POLYTECHNIQUE MONTRÉAL

MTH8302
Modèles de régression et d'analyse de la variance
Devoir 2
distribution : 31 mai 2018
remise : 12 juin 2018 - 23h59 (au plus tard)

Ce travail est réalisé individuellement par chaque étudiant inscrit au cours.
Chaque étudiant le fait SEUL sans demander de l'aide à d'autres.
En apposant sa signature ci-dessous, l'étudiant (e) certifie sur son honneur avoir fait ce travail SEUL.
L'obtention des résultats présentés et la rédaction de ce travail ne fait l'objet d'aucun plagiat, partiel ou total.

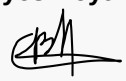
Information concernant le plagiat à Polytechnique : http://www.polymtl.ca/etudes/ppp/index.php

Exigences pour la rédaction du rapport consulter la page 4 du plan de cours

http://www.groupes.polymtl.ca/mth6301/mth8302/Autres/2018-MTH8302-PlanCours.pdf

Compléter l'information suivante et transmettez cette page comme la page 1 de votre rapport de devoir.

MTH8302 Modèles de régression et d'analyse de variance
NOM __BETTACHE_____ PRÉNOM __Lyes Heythem___

MATRICULE _1923715__ SIGNATURE 

- **Transmettre votre rapport par courriel à bernard.clement@polymtl.ca**
- **Nom suggéré pour le fichier à transmettre : NomFamille-matricule-MTH8302-Devoir2.pdf**

TABLEAU CORRECTION

	valeur	obtenu
No 5-BostonHousing	30	
No 6-BodyFat-Femme	30	
No 7-Penta	30	
Qualité	10	
TOTAL	100	

- **Les données pour la réalisation du devoir sont disponibles sur le site WEB du cours**

http://www.groupes.polymtl.ca/mth6301/MTH8302.htm/

Remarque : dans la 1er partie excersice 1, quand on a utilisé statistica pour développer la méthode *Forward Stepwise (ou Backward Stepwise), il y a 2 chemain :*
1-Advanced Models->General regression->Multiple regression->Option/Forward(Backward)
2- Multiple regression->Advanced option-> Forward(Backward)
J'ai utilisé les 2 et j'ai trouvé des résultats différent (dans ce rapport j'ai utilisé chemain 2)

No 5 Étude de modélisation avec plusieurs méthodes
Données = BostonHousing.sta

Réponse

5a) Modèle de Régression Ordinaire (MRO) (RÉGRESSION LINÉAIRE MULTIPLE)

$$Y= b0+ b1*x1+ b2*x2+ \dots + b11*x11 + e$$

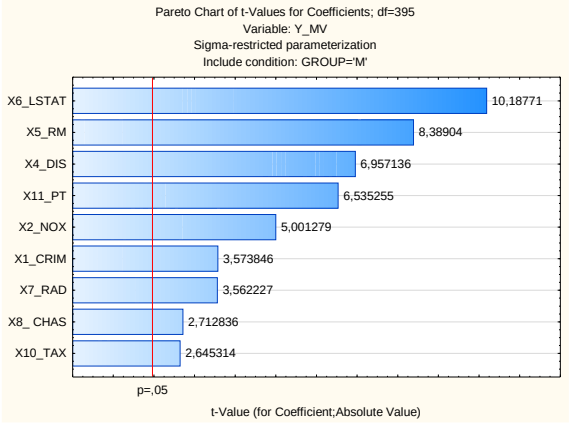
Var	Nom	coefficient	MRO ordinaire
X0	GENÉRAL intercepte	b0	41,4087
X1	CRIM	b1	-0,1251
X2	NOX	b2	-21,1251
X3	AGE	b3	0,0040
X4	DIS	b4	-1,2916
X5	RM	b5	3,9002
X6	LSTAT	b6	-0,5639
X7	RAD	b7	0,2685
X8	CHAS	b8	2,7103
X9	INDUS	b9	-0,0022
X10	TAX	b10	-0,0102
X11	PT	b11	-0,9708
		SS resid résiduelle	9826,13
		MSE = sigma^2 (ANOVA)	25,003
		R^2	0,71963493
		R^2ajusté	0,71178756

Ce tableau contient les coefficients de notre modèle Y basé sur les 11 variables X1,..., X11

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
R= ,84831299 R²= ,71963493 Adjusted R²= ,71178756
F(11,393)=91,704 p<0,0000 Std.Error of estimate: 5,0003
Include condition: GROUP='M'

N=405	b*	Std.Err. of b*	b	Std.Err. of b	t(393)	p-value
Intercept			41,4087	5,811334	7,12551	0,000000
X1_CRIM	-0,123260	0,034611	-0,1251	0,035134	-3,56129	0,000414
X2_NOX	-0,266628	0,056691	-21,1251	4,491677	-4,70317	0,000004
X3_AGE	0,012167	0,044953	0,0040	0,014906	0,27067	0,786790
X4_DIS	-0,294117	0,047968	-1,2916	0,210649	-6,13146	0,000000
X5_RM	0,295059	0,036317	3,9002	0,480050	8,12453	0,000000
X6_LSTAT	-0,432853	0,044747	-0,5639	0,058293	-9,67325	0,000000
X7_RAD	0,249104	0,072119	0,2685	0,077723	3,45408	0,000612
X8_CHAS	0,073912	0,027634	2,7103	1,013346	2,67462	0,007794
X9_NDUS	-0,001657	0,051909	-0,0022	0,070207	-0,03192	0,974552
X10_TAX	-0,183696	0,076450	-0,0102	0,004236	-2,40284	0,016731
X11_PT	-0,223687	0,034786	-0,9708	0,150978	-6,43035	0,000000

D'après les P-value on remarque que tous les variables Xi sont significatives sauf les variables X3_AGE et X9 INDUS qui ne sont pas Significatives. (Variables Significatives p-level≤ 0,05)



D'après ce graphe on a confirmé la remarque de tableau précédant

On remarque que R2 élevé (0.7196) et R2adj légèrement inférieur à R2

5b)

Les données ne présentent pas un problème de multi colinéarité

La preuve :

Pour trouver si nous avons la multicolinéarité on a utilisé les critéres de Détection multicolinéarité

Critère 1 : matrice de corrélation des variables X

Correlations (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Marked correlations are significant at p < ,05000
N=405 (Casewise deletion of missing data)
Include condition: GROUP='M'

Variable	X1_CRIM	X2_NOX	X3_AGE	X4_DIS	X5_RM	X6_LSTAT	X7_RAD	X8_ CHAS	X9_NDUS	X10_TAX	X11_PT	Y_MV
X1_CRIM	1,000000	0,405092	0,341859	-0,371778	-0,183302	0,424085	0,608001	-0,049381	0,394384	0,566047	0,289547	-0,377012
X2_NOX	0,405092	1,000000	0,721333	-0,770838	-0,289010	0,590483	0,610317	0,117650	0,761074	0,666941	0,164000	-0,421667
X3_AGE	0,341859	0,721333	1,000000	-0,728034	-0,231930	0,596188	0,458214	0,111474	0,633658	0,501316	0,243265	-0,359839
X4_DIS	-0,371778	-0,770838	-0,728034	1,000000	0,192688	-0,484251	-0,502597	-0,117004	-0,705015	-0,540198	-0,220393	0,230695
X5_RM	-0,183302	-0,289010	-0,231930	0,192688	1,000000	-0,605808	-0,174608	0,077613	-0,386300	-0,265062	-0,341690	0,685445
X6_LSTAT	0,424085	0,590483	0,596188	-0,484251	-0,605808	1,000000	0,469223	-0,042379	0,596729	0,528165	0,377576	-0,740351
X7_RAD	0,608001	0,610317	0,458214	-0,502597	-0,174608	0,469223	1,000000	0,006493	0,595796	0,909233	0,480713	-0,364852
X8_ CHAS	-0,049381	0,117650	0,111474	-0,117004	0,077613	-0,042379	0,006493	1,000000	0,081321	-0,013160	-0,109490	0,154035
X9_NDUS	0,394384	0,761074	0,633658	-0,705015	-0,386300	0,596729	0,595796	0,081321	1,000000	0,715589	0,364074	-0,468867
X10_TAX	0,566047	0,666941	0,501316	-0,540198	-0,265062	0,528165	0,909233	-0,013160	0,715589	1,000000	0,462376	-0,452231
X11_PT	0,289547	0,164000	0,243265	-0,220393	-0,341690	0,377576	0,480713	-0,109490	0,364074	0,462376	1,000000	-0,473461
Y_MV	-0,377012	-0,421667	-0,359839	0,230695	0,685445	-0,740351	-0,364852	0,154035	-0,468867	-0,452231	-0,473461	1,000000

