Exercice 7 : Étude de prédiction d'activité biologique : modélisation PLS

7a)

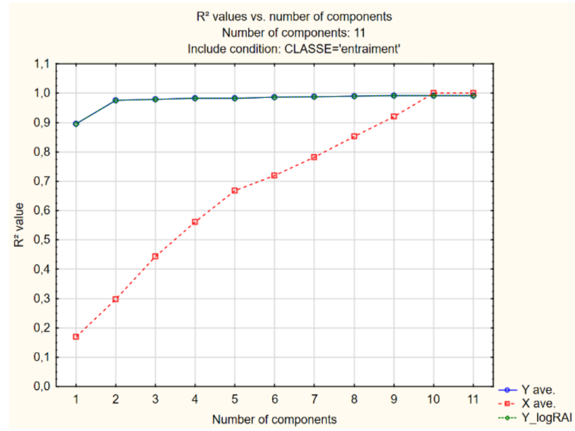
Tout d'abord, nous allons développer un premier modèle PLS (M1) en appliquant un filtre pour ne retenir que les observations de la classe Test. Ce premier modèle comprendra toutes les composantes. la régression PLS (Partial Least Squares) permet d'explorer la relation entre la variable de réponse Y_logRAI et les variables indépendantes S1 à P5.

Figure 20: Summary of PLS

Summary of PLS (Penta.sta in 2023-MTH8302-Dev
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Increase R^2 of Y	Average R^2 of Y	Increase R^2 of X	Average R^2 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

Figure 21: R^2 vs number of components



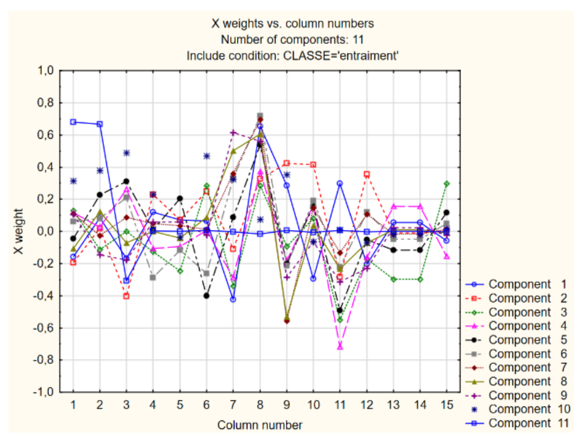
Comme le montre la Figure 21, nous avons la valeur du coefficient de détermination (R2) en fonction du numéro de la composante. Nous observons une augmentation du R2, cependant, à partir d'un certain seuil, la valeur de R2 n'augmente plus. L'objectif est donc de choisir un nombre de composantes qui donne une valeur de R2 suffisamment élevée. Dans le prochain modèle, M2, nous ne prendrons en compte que 2 composantes. De plus, le tableau de la Figure 20 présente un résumé des valeurs de R2 pour X et Y en fonction du nombre de composantes. Les colonnes