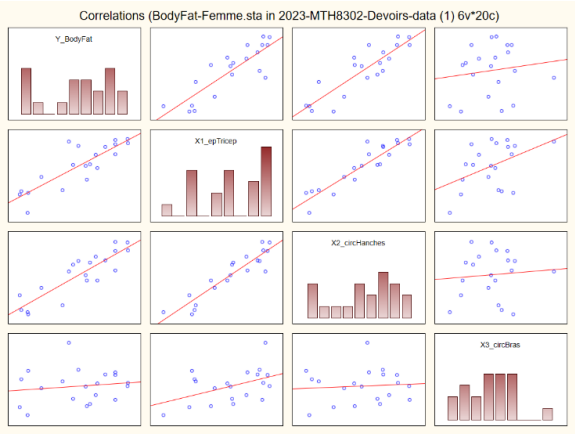
6a)

Commençons par calculer la matrice de corrélation afin de vérifier s'il existe une forte corrélation entre les variables et de déterminer s'il y a un possible problème de multicollinéarité.

Figure 1: Matrice de corrélation

Variable	X1_epTricep	X2_circHanches	X3_circBras	Y_BodyFat
	Correlations (BodyFat-Femme.sta in 2023-MTH8302-Devoirs-data (1)) Marked correlations are significant at p < ,05000 N=20 (Casewise deletion of missing data)			
X1_epTricep	1,000000	0,923843	0,457777	0,843265
X2_circHanches	0,923843	1,000000	0,084667	0,878090
X3_circBras	0,457777	0,084667	1,000000	0,142444
Y_BodyFat	0,843265	0,878090	0,142444	1,000000

Figure 2: Scattergramme global



Dans la matrice de corrélation que nous avons calculée, nous avons observé une forte corrélation positive entre X2_circHanches et X1_epTricep. Nous avons également remarqué une corrélation positive significative entre X1 et Y_BodyFat ainsi que entre X2 et Y_BodyFat, avec des coefficients respectifs de 0,843 et 0,878. De plus, nous avons observé sur le scatterplot que X1 et X2 augmentent ensemble. Il est important de noter qu'il n'y a aucune corrélation supérieure ou égale à 0,95, le maximum étant de r = 0,92 entre X1 et X2. Cependant, afin de confirmer l'existence d'une possible multicollinéarité, il est nécessaire de calculer le VIF et l'IC.

Calculons maintenant le variance inflation factor (VIF).

Le critère 2 pour détecter la présence d'une multicollinéarité est que tous les VIF (variance inflation factor) doivent être inférieurs à 10 pour toutes les variables incluses dans le modèle.
VIF j = 1 / (1 – R2 j)

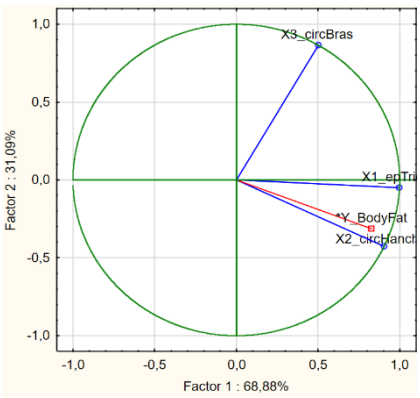
	R2	VIF
X1	0,711	3,461
X2	0,771	4,368
X3	0,020	1,021

Dans notre cas, nous avons constaté que tous les VIF étaient inférieurs à 10 pour toutes les variables incluses dans le modèle, ce qui signifie que Critere 2 est non satisfait VIF<10.

Calculons maintenant l'IC.

	Eigenvalue	% Variance	Eigenvalue Cum	% Cumulative	IC
X1	2,066473	68,88242	2,066473	68,8824	1,00
X2	0,932801	31,09336	2,999273	99,9758	2,22
X3	0,000727	0,02422	3,000000	100,0000	2843,95

Figure 3: Cercle de correlation



Le critère 3 pour détecter la présence d'une multicollinéarité est que l'IC (indice de conditionnement) doit être supérieur à 100. L'IC mesure la gravité de la multicollinéarité en évaluant la stabilité numérique de la matrice de régression.