D'après le matrice de corrilation les données ne présentent pas le problème de multi colinéarité, si on prendre rij≥ 0.95. Puisque le critére 1 pour la détection de multicolonéarité n'est pas presente (rij<0.95)

(R = (rij)matrice de corrélation des variables X critère1 ri j≥ 0.95 nécessaire mais non suffisant)

Mais on remarque pour rij≥ 0.70 il y a une corrélation entre quelque variable donc, on vérifie les deux autres critères

Critère 2 : Variance inflation factors

Collinearity statistics for terms in the equation (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization
Include condition: GROUP='M'

Effect	Tolerance	Variance Infl fac	R square	Y_MV Beta in	Y_MV Partial	Y_MV Semi-par	Y_MV t	Y_MV p
X1_CRIM	0,5955250	1,6791907	0,4044750	-0,1232604	-0,1768132	-0,0951204	-3,561295	0,0004143
X2_NOX	0,2219728	4,5050560	0,7780272	-0,2666284	-0,2308362	-0,1256193	-4,703169	0,0000036
X3_AGE	0,3630338	2,8325899	0,6369662	0,0121672	0,0136520	0,0072294	0,270666	0,7867896
X4_DIS	0,3100423	3,2253667	0,6899577	-0,2941165	-0,2954812	-0,1637683	-6,131464	0,0000000
X5_RM	0,5408926	1,8487958	0,4591074	0,2950586	0,3792172	0,2170021	8,124529	0,0000000
X6_LSTAT	0,3562825	2,8067617	0,6437175	-0,4328533	-0,4385296	-0,2583676	-9,673248	0,0000000
X7_RAD	0,1371627	7,2906102	0,8628373	0,2491040	0,1716495	0,0922569	3,454085	0,0006123
X8_ CHAS	0,9341780	1,0704598	0,0658220	0,0739117	0,1337053	0,0714377	2,674620	0,0077938
X9_NDUS	0,2647512	3,7771310	0,7352488	-0,0016569	-0,0016101	-0,0008526	-0,031920	0,9745521
X10_TAX	0,1220624	8,1925329	0,8779376	-0,1836958	-0,1203265	-0,0641786	-2,402838	0,0167312
X11_PT	0,5895473	1,6962188	0,4104527	-0,2236872	-0,3085425	-0,1717514	-6,430351	0,0000000

D'après le tableau on a max VIF =9,1925<10 donc les données ne présentent pas un problème de multicolinéarité (max VIF j≥ 10 c-à-d R2j≥ 0,90)

Critère 3 : Indice Conditionnement

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %	IC
1	5,462799	49,66181	5,46280	49,6618	1
2	1,372640	12,47854	6,83544	62,1404	3,97977587
3	1,129551	10,26864	7,96499	72,4090	4,83625771
4	0,845876	7,68978	8,81087	80,0988	6,45815553
5	0,655850	5,96228	9,46672	86,0611	8,32933748
6	0,500154	4,54685	9,96687	90,6079	10,9222384
7	0,348620	3,16927	10,31549	93,7772	15,669793
8	0,240779	2,18890	10,55627	95,9661	22,6880139
9	0,211233	1,92030	10,76750	97,8864	25,8615174
10	0,163825	1,48932	10,93133	99,3757	33,3452692
11	0,068673	0,62430	11,00000	100,0000	79,5475137

On a vérifié aussi le Critère 3 ou on a trouvé que les données ne présentent pas un problème de multicolinéarité (IC = λ1/λk < 100 k = 2, 3, ...)

D'après les 3 critéres les données ne présentent pas le problème de multi colinéarité.

5c)
Modèle de Régression avec Sélection pas à pas Avant (Forward Stepwise) (MRF)

Y= b0+ b1*x1+ b2*x2+b4*x4 … +b8*x8+b10*x10+ b11*x11 + e

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
R= ,84828194 R²= ,71958226 Adjusted R²= ,71319299
F(9,395)=112,62 p<0,0000 Std.Error of estimate: 4,9881
Include condition: GROUP='M'

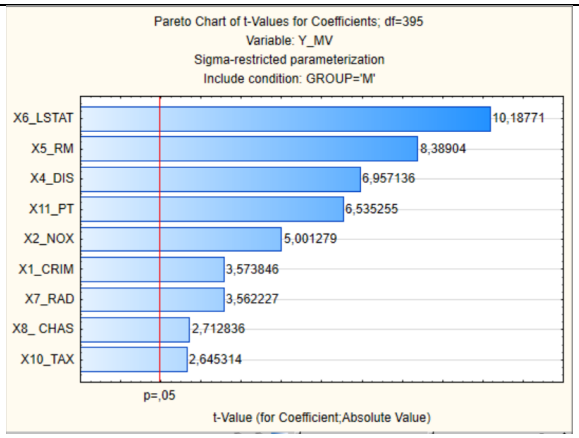
N=405	b*	Std.Err. of b*	b	Std.Err. of b	t(395)	p-value
Intercept			41,3312	5,781619	7,1487	0,000000
X6_LSTAT	-0,428992	0,042109	-0,5589	0,054856	-10,1877	0,000000
X5_RM	0,296865	0,035387	3,9241	0,467760	8,3890	0,000000
X11_PT	-0,222863	0,034102	-0,9673	0,148007	-6,5353	0,000000
X4_DIS	-0,298273	0,042873	-1,3098	0,188273	-6,9571	0,000000
X2_NOX	-0,263239	0,052634	-20,8566	4,170248	-5,0013	0,000001
X8_ CHAS	0,074347	0,027406	2,7263	1,004961	2,7128	0,006963
X1_CRIM	-0,123308	0,034503	-0,1252	0,035024	-3,5738	0,000395
X7_RAD	0,248692	0,069814	0,2680	0,075239	3,5622	0,000413
X10_TAX	-0,184822	0,069868	-0,0102	0,003871	-2,6453	0,008487

Summary of stepwise regression; variable: Y_MV (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Forward stepwise
P to enter: ,05, P to remove: ,05
Include condition: GROUP='M'

Effect	Steps	Degr. of Freedom	F to remove	P to remove	F to enter	P to enter	Effect status
X1_CRIM	Step Number 1	1			66,7724	0,000000	Out
X2_NOX		1			87,1500	0,000000	Out
X3_AGE		1			59,9438	0,000000	Out
X4_DIS		1			22,6534	0,000003	Out
X5_RM		1			357,1414	0,000000	Out
X6_LSTAT		1			488,8275	0,000000	Entered
X7_RAD		1			61,8841	0,000000	Out
X8_ CHAS		1			9,7942	0,001878	Out
X9_NDUS		1			113,5585	0,000000	Out
X10_TAX		1			103,6076	0,000000	Out
X11_PT		1			116,4407	0,000000	Out

…

Effect	Steps	Degr. of Freedom	F to remove	P to remove	F to enter	P to enter	Effect status
X3_AGE		1			0,0833	0,773004	Out
X6_LSTAT	Step Number 10	1	103,7894	0,000000			In
X5_RM		1	70,3760	0,000000			In
X11_PT		1	42,7096	0,000000			In
X4_DIS		1	48,4017	0,000000			In
X2_NOX		1	25,0128	0,000001			In
X8_ CHAS		1	7,3595	0,006963			In
X1_CRIM		1	12,7724	0,000395			In
X7_RAD		1	12,6895	0,000413			In
X10_TAX		1	6,9977	0,008487			In
X9_NDUS		1			0,0006	0,980928	Out
X3_AGE		1			0,0730	0,787163	Out



On remarque que la méthode de Régression avec Forward Stepwise a éliminé (négligé) les variables X3_AGE et X9 INDUS qui ne sont pas significatives.
Forward Stepwise ne vérifie pas l'états de toutes les variables après la sélection chaque step.

d)
Modèle de Régression avec Sélection pas à pas Arrière (Backward Stepwise) (MRB)

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
R= ,83553061 R²= ,69811140 Adjusted R²= ,69432833
F(5,399)=184,54 p<0,0000 Std.Error of estimate: 5,1495
Include condition: GROUP='M'

N=405	b*	Std.Err. of b*	b	Std.Err. of b	t(399)	p-value
Intercept			36,8479	5,268551	6,9939	0,000000
X2_NOX	-0,246248	0,047446	-19,5103	3,759152	-5,1901	0,000000
X4_DIS	-0,291856	0,043995	-1,2817	0,193201	-6,6338	0,000000
X5_RM	0,316397	0,035411	4,1822	0,468074	8,9350	0,000000
X6_LSTAT	-0,463725	0,042430	-0,6041	0,055274	-10,9291	0,000000
X11_PT	-0,214198	0,030493	-0,9297	0,132344	-7,0246	0,000000

Analysis of Variance; DV: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
Include condition: GROUP='M'

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	24467,15	5	4893,429	184,5359	0,00
Residual	10580,48	399	26,517		
Total	35047,62				

On remarque que la méthode de Régression avec Backward Stepwise a éliminé les variables X1_CRIM, X3_AGE ,X7_RAD ,X8_CHAS ,X9 INDUS et X10_TAX qui ne sont pas significatives

Tableau 5d : synthèse des modèles

Var	Nom	coefficient	MRO ordinaire	MRF sélection avant	MRB sélection arrière
X0	GENÉRAL intercepte	b0	41,4087	41,3312	36,8479
X1	CRIM	b1	-0,1251	-0,1252	
X2	NOX	b2	-21,1251	-20,8566	-19,5103
X3	AGE	b3	0,0040		
X4	DIS	b4	-1,2916	-1,3098	-1,2817
X5	RM	b5	3,9002	3,9241	4,1822
X6	LSTAT	b6	-0,5639	-0,5589	-0,6041
X7	RAD	b7	0,2685	0,2680	
X8	CHAS	b8	2,7103	2,7263	
X9	INDUS	b9	-0,0022		
X10	TAX	b10	-0,0102	-0,0102	
X11	PT	b11	-0,9708	-0,9673	-0,9297
		SS resid résiduelle	9826,13	9827,976	10580,48
		MSE = sigma^2 (ANOVA)	25,003	24,88095	26,517
		R^2	0,71963493	0,719582	0,69811140
		R^2 ajusté	,71178756	0,713193	0,69432833

5e)
Comparez les prédictions des 3 modèles sur l'ensemble test T constitué des 101 observations. Choisir le meilleur modèle selon des critères; préciser la nature de ces critères.

Var	Nom	coefficient	MRO ordinaire	MRF sélection avant	MRB sélection arrière
X0	GENÉRAL intercepte	b0	37,30622	33,70878	27,50599
X1	CRIM	b1	0,15954	0,12549	
X2	NOX	b2	-4,99851		
X3	AGE	b3	-0,03647	0,028814	
X4	DIS	b4	-0,92637	0,393779	
X5	RM	b5	4,16260	4,23967	4,35412
X6	LSTAT	b6	-0,50708	-0,49156	-0,45228
X7	RAD	b7	0,05491		
X8	CHAS	b8	3,83516	4,09416	
X9	INDUS	b9	-0,04094	0,100942	
X10	TAX	b10	-0,00483		
X11	PT	b11	-1,35093	-1,32519	-1,45293
		SS resid résiduelle	1508,168	1522,431	1769,708
		MSE = sigma^2 (ANOVA)	16,94570	1654,82	18,244
		R^2	0,802669	0, 80080241	0,76844831
		R^2 ajusté	0,778279	0, 78348088	0, 76128691
		F	32,910728	46,232	107,30

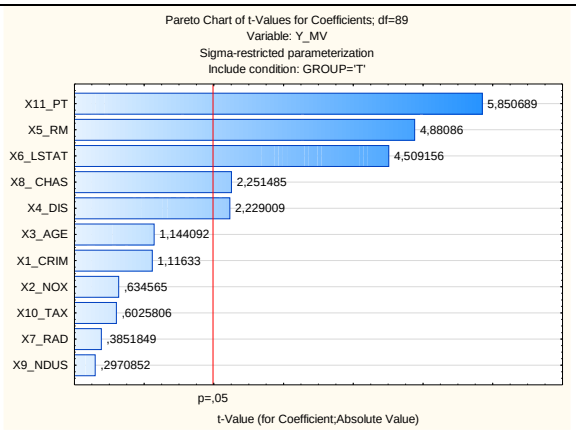
MQR

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
R= ,89591780 R²= ,80266870 Adjusted R²= ,77827944
F(11,89)=32,911 p<0,0000 Std.Error of estimate: 4,1165
Include condition: GROUP='T'

N=101	b*	Std.Err. of b*	b	Std.Err. of b	t(89)	p-value
Intercept			37,30622	9,758018	3,82314	0,000244
X1_CRIM	0,105387	0,094405	0,15954	0,142915	1,11633	0,267286
X2_NOX	-0,062573	0,098607	-4,99851	7,877061	-0,63456	0,527340
X3_AGE	-0,118951	0,103970	-0,03647	0,031879	-1,14409	0,255653
X4_DIS	-0,217179	0,097433	-0,92637	0,415595	-2,22901	0,028331
X5_RM	0,332211	0,068064	4,16260	0,852841	4,88086	0,000005
X6_LSTAT	-0,414082	0,091831	-0,50708	0,112456	-4,50916	0,000020
X7_RAD	0,056561	0,146842	0,05491	0,142562	0,38518	0,701019
X8_CHAS	0,111972	0,049732	3,83516	1,703389	2,25149	0,026816
X9_NDUS	-0,031720	0,106772	-0,04094	0,137812	-0,29709	0,767094
X10_TAX	-0,094350	0,156577	-0,00483	0,008011	-0,60258	0,548320
X11_PT	-0,346350	0,059198	-1,35093	0,230901	-5,85069	0,000000

Variables currently in the Equation; DV: Y_MV (BostonHousing.sta in Workbook1_(Recovered
Include condition: GROUP='T'

Variable	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(89)	p-value
X1_CRIM	0,105387	0,117511	0,052565	0,248781	0,751219	1,11633	0,267286
X2_NOX	-0,062573	-0,067112	-0,029880	0,228027	0,771973	-0,63456	0,527340
X3_AGE	-0,118951	-0,120391	-0,053872	0,205112	0,794888	-1,14409	0,255653
X4_DIS	-0,217179	-0,229943	-0,104958	0,233557	0,766443	-2,22901	0,028331
X5_RM	0,332211	0,459513	0,229826	0,478597	0,521403	4,88086	0,000005
X6_LSTAT	-0,414082	-0,431242	-0,212324	0,262920	0,737080	-4,50916	0,000020
X7_RAD	0,056561	0,040796	0,018137	0,102826	0,897174	0,38518	0,701019
X8_CHAS	0,111972	0,232138	0,106016	0,896453	0,103547	2,25149	0,026816
X9_NDUS	-0,031720	-0,031475	-0,013989	0,194487	0,805513	-0,29709	0,767094
X10_TAX	-0,094350	-0,063744	-0,028374	0,090437	0,909563	-0,60258	0,548320
X11_PT	-0,346350	-0,527045	-0,275493	0,632689	0,367311	-5,85069	0,000000



Collinearity statistics for terms in the equation (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization
Include condition: GROUP='T'