Dans notre le Critère 3 est satisfait, IC > 100 donc on peut dire qu'il y a multicollinéarité.

On peut donc dire que la forte corrélation positive entre les variables observée dans la matrice de corrélation et sur le scattergramme est valide. On verra par la suite que le modèle de régression ordinaire (MRO) n'est pas satisfaisant.

6b) Modèle de Régression Ordinaire (MRO)

Figure 4: Tableau D'ANOVA

Analysis of Variance; DV: Y_BodyFat (BodyFat-Femme

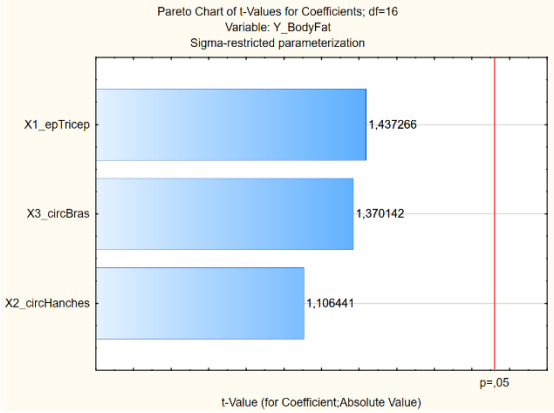
Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	396,9846	3	132,3282	21,51571	0,000007
Residual	98,4049	16	6,1503		
Total	495,3895				

Figure 5: Regression Summary

Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Femme
R= ,89518632 R²= ,80135855 Adjusted R²= ,76411328
F(3,16)=21,516 p<,00001 Std.Error of estimate: 2,4800

N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			117,0847	99,78240	1,17340	0,257808
X1_epTricep	4,26370	2,966538	4,3341	3,01551	1,43727	0,169911
X2_circHanches	-2,92870	2,646956	-2,8568	2,58202	-1,10644	0,284894
X3_circBras	-1,56142	1,139602	-2,1861	1,59550	-1,37014	0,189563

Figure 6: Diagramme de Pareto



On remarque sur le diagramme de Pareto, Figure 6, ainsi que sur le tableau de régression, Figure 5, qu'aucune des variables n'est significative au seuil alpha de 5 %. Toutefois, le tableau ANOVA, Figure 4, nous indique que le modèle global est significatif.
Un R2 de 0.80 indique que le modèle de régression explique 80% de la variance totale de la variable dépendante, ce qui est considéré comme un bon ajustement du modèle.

Analyse des résidus:

Figure 7: Normal Plot

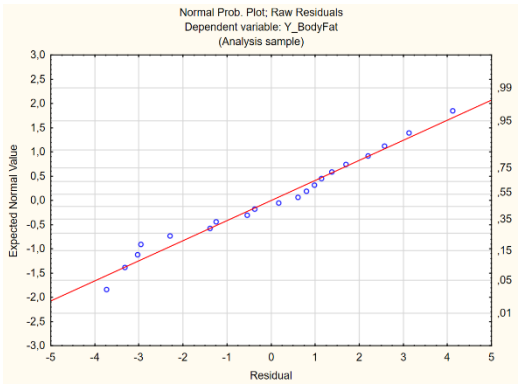


Figure 8: Predicted vs residuals

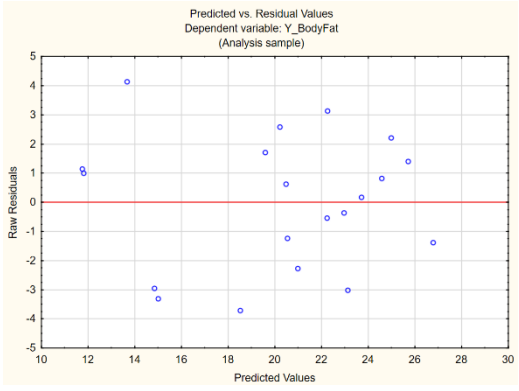
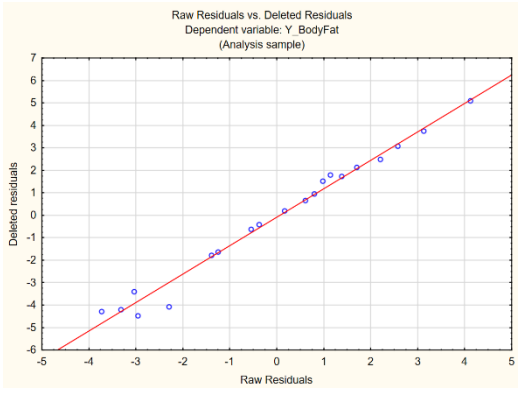