Effect	Tolernce	Variance Infl fac	R square	Y_MV Beta in	Y_MV Partial	Y_MV Semi-par	Y_MV t	Y_MV p
X1_CRIM	0,2487806	4,019606	0,7512194	0,1053870	0,1175109	0,0525649	1,116330	0,2672859
X2_NOX	0,2280270	4,385446	0,7719730	-0,0625729	-0,0671121	-0,0298799	-0,634565	0,5273400
X3_AGE	0,2051118	4,875389	0,7948882	-0,1189511	-0,1203914	-0,0538721	-1,144092	0,2556532
X4_DIS	0,2335566	4,281617	0,7664434	-0,2171794	-0,2299432	-0,1049578	-2,229009	0,0283314
X5_RM	0,4785972	2,089440	0,5214028	0,3322112	0,4595131	0,2298261	4,880860	0,0000046
X6_LSTAT	0,2629196	3,803444	0,7370804	-0,4140824	-0,4312416	-0,2123236	-4,509156	0,0000198
X7_RAD	0,1028263	9,725135	0,8971737	0,0565614	0,0407955	0,0181373	0,385185	0,7010194
X8_CHAS	0,8964528	1,115508	0,1035472	0,1119717	0,2321376	0,1060162	2,251485	0,0268164
X9_NDUS	0,1944867	5,141739	0,8055133	-0,0317204	-0,0314754	-0,0139889	-0,297085	0,7670937
X10_TAX	0,0904374	11,057376	0,9095626	-0,0943505	-0,0637435	-0,0283738	-0,602581	0,5483200
X11_PT	0,6326885	1,580557	0,3673115	-0,3463499	-0,5270451	-0,2754926	-5,850689	0,0000001

D'après le graphe et p-valeu on remarque qu'il y a des variables qui sont non significatifs, et on a aussi max VIF=11,05 donc on a le probléme de multicolinéarité. Le modèle ne satisfaisant pas

MRF

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
R= ,89487564 R²= ,80080241 Adjusted R²= ,78348088
F(8,92)=46,232 p<0,0000 Std.Error of estimate: 4,0679
Include condition: GROUP='T'

N=101	b*	Std.Err. of b*	b	Std.Err. of b	t(92)	p-value
Intercept			33,70878	7,932084	4,24968	0,000051
X5_RM	0,338362	0,065560	4,23967	0,821465	5,16111	0,000001
X11_PT	-0,339752	0,055712	-1,32519	0,217304	-6,09833	0,000000
X6_LSTAT	-0,401407	0,089531	-0,49156	0,109639	-4,48342	0,000021
X8_CHAS	0,119534	0,047801	4,09416	1,637227	2,50067	0,014166
X1_CRIM	0,082898	0,066917	0,12549	0,101302	1,23882	0,218563
X4_DIS	-0,223459	0,092319	-0,95315	0,393779	-2,42052	0,017463
X3_AGE	-0,151513	0,093971	-0,04646	0,028814	-1,61233	0,110315
X9_NDUS	-0,095175	0,078206	-0,12284	0,100942	-1,21697	0,226729

Analysis of Variance; DV: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
Include condition: GROUP='T'

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	6120,389	8	765,0486	46,23162	0,000000
Residual	1522,431	92	16,5482		
Total	7642,820				

D'après les tableaux on remarque que la méthode de Régression avec Forward Stepwise a éliminé (négligé) quelques variables qui ne sont pas significatives mais il n'a pas éliminé toutes les variables puisque la méthode de Forward Stepwise ne vérifie pas l'états de toutes les variables après la sélection chaque step.

MRB

Regression Summary for Dependent Variable: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
R= ,87661183 R²= ,76844831 Adjusted R²= ,76128691
F(3,97)=107,30 p<0,0000 Std.Error of estimate: 4,2713
Include condition: GROUP='T'

N=101	b*	Std.Err. of b*	b	Std.Err. of b	t(97)	p-value
Intercept			27,50599	7,645390	3,59772	0,000507
X5_RM	0,347496	0,066119	4,35412	0,828469	5,25562	0,000001
X6_LSTAT	-0,369334	0,064799	-0,45228	0,079352	-5,69966	0,000000
X11_PT	-0,372500	0,054104	-1,45293	0,211030	-6,88492	0,000000

Analysis of Variance; DV: Y_MV (BostonHousing.sta in Workbook1_(Recovered))
Include condition: GROUP='T'

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	5873,112	3	1957,704	107,3043	0,000000
Residual	1769,708	97	18,244		
Total	7642,820				

D'après les tableaux on remarque que la méthode de Régression avec Backward Stepwise a éliminé les variables qui ne sont pas significatives et résoudre le probléme de multicolinéarité.

Backward Stepwise vérifie l'états de toutes les variables après la sélection pour chaque step

D’après les résultats precedantes on remarque que la méthode de Régression avec Backward Stepwise plus rapide et éfficase et contien moins des étapes par rapport Régression avec Forward Stepwisea et Régression Ordinaire.
Et on remarque que avec la méthode Backward Stepwise on a éliminé le probléme des variables non significatifs qui aparaitre dans le modèle de Régression Ordinaire et Régression avec Forward Stepwisea, et aussi on remarque qu’il a éliminé le probléme de multicolinéarité, Et aussi R2adj légèrement inférieur à R2.

No 6 Étude d'un modèle de régression multiple problématique

Données = BodyFat-F.sta

Réponse

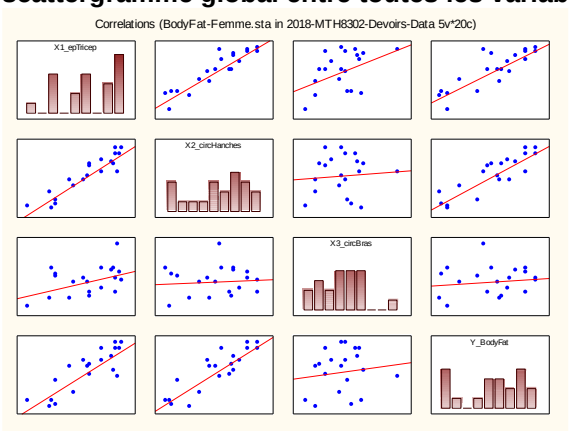
6a)

matrice de corrélation

Correlations (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Marked correlations are significant at p < ,05000
N=20 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	X1_epTricep	X2_circHanches	X3_circBras	Y_BodyFat
X1_epTricep	25,30500	5,023259	1,000000	0,923843	0,457777	0,843265
X2_circHanches	51,17000	5,234612	0,923843	1,000000	0,084667	0,878090
X3_circBras	27,62000	3,647147	0,457777	0,084667	1,000000	0,142444
Y_BodyFat	20,19500	5,106186	0,843265	0,878090	0,142444	1,000000

scattergramme global entre toutes les variables



D'après la matrice de corrélation et le scattergramme on remarque que La corrélation est positive entre les variables et est valide
on remarque qu'il y a une forte corrélation entre (X2 et X1 ,…) ce qui implique qu'il y a un problème de multicolinéarité

6b)
Modèle de régression multiple ordinaire (MRO) entre Y et X1, X2, X3.

Y= b0+ b1*X1+ b2*X2+b3*X3

Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
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

On remarque que le signe de coefficient de variable X1_epTricep est positive cela implique que X1_epTricep évalué dans le même sens que Y_bodyFat (si X1 augmente donc Y augmente).
Et le signe des coefficients des variables X2_circHanches, X3_circBras est négative, implique que les variables évalués dans le sens inverse de Y_bodyFat (si X2 augmente donc Y diminué).
Et aussi on remarque que la variable X1_epTricep a un plus grand impact sur Y_bodyFat par unité par rapport les autres variables.

Analysis of Variance; DV: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)

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				

Collinearity statistics for terms in the equation (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization

Effect	Tolerce	Variance Infl fac	R square	Y_BodyFat Beta in	Y_BodyFat Partial	Y_BodyFat Semi-par	Y_BodyFat t	Y_BodyFat p
X1_epTricep	0,0014107	708,84291	0,9985893	4,263705	0,3381500	0,1601445	1,437266	0,1699111
X2_circHanches	0,0017720	564,34335	0,9982280	-2,928701	-0,2665991	-0,1232830	-1,106441	0,2848944
X3_circBras	0,0095597	104,60691	0,9904403	-1,561417	-0,3240520	-0,1526654	-1,370142	0,1895628

On remarque aussi qu'il y a aucun coefficient significatif (p-value >0,05)

b0	**117,0847**
b1	**4,3341**
b2	**-2,8568**
b3	**-2,1861**

Le modèle ne satisfaisant pas à cause de présentent un problème de multicolinéarité entre les variable Xi (max VIF=708,84 >10)

Pour obtenir un modèle plus satisfaisant on utilise des méthodes pour remédier aux problèmes de multicolinéarité comme la régression en composantes principales (ACP), la régression PLS (« Partial Least Square »), la régression RIDGE, la méthode de sélection de variables…

Et pour cet exercice on utilise les 2 méthodes suivant pour résoudre le problème de multicolinéarité : Mod1 : régression RIDGE et Mod2 : régression en composantes principales

6c)
Mod1 : régression RIDGE

K=0

Ridge Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
I=0,0000 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

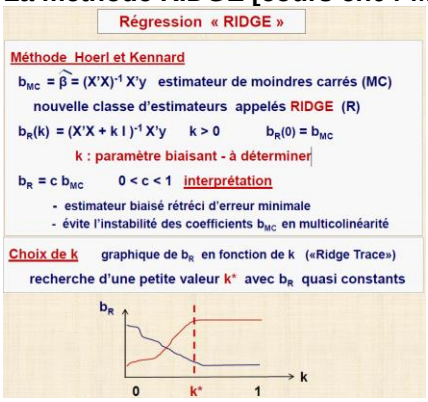
K=0,1

Ridge Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
I=0,0000 R= ,89518632 R²= ,80135855 Adjusted R²= ,76411328
F(3,16)=21,516 p<,00001 Std.Error of estimate: 2,4800

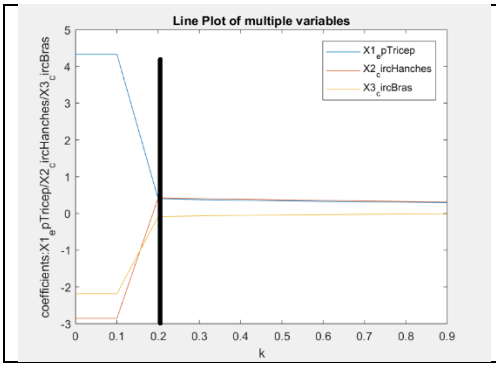
N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			117,0847	99,78240	1,17340	0,257808
X1_epTricep	4,26370	2,966538	4,3341	3,01551	1,43727	0,169911
X2_circHanches	-2,92870	2,646956	-2,8568	2,58202	-1,10644	0,284894
X3_circBras	-1,56142	1,139602	-2,1861	1,59550	-1,37014	0,189563

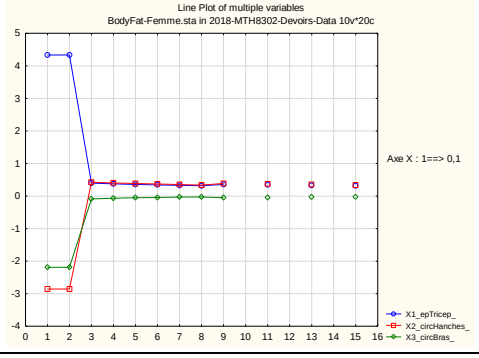
…

La méthode RIDGE [cours ch04-Multiple2 p21]



k	X1	X2	X3
0,0	4,3341	-2,8568	-2,1861
0,1	4,3341	-2,8568	-2,1861
0,2	0,39789	0,42405	-0,08581
0,3	0,37644	0,40521	-0,06704
0,4	0,35877	0,38685	-0,05270
0,5	0,34330	0,36976	-0,04130
0,6	0,32939	0,35401	-0,03205
0,7	0,31673	0,33953	-0,02443





D'après le graphe on a prend K=0,24

Y_BodyFat_pred = b0 + b1*X1_epTricep_cr +b1*X2_cirHanches + b1*X3_circBras

Ridge Regression Summary for Dependent Variable: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
l=,25000 R= ,82838405 R²= ,68622013 Adjusted R²= ,62738640
F(3,16)=11,664 p<,00027 Std.Error of estimate: 3,1169

N=20	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			-8,71372	9,023992	-0,965617	0,348606
X1_epTricep	0,380223	0,210714	0,38650	0,214193	1,804446	0,090017
X2_circHanches	0,425122	0,196527	0,41469	0,191705	2,163175	0,046009
X3_circBras	-0,054086	0,142269	-0,07572	0,199184	-0,380166	0,708821

Analysis of Variance; DV: Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Ridge regression, lambda=,2500000

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	339,9462	3	113,3154	11,66372	0,000266
Residual	155,4433	16	9,7152		
Total	495,3895				

R^2	0,686220
R^2ajusté	0,627386

On remarque que la régression RIDGE résoudre le problème de signification avec la variable X2_circHanches mais les variables X1_epTricep et X3_circBras sont non significatifs
Et on remarque aussi qu'il y a 45,72% de variabilité de variable qui n'est pas expliqué par notre modèle

6c)

Mod2 : régression en composantes principales

Au début on utilise la méthode ACP pour exprime les données initiales X dans l'espace des facteurs (F1, F2,..), on exprime les variables initiales X sous forme centrées-réduites Xcr
On criée un fichier initial qui contient les données dans
-l'espace initial (X) -l'espace des facteurs (F) -l'espace des variables centrés-réduites (Xcr)

12 X1_epTricep_cr	13 X2_circHanches_cr	14 X3_circBras_cr	15 Facteur1	16 Facteur2	17 Facteur3
-1,2	-1,5	0,4	-1,63189	1,10075	0,046262
-0,1	-0,3	0,2	-0,19303	0,26387	0,037463
1,1	0,1	2,6	1,72932	2,19005	-0,024545
0,9	0,6	1,0	1,33021	0,54703	-0,002570
-1,2	-1,7	0,9	-1,62359	1,62286	-0,036284
0,1	0,5	-1,1	-0,00515	-1,19609	0,003308
1,2	1,4	-0,0	1,72239	-0,68255	-0,024221
0,5	0,2	0,8	0,75518	0,62824	0,032686
-0,6	-0,2	-1,2	-1,01789	-0,94745	0,030129
0,0	0,4	-0,8	0,03793	-0,89109	-0,044847
1,2	1,0	0,7	1,68155	0,07017	-0,015348
1,0	1,1	0,2	1,43448	-0,34907	0,000371
-1,3	-0,9	-1,3	-1,91602	-0,67657	-0,024716
-1,1	-1,3	0,3	-1,51969	0,88331	-0,022057
-2,1	-1,6	-1,7	-3,10224	-0,73369	-0,017750
0,8	0,6	0,7	1,20528	0,29584	0,017606
0,5	0,8	-0,5	0,64454	-0,84331	-0,018437
1,0	1,4	-0,8	1,28207	-1,41630	0,017912
-0,5	-0,6	-0,1	-0,76702	0,14812	0,030217
-0,0	-0,0	-0,0	-0,04642	-0,01414	0,014821

Factor coordinates of the variables, based on correlations (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Active and Supplementary variables
*Supplementary variable

Variable	Factor 1	Factor 2	Factor 3
X1_epTricep	0,998641	-0,048393	0,019342
X2_circHanches	0,904817	-0,425451	-0,017255
X3_circBras	0,500494	0,865708	-0,007399
*Y_BodyFat	0,826492	-0,312046	0,144559

Ensuite on régresse Y sur les facteurs F avec la méthode progressive
Régresser Y sur F1, régresser Y sur F1 et F2,…, etc
Arrêter lorsqu'un nouveau F devient non significatif.