Figure 9: Residuals vs deleted



D'après la figure 8 (Predicted vs Residuals) on remarque que les observations sont distribuées à peu près symétriquement par rapport à 0 et de façon aléatoire, ce qui atteste d'une bonne homogeneité et donc constance de la variance. On peut également observer sur la figure 7 des résidus sur une échelle normale que les observations sont à peu pres aligné avec la droite normale ce qui veut dire qu'ils suivent une distribution normale. Enfin pour la figure 9 (Residuals vs deleted) on observe que les observations suivent la droite et qu'il n'y'a pas de données aberantes. On peut conclure de cette analyse que la variance est a peu prêt constante mais avec un meilleur ajustement on peut encore améliorer notre modèle.

Coefficient du modèle MRO

Coefficient pour X1_epTricep : 4.3341

Coefficient pour X2_circHanches : -2.8568

Coefficient pour X3_circBRas : -2.1861

Les coefficients de régression pour X1, X2 et X3 ont été estimés comme suit: le coefficient de régression pour X1 est positif, ce qui suggère une relation positive entre X1 et Y. Autrement dit, une épaisseur élevée du triceps peut indiquer une augmentation du pourcentage de graisse corporelle. Les coefficients pour X2 et X3 sont négatifs, ce qui suggère une relation négative entre ces variables et Y. Cependant, ces résultats semblent contre-intuitifs car on pourrait s'attendre à ce qu'une augmentation de X2 et X3 soit associée à une augmentation de Y (le pourcentage de graisse corporelle). Ces résultats peuvent indiquer que l'estimation des signes des coefficients est erronée ou que la relation entre ces variables et Y est non linéaire ou encore qu'il y a des variables manquantes qui peuvent influencer les résultats.

Le modèle global est significatif, mais aucun des coefficients n'est significatif au seuil de 5 %. On peut voir sur le diagramme de Pareto, Figure 5, qu'aucun des facteurs n'est significatif. Tout d'abord, nous avons vu, à l'aide de la matrice de corrélation, qu'il n'y avait aucune corrélation supérieure ou égale à 0,95. Le maximum est r = 0,92 entre X1 et X2 (critère 1 non satisfait).
Aucun des coefficients du modèle n'est significatif, ce qui suggère que les variables explicatives ne contribuent pas de manière significative à la prédiction de la variable dépendante. De plus, un problème de multicollinéarité est observé, ce qui indique que certaines variables explicatives sont fortement corrélées entre elles. En effet, comme le montre le cercle de corrélation (Figure 3), La perpendicularité entre X2 et X3 indique qu'il n'y a pas de lien entre ces dernières. En revanche, les variables X1 et X2 pointent dans la même direction, ce qui indique une forte corrélation entre elles. Le modèle n'est donc pas satisfaisant, et des ajustements sont nécessaires pour obtenir un modèle plus satisfaisant.

Si on souhaite conserver toutes les variables et éviter la multicolinéarité, on peut utiliser des méthodes de régression pénalisée telles que la régression Ridge, Lasso, PLS ou Elastic Net. Ces méthodes ajoutent une pénalité à la fonction de coût de la régression afin de réduire les coefficients des variables qui contribuent à la multicolinéarité. On peut transformer les variables explicatives en utilisant des techniques telles que la standardisation pour améliorer l'ajustement du modèle. Ensuite, ajuster le modèle en utilisant des techniques telles que la régression ridge ou la régression Lasso pour pénaliser les coefficients des variables moins importantes et améliorer ainsi la prédiction de la variable dépendante.

Pour améliorer notre modèle en incorporant toutes les variables X, nous allons considérer deux modèles : le modèle de régression RIDGE et le modèle d'analyse en composantes principales (ACP).