Régresser Y sur F1

Parameter Estimates (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization

Effect	Y_BodyFat Std.Err	Y_BodyFat t	Y_BodyFat p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Y_BodyFat Beta (ß)	Y_BodyFat St.Err.ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	0,660376	30,58107	0,000000	18,80760	21,58240				
Facteur1	0,471318	6,22882	0,000007	1,94555	3,92596	0,826492	0,132688	0,547724	1,105260

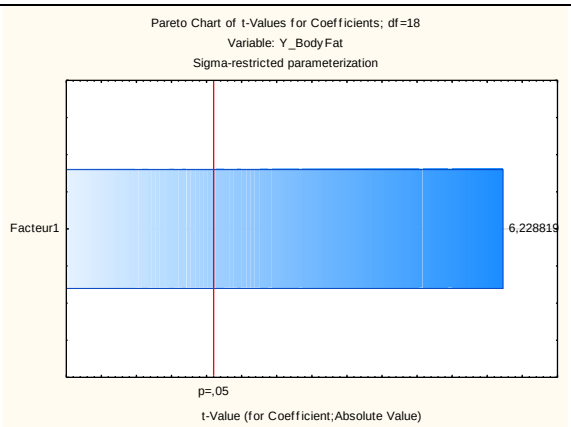
Régresser Y sur F1 et F2

Parameter Estimates (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization

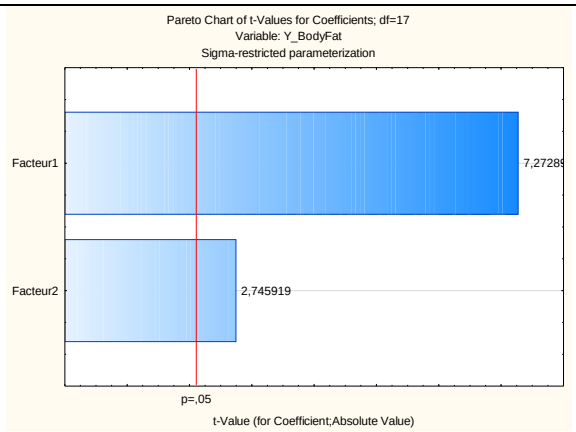
Effect	Y_BodyFat Param.	Y_BodyFat Std.Err	Y_BodyFat t	Y_BodyFat p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Y_BodyFat Beta (ß)	Y_BodyFat St.Err.ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	20,19500	0,565574	35,70708	0,000000	19,00174	21,38826				
Facteur1	2,93576	0,403657	7,27290	0,000001	2,08412	3,78740	0,826492	0,113640	0,586732	1,066251
Facteur2	-1,64976	0,600805	-2,74592	0,013785	-2,91735	-0,38217	-0,312046	0,113640	-0,551806	-0,072287

Univariate Tests of Significance for Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 2,5293

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	8156,760	1	8156,760	1274,996	0,000000
Facteur1	338,395	1	338,395	52,895	0,000001
Facteur2	48,237	1	48,237	7,540	0,013785
Error	108,757	17	6,397		



D'après le graphe et p-valeu on remarque que le variable Facteur1 est significatif



D'après le graphe et p-valeu on remarque que les variables Facteur1 et Facteur2 sont significatifs

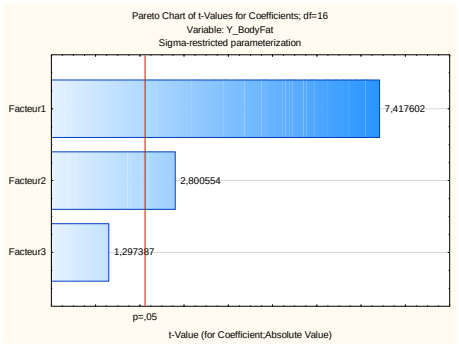
Régresser Y sur F1 et F2 et F3

Parameter Estimates (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization

Effect	Y_BodyFat Param.	Y_BodyFat Std.Err	Y_BodyFat t	Y_BodyFat p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	Y_BodyFat Beta (ß)	Y_BodyFat St.Err.ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept	20,19680	0,55454	36,41753	0,000000	19,0194	21,37367				
Facteur1	2,93576	0,39578	7,41760	0,000001	2,0967	3,77478	0,826492	0,111423	0,590285	1,062698
Facteur2	-1,64976	0,58908	-2,80055	0,012827	-2,8986	-0,40096	-0,312046	0,111423	-0,548252	-0,075840
Facteur3	27,38341	21,10659	1,29739	0,212894	-17,3606	72,12738	0,144559	0,111423	-0,091647	0,380765

Univariate Tests of Significance for Y_BodyFat (BodyFat-Femme.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 2,4800

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	8156,761	1	8156,761	1326,237	0,000000
Facteur1	338,395	1	338,395	55,021	0,000001
Facteur2	48,237	1	48,237	7,843	0,012827
Facteur3	10,352	1	10,352	1,683	0,212894
Error	98,405	16	6,150		



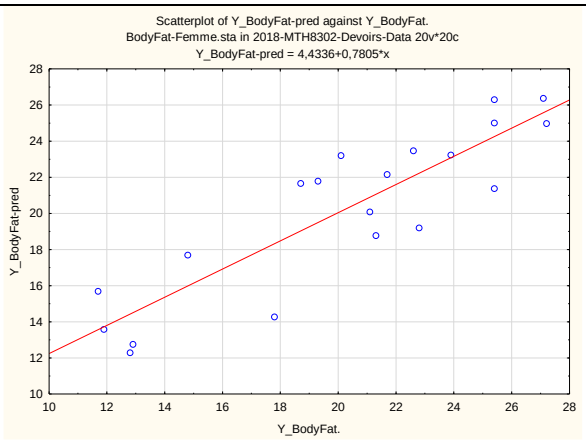
D'après le graphe et p-valeu on remarque que le variable Facteur3 est non significatif

Et après on fait la Comparaissant entre Yprédit et Yobservé et on modifie l'étape de régression et prend le meilleur choix.

Et a la fin nous avons trouvé que Y = b0 + b1*F1 + b2*F2 c'est le meilleur choix

Pour régresser Y sur F1 et F2
Nous avons trouvé
Y = b0 + b1*F1 + b2*F2

Var	Coefficient	MR
F0	b0	20,19500
F1	b1	2,93576
F2	b2	-1,64976
	SS resid résiduelle	108,7572 28,12%
	MSE = sigma^2 (ANOVA)	6,397480
	R^2	0,780461
	R^2ajusté	0,754633



D'après le graphe et le tableau on reamrque que on a une bonne corrélation entre les facteurs et Y_pred (R^2ajusté=0,75). Et aussi qu'il y a 28% de variabilité de variable qui n'est pas expliqué par notre modèle

Variable	Factor 1	Factor 2	Factor 3
X1_epTricep	0,998641	-0,048393	0,019342
X2_circHanches	0,904817	-0,425451	-0,017255
X3_circBras	0,500494	0,865708	-0,007399
*Y_BodyFat	0,826492	-0,312046	0,144559

Y_BodyFat_pred = b0 + b1*F1 + b2*F2=b0+(b1*0,998-b2*0,048)*X1_epTricep_cr +
(b1*0,904-b2*0,425)*X2_cirHanches + (b1*0,5-b2*0,312)*X3_circBras

6d)

	Mod1	Mod2
R^2	0,686220	0,780461
R^2ajusté	0,627386	0,754633
SSresid(%)	45,72	28,12

D'après la valeure de R^2ajusté et SSresid(%) on remarque que le Mod2 (régression en composantes principales) le meilleur choix
Et on remarque aussi qu'il y a des variables non significatives dans le Mod1 (régression RIDGE) par contre le Mod2 qui toutes ces variables sont significatives.

No 7 <u>Étude de prédiction d'activité biologique : modélisation PLS</u>

Données = Penta.sta

Réponse

7a)

M1 : Modèle PLS (modèle avec toutes les composantes)

Predictor weights (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"
Compo 1	-0,157641	0,085681	-0,169313	0,121527	0,071134	0,065188	-0,424809	0,653701	0,285046	-0,293407
Compo 2	-0,193810	0,021349	-0,404829	0,229796	0,071941	0,250337	-0,108375	0,325934	0,425376	0,414880
Compo 3	0,130046	-0,113962	0,001702	-0,125617	-0,246332	0,283583	-0,340124	0,286179	-0,092797	0,081971
Compo 4	0,118332	0,025651	0,263189	-0,108027	-0,094759	0,006084	-0,288687	0,372122	-0,179677	0,161053
Compo 5	-0,045868	0,228215	0,310581	0,008716	0,204099	-0,399491	0,087050	0,538996	-0,199468	0,159731
Compo 6	0,061752	0,080174	0,209562	-0,288848	-0,115836	-0,263064	0,331086	0,719173	-0,212239	0,191910
Compo 7	0,110030	-0,026349	0,087751	0,049820	0,035832	0,013176	0,359529	0,696413	-0,556903	0,145516
Compo 8	-0,108092	0,122918	-0,069392	0,000680	-0,042247	0,086976	0,501348	0,605843	-0,529243	0,035130
Compo 9	0,106628	-0,146492	-0,178732	0,057439	0,059039	-0,020814	0,615097	0,559399	-0,282888	-0,062998
Compo 10	0,312921	0,378214	0,489695	0,231033	-0,035346	0,470374	0,322718	0,073336	0,351522	-0,064007
Compo 11	0,679530	0,665916	-0,307054	0,004732	0,001234	0,005658	-0,003764	-0,017029	0,005456	-0,006089

"S5"	"L5"	"P5"
0,056905	0,056905	-0,056905
-0,008193	-0,008193	0,008193
-0,299076	-0,299076	0,299076
0,156438	0,156438	-0,156438
-0,115688	-0,115688	0,115688
-0,048718	-0,048718	0,048718
-0,014484	-0,014484	0,014484
0,009015	0,009015	-0,009015
0,021920	0,021920	-0,021920
-0,017691	-0,017691	0,017691
0,001854	0,001854	-0,001854

PLS regression coefficients (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Interc.	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"
Y_logRAI	0,797598	0,126345	0,268079	-0,442360	0,103932	0,058552	0,060840	-0,049595	0,362006	0,117779	-0,031100

"L4"	"P4"	"S5"	"L5"	"P5"
-0,057821	-0,328174	0,011786	0,090855	-0,008913

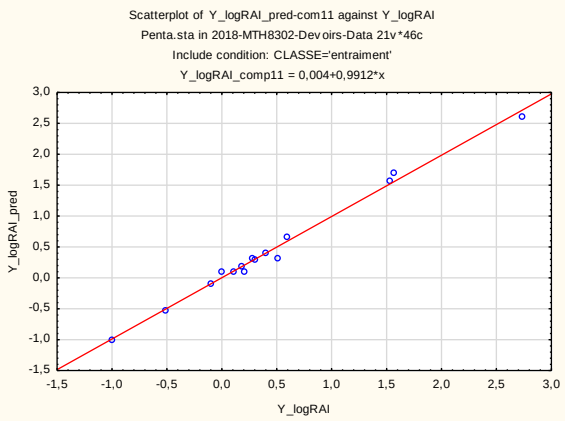
Y_logRAI-pred=b0+b1*S1+…b15*P5

Summary of PLS (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Increase R² of Y	Average R² of Y	Increase R² of X	Average R² of X
Comp 1	0,896399	0,896399	0,169014	0,169014
Comp 2	0,078368	0,974767	0,127721	0,296735
Comp 3	0,004636	0,979403	0,146554	0,443289
Comp 4	0,002485	0,981889	0,118421	0,561710
Comp 5	0,001494	0,983383	0,105894	0,667605
Comp 6	0,002617	0,986001	0,051876	0,719481
Comp 7	0,002428	0,988428	0,061873	0,781354
Comp 8	0,001926	0,990354	0,072252	0,853606
Comp 9	0,000725	0,991080	0,067285	0,920891
Comp 10	0,000000	0,991080	0,079076	0,999967
Comp 11	0,000099	0,991179	0,000033	1,000000

D'après les R2 de de Y(cumul) on remarque que nous avons une bonne corrélation entres les composantes et Y

Y_logRAI-OBSV	Y_logRAI-pred-com11	écart
0,00	0,10001	-0,10001
0,28	0,32348	-0,04348
0,20	0,09999	0,10001
0,51	0,32345	0,18655
0,11	0,11000	0,00000
2,73	2,60227	0,12773
0,18	0,18000	0,00000
1,53	1,58613	-0,05613
-0,10	-0,10000	0,00000
-0,52	-0,52000	0,00000
0,40	0,40000	0,00000
0,30	0,30000	0,00000
-1,00	-1,00000	0,00000
1,57	1,71053	-0,14053
0,59	0,66413	-0,07413



On remarque que le modèle prédit a bien suivi notre modèle réel

7b)
modèle à 2 composantes semble un bon choix R2 de Y = 0,9747

Summary of PLS (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Increase R² of Y	Average R² of Y	Increase R² of X	Average R² of X
Comp 1	0,896399	0,896399	0,169014	0,169014
Comp 2	0,078368	0,974767	0,127721	0,296735
Comp 3	0,004636	0,979403	0,146554	0,443289
Comp 4	0,002485	0,981889	0,118421	0,561710
Comp 5	0,001494	0,983383	0,105894	0,667605
Comp 6	0,002617	0,986001	0,051876	0,719481
Comp 7	0,002428	0,988428	0,061873	0,781354
Comp 8	0,001926	0,990354	0,072252	0,853606
Comp 9	0,000725	0,991080	0,067285	0,920891
Comp 10	0,000000	0,991080	0,079076	0,999967
Comp 11	0,000099	0,991179	0,000033	1,000000

On a basé sur le choix des composantes sur les deux critères suivants :
Nombre de composante petit et R2 de Y (cumul) max
Et comme vous voyez dans le tableau nous remarquons que R2 de Y(cumul) de la composante2 augmente de 0,07 par rapport R2 de Y(cumul) de composante1, et après la composante2 on remarque que les autres composantes augmentent de façon presque constante.
avec R2 de Y (cumul)=0,97 qui montre que notre modèle presque prédit tous les variables

M2 : Modèle PLS (modèle avec les 2 premières composantes)

Predictor weights (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"
Compo 1	-0,157641	0,085681	-0,169313	0,121527	0,071134	0,065188	-0,424809	0,653701	0,285046
Compo 2	-0,193810	0,021349	-0,404829	0,229796	0,071941	0,250337	-0,108375	0,325934	0,425376

"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
-0,293407	0,298287	-0,203095	0,056905	0,056905	-0,056905
0,414880	-0,283113	0,355025	-0,008193	-0,008193	0,008193

PLS regression coefficients (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Interc.	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"
Y_logRAI	-0,437289	-0,083797	0,039170	-0,296360	0,145064	0,085710	0,106487	-0,106782	0,209140

"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
0,515502	-0,101102	0,117977	-0,254619	0,028722	0,221406	-0,021719

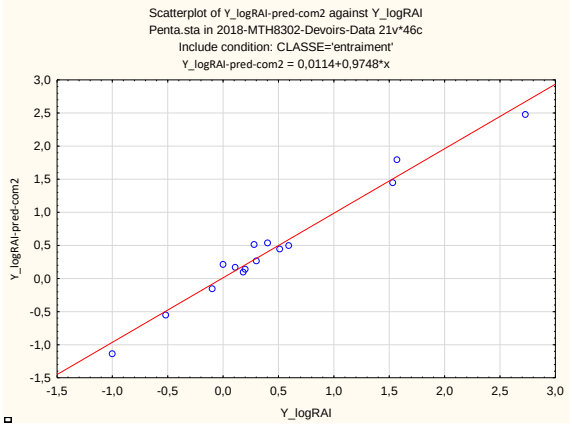
Y_logRAI-pred=b0+b1*S1+…b15*P5

Summary of PLS (Penta.sta in 2018-MTH8302-Devoirs-Data)
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Increase R² of Y	Average R² of Y	Increase R² of X	Average R² of X
Comp 1	0,896399	0,896399	0,169014	0,169014
Comp 2	0,078368	0,974767	0,127721	0,296735