6c)

Modèle1 : Régression Ridge

J'ai choisi k=1 pour le modèle de Ridge car c'est la valeur optimale qui minimise l'erreur de prédiction et qui donne le meilleur coefficient de détermination. Cela signifie que ce modèle produit des prévisions plus précises et explique mieux la variance des données que les autres valeurs de k testées. En utilisant une méthode de validation croisée, j'ai constaté que la valeur de k=1 a produit le plus faible RMSE et le plus élevé R^2, ce qui indique une bonne performance de prédiction et un ajustement satisfaisant aux données.

Figure 10: Regression Summary

N=20	Ridge Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Fen l=,10000 R= ,86003142 R²= ,73965404 Adjusted R²= ,69083917 F(3,16)=15,152 p<,00006 Std.Error of estimate: 2,8392					
	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			-9,96277	11,22102	-0,887867	0,387765
X1_epTricep	0,423354	0,294687	0,43034	0,29955	1,436623	0,170091
X2_circHanches	0,448960	0,268753	0,43795	0,26216	1,670528	0,114257
X3_circBras	-0,081246	0,160434	-0,11375	0,22462	-0,506410	0,619476

Figure 11: Tableau D'ANOVA

Effect	Analysis of Variance; DV: Y_BodyFat (BodyFat-Femme Ridge regression, lambda=,1000000				
	Sums of Squares	df	Mean Squares	F	p-value
Regress.	366,4168	3	122,1389	15,15223	0,000062
Residual	128,9727	16	8,0608		
Total	495,3895				

Y_RIDGE = -9.96277 + 0.43*X1 + 0.43795*X2 – 0.11375*X3

Analyse des résidus:

Figure 12: Normal Plot

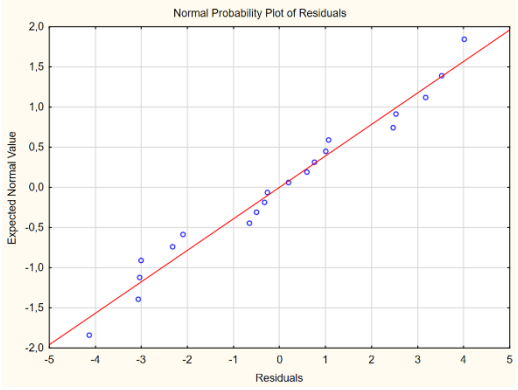


Figure 13: Predicted vs residuals

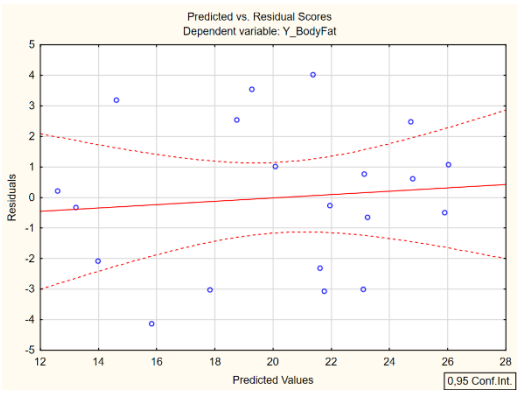
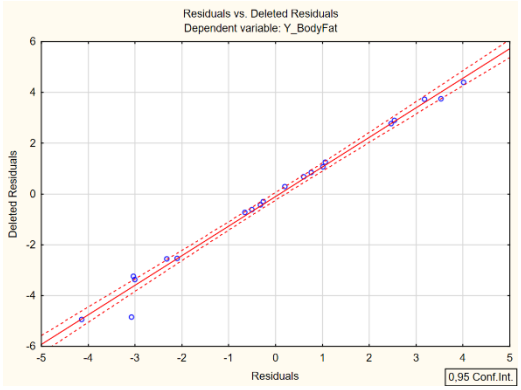


Figure 14: Residuals vs Deleted



Sur la figure "Predicted vs Residuals" que les résidus sont aléatoires et répartis uniformément autour de zéro, ce qui suggère une bonne homogénéité et une constance de la variance. En outre, en examinant la figure des résidus sur une échelle normale, on remarque que les observations sontapproximativement alignées avec la droite normale, indiquant qu'elles suivaient une distribution normale. Enfin, sur la figure "Residuals vs Deleted", les observations suivent la droite et il n'y a pas de données aberrantes, ce qui confirme la validité du modèle et la pertinence de sa capacité à expliquer les variations des données.

6d)

Modèle 2 : Régression Analyse en composantes principales (ACP)

La première étape de l'Analyse en Composantes Principales (ACP) est de centrer et de réduire les variables.

Figure 15 : Tableau Standardisé

1 ID	2 Genre	3 X1	4 X2	5 X3
1	femme	-1,2	-1,5	0,4
2	femme	-0,1	-0,3	0,2
3	femme	1,1	0,1	2,6
4	femme	0,9	0,6	1,0
5	femme	-1,2	-1,7	0,9
6	femme	0,1	0,5	-1,1
7	femme	1,2	1,4	-0,0
8	femme	0,5	0,2	0,8
9	femme	-0,6	-0,2	-1,2
10	femme	0,0	0,4	-0,8
11	femme	1,2	1,0	0,7
12	femme	1,0	1,1	0,2
13	femme	-1,3	-0,9	-1,3
14	femme	-1,1	-1,3	0,3
15	femme	-2,1	-1,6	-1,7
16	femme	0,8	0,6	0,7
17	femme	0,5	0,8	-0,5
18	femme	1,0	1,4	-0,8
19	femme	-0,5	-0,6	-0,1
20	femme	-0,0	-0,0	-0,0

Figure 16 : Graphique des projections

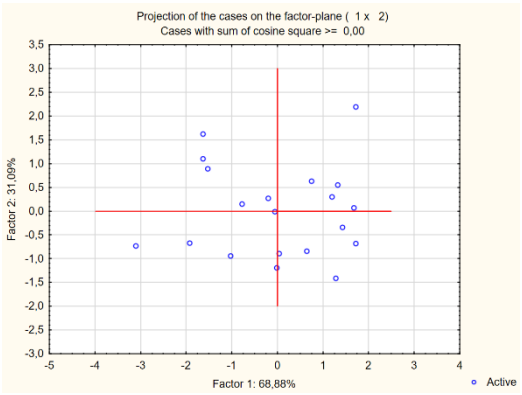
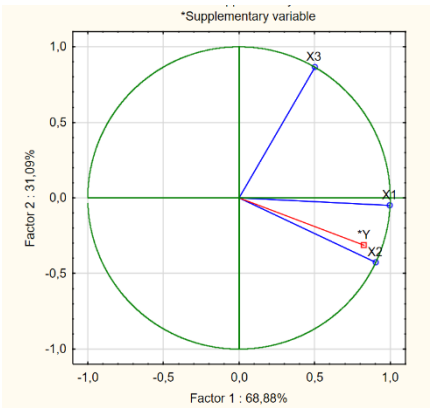


Figure 17: Cercle de correlation



Resultat de l'Analyse en Composantes Principales

Tableau des facteurs scores

Case	Factor scores, based on correlations		
	Factor 1	Factor 2	Factor 3
1	-1,13521	1,13971	1,71620
2	-0,13428	0,27321	1,38979
3	1,20298	2,26756	-0,91055
4	0,92535	0,56640	-0,09535
5	-1,12944	1,68030	-1,34605
6	-0,00358	-1,23842	0,12271
7	1,19816	-0,70671	-0,89854
8	0,52534	0,65047	1,21258
9	-0,70809	-0,98098	1,11772
10	0,02639	-0,92262	-1,66370
11	1,16975	0,07265	-0,56939
12	0,99788	-0,36142	0,01375
13	-1,33286	-0,70051	-0,91691
14	-1,05716	0,91457	-0,81827
15	-2,15804	-0,75966	-0,65847
16	0,83844	0,30631	0,65315
17	0,44837	-0,87316	-0,68397
18	0,89186	-1,46643	0,66450
19	-0,53357	0,15336	1,12097
20	-0,03229	-0,01464	0,54984

Variable	Factor score coefficients, based on co		
	Factor 1	Factor 2	Factor 3
X1	0,483259	-0,051879	26,6197
X2	0,437856	-0,456100	-23,7475
X3	0,242197	0,928074	-10,1826