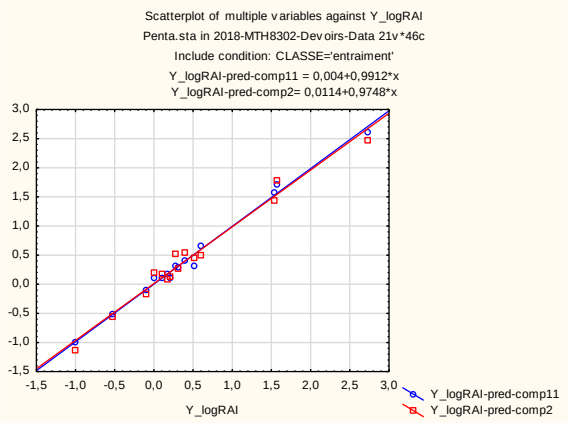
D'après les R2 de de Y(cumul) on remarque que nous avons une bonne corrélation entres les composantes et Y

Y_logRAI-OBSV	Y_logRAI-pred-com2	écart
0,00	0,21319	-0,21319
0,28	0,51533	-0,23533
0,20	0,14380	0,05620
0,51	0,44595	0,06405
0,11	0,17156	-0,06156
2,73	2,48083	0,24917
0,18	0,09637	0,08363
1,53	1,44762	0,08238
-0,10	-0,15456	0,05456
-0,52	-0,54923	0,02923



On remarque que le modèle prédit a bien suivi notre modèle réel (il y a une linéarisation entre Y et sa valeur prédit)

0,40	0,53927	-0,13927
0,30	0,26851	0,03149
-1,00	-1,13327	0,13327
1,57	1,79743	-0,22743
0,59	0,49719	0,09281



On remarque que les deux modèles prédits sont bien suivi le modèle réel pour les deux cas de composantes (2 et 11)
Pour les deux composantes on a presque la même résultat

7c)
Avec les variables S1 P1 S3 P3 L3 S4 L4 P4 on a trouvé R2 de Y pour composant2 égale 0,947 est ca signifie que avec c'est variable on a exprimé 95% de valeur prédit de Y et a cause de ca on peut éliminé les autres variables qui ne contribué pas avec les valeurs prédites

M3 : Modèle PLS (modèle avec les 2 premières composantes et les variables S1 P1 S3 P3 L3 S4 L4 P4)

Summary of PLS (Penta.sta in Workbook1_(Recovered))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Increase R² of Y	Average R² of Y	Increase R² of X	Average R² of X	NewVar
Comp 1	0,852271	0,852271	0,322533	0,322533	
Comp 2	0,094832	0,947103	0,221964	0,544498	

D'après les R2 de de Y(cumul) on remarque que nous avons une bonne corrélation entres les composantes et Y

Predictor weights (Penta.sta in Workbook1_(Recovered))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

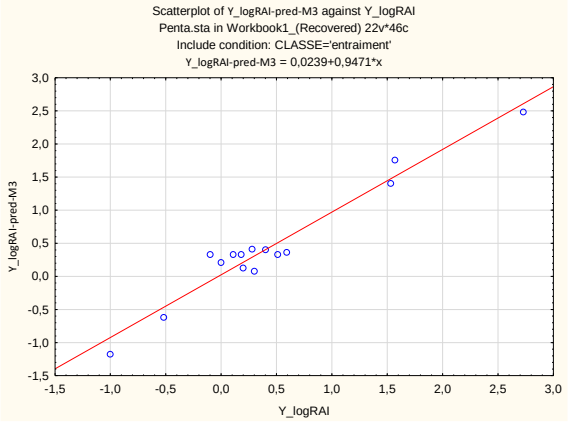
	"S1"	"P1"	"S3"	"P3"	"L3"	"S4"	"L4"	"P4"
Compo 1	-0,160987	-0,172906	-0,433825	0,291095	0,667575	-0,299635	0,304618	-0,207405
Compo 2	-0,298486	-0,301913	-0,153813	0,466094	0,393983	0,441079	-0,281950	0,386761

PLS regression coefficients (Penta.sta in Workbook1_(Recovered))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Interc.	"S1"	"P1"	"S3"	"P3"	"L3"	"S4"	"L4"	"P4"
Y_logRAI	0,241943	-0,103055	-0,276803	-0,111584	0,560382	0,220074	-0,076553	0,107457	-0,144669

Y_logRAI-pred=0,24-0,1*S1-0,27*P1-0,11*S3+0,56*P3+0,22*L3-0,07*S4+0,1*L4-0,14*P4

Y_logRAI-OBSV	Y_logRAI-pred-com2	écart
0,00	0,21108	-0,21108
0,28	0,41472	-0,13472
0,20	0,12791	0,07209
0,51	0,33154	0,17846
0,11	0,33154	-0,22154
2,73	2,48272	0,24728
0,18	0,33154	-0,15154
1,53	1,40562	0,12438

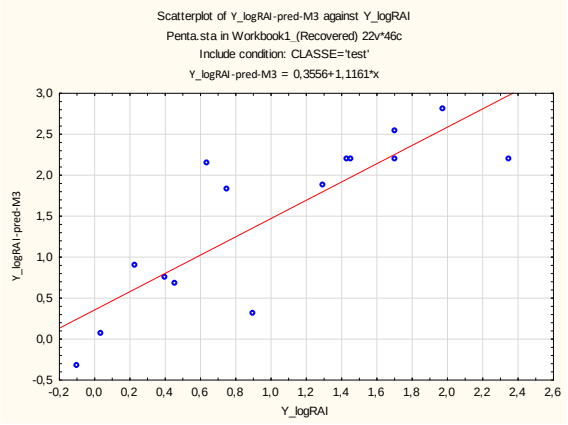


On remarque que le modèle prédit n'a pas bien suivi notre modèle réel (il y a une quelque erreur entre Y et sa valeur prédit)

-0,10	0,33154	-0,43154
-0,52	-0,61608	0,09608
0,40	0,40105	-0,00105
0,30	0,08113	0,21887
-1,00	-1,17345	0,17345
1,57	1,75619	-0,18619
0,59	0,36295	0,22705

7d)
Y_logRAI-pred=0,24-0,1*S1-0,27*P1-0,11*S3+0,56*P3+0,22*L3-0,07*S4+0,1*L4-0,14*P4

Y_logRAI-OBSV	Y_logRAI-pred-com2	écart
-0,10	-0,327235642	0,23
0,46	0,686589345	-0,23
0,75	1,82393778	-1,07
1,43	2,20	-0,78
1,45	2,20	-0,76
1,71	2,20	-0,50
0,04	0,05	-0,01
0,23	0,89	-0,66
1,30	1,86	-0,57
2,35	2,20	0,14
1,98	2,80	-0,82
1,71	2,53	-0,82
0,90	0,30	0,60
0,64	2,14	-1,51
0,40	0,75	-0,35
	-1,89	1,89



On remarque que le modèle prédit n'a pas bien suivi notre modèle réel (il y a une grand erreur entre Y et sa valeur prédit)

Pour que la régéssion soit plus rapide et éfficase on élimine les variables qui ne contribué pas avec les valeurs prédites et on prend le nombre de composante petit et R2 de Y (cumul) max, avec cela nous obtenons le même résultat

Conclusion générale

Dans cet devoir nous avons tretié des données réels ou on a utilisé la régression multiple ordinaire et Régression avec Forward Stepwise et Régression avec Backward Stepwise. Et on a tretié le probléme de multicolinéarité ou on a utilisé les méthodes (RIDGE, ACP) pour résoudre le probléme. Et on a bien étudié la régression PLS