Les pourcentages de variance expliquée par chaque facteur peuvent être visualisés sur le cercle de corrélation ainsi que sur le graphique des projections orthogonales de l'analyse en composantes principales (ACP). En effet, le cercle de corrélation permet de représenter les corrélations entre les variables et les projections orthogonales permettent de visualiser la contribution des variables à chaque facteur. Le première facteur explique 68.9% de la variance totale des données, ce qui représente la majorité de la variance. Le deuxième facteur explique quant à lui 31% de la variance totale, tandis que le troisième facteur explique seulement 0.02% de la variance.
Comme nous l'avons vu précédemment, le cercle de corrélation (Figure 17) montre que la perpendicularité entre les variables X2 et X3 indique qu'il n'y a pas de lien entre ces dernières. En revanche, les variables X1 et X2 pointent dans la même direction, ce qui indique une forte corrélation entre ces dernieres.
L'analyse des résidus du modèle ACP est similaire à celle du modèle Ridge, avec une constance de la variance satisfaisante, le respect de l'hypothèse de normalité et aucune donnée aberrante.

Régression « stepwise forward »

3 FS1	4 FS2	5 FS3	6 Y_BodyFat
-1,13521	1,13971	1,71620	11,9
-0,13428	0,27321	1,38979	22,8
1,20298	2,26756	-0,91055	18,7
0,92535	0,56640	-0,09535	20,1
-1,12944	1,68030	-1,34605	12,9
-0,00358	-1,23842	0,12271	21,7
1,19816	-0,70671	-0,89854	27,1
0,52534	0,65047	1,21258	25,4
-0,70809	-0,98098	1,11772	21,3
0,02639	-0,92262	-1,66370	19,3
1,16975	0,07265	-0,56939	25,4
0,99788	-0,36142	0,01375	27,2
-1,33286	-0,70051	-0,91691	11,7
-1,05716	0,91457	-0,81827	17,8
-2,15804	-0,75966	-0,65847	12,8
0,83844	0,30631	0,65315	23,9
0,44837	-0,87316	-0,68397	22,6
0,89186	-1,46643	0,66450	25,4
-0,53357	0,15336	1,12097	14,8
-0,03229	-0,01464	0,54984	21,1

Figure 18: Regression Summary

Regression Summary for Dependent Variable: Y_BodyFat (ACP-BodyFAt in
R= ,89518632 R²= ,80135855 Adjusted R²= ,76411328
F(3,16)=21,516 p<,00001 Std.Error of estimate: 2,4800

N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			20,19500	0,554541	36,41753	0,000000
FS1	0,826492	0,111423	4,22022	0,568947	7,41760	0,000001
FS2	-0,312046	0,111423	-1,59337	0,568947	-2,80055	0,012827
FS3	0,144559	0,111423	0,73814	0,568947	1,29739	0,212894

Figure 19: Tableau D'ANOVA

Analysis of Variance; DV: Y_BodyFat (ACP-BodyFAt ir

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	396,9846	3	132,3282	21,51571	0,000007
Residual	98,4049	16	6,1503		
Total	495,3895				

Etape 4 : relation entre les FS et les variables centrées-réduites

FS1: 0.4832*X1 + 0.4378*X2 + 0.242*X3

FS2: -0.05187*X1 -0.4561*X2 + 0.928*X3

Modèle de Y sur les variables d'origine centrées-réduites

Y= 20.19 + 4.22*(0.4832*X1 + 0.4378*X2 + 0.242*X3) -1.59*(-0.05187*X1 -0.4561*X2 + 0.928*X3)

Etape 5 : modèle de Y sur les variables d'origine centrées-réduites :

Y = 20.19+ 2.1215773X1 + 2.572715X2 - 0.45428*X3

Etape 6 : Modèle de Y_BodyFat sur les variables d'origine : X1_epTricep, X2_circHanches et X3_circBras

Relations entre les variables centrées-réduites et les variables d'origine:

X1= (X1_epTricep-25,3)/ 5,0 = 0.2*X1_epTricep - 5.06

X2= (X2_circHanches-51,2)/ 5,2= 0.1923* X2_circHanches – 9.846

X3= (X3_circBras-27,6)/ 3,6= 0.28* X3_circBras – 7.7

	X1_epTricep	X2_circHanches	X3_circBras
SD	5,0	5,2	3,6
Moyenne	25,3	51,2	27,6

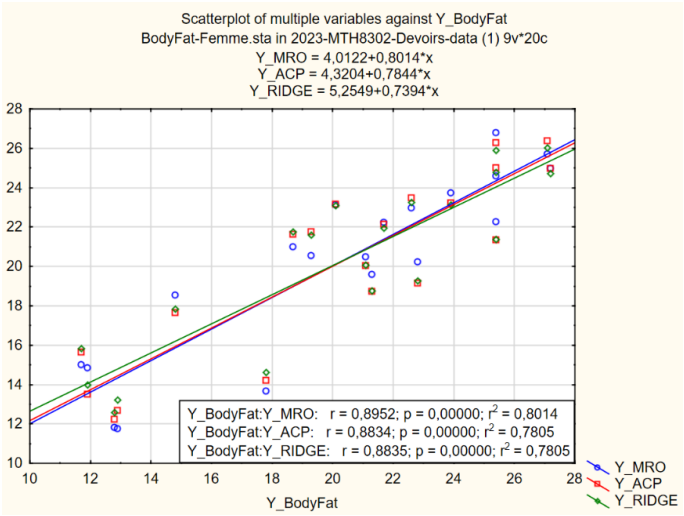
Y_BodyFat= 20.19+2.1215773*(0.2*X1_epTricep - 5.06)+ 2.572715*(0.1923* X2_circHanches – 9.846) - 0.45428*(0.28* X3_circBras – 7.7)

Équation de prédiction du modèle ACP:

Y_BodyFat = - 12.37817702+ 0.42431546*X1_epTricep + 0.49473309*X2_circHanches - 0.1271984*X3_circBras

6e) Comparaison des 2 modèles

6 Y_BodyFat	7 Y_MRO	8 Y_ACP	9 Y_RIDGE
11,9	14,85606	13,5174972	13,98775
22,8	20,22031	19,1531278	19,26039
18,7	20,98795	21,6186142	21,759085
20,1	23,12893	23,1745602	23,09429
12,9	11,75856	12,6735541	13,216845
21,7	22,24557	22,1358102	21,95486
27,1	25,71628	26,3765383	26,019805
25,4	22,27215	21,3435473	21,370675
21,3	19,59647	18,735373	18,754935
19,3	20,55017	21,7555672	21,611555
25,4	24,59733	25,0039747	24,7857
27,2	24,99415	24,9726644	24,72187
11,7	15,01087	15,6360476	15,826655
17,8	13,67345	14,2101659	14,61237
12,8	11,81327	12,2326057	12,59282
23,9	23,72912	23,2239373	23,122835
22,6	22,97546	23,4651022	23,24349
25,4	26,78798	26,2984283	25,88885
14,8	18,5277	17,6528422	17,824795
21,1	20,48947	20,0480042	20,080555



Calcul du AICc pour ACP

n	20
K	3
AICc	98,1213

Calcul du AICc pour RIDGE

n	20
K	4
AICc	101,47208

Tableau récapitulatif

	R^2	R^2 adjusted	MSE	SS_resid	AICc
Modèle RIDGE	0.74	0.69	8.06	128.9	101.5
Modèle ACP	0.80	0.76	6.15	98.4	98.12

Comme on peut le voir à l'aide des critères de comparaison choisis (R2, R2 ajusté, MSE, SS_Res et AICc), le modèle ACP semble être meilleur que le modèle Ridge. Avant même de calculer l'AICc, on peut constater que le R2 et le R2 ajusté sont plus élevés pour le modèle ACP, avec 80% de la variance totale expliquée par le modèle, contre 74% pour le modèle Ridge. De plus, les erreurs MSE et SS_Res sont beaucoup plus faibles pour le modèle ACP. Enfin, après avoir calculé l'AICc, on constate que le modèle ACP a le AICc le plus faible, ce qui renforce l'idée que c'est le meilleur modèle des deux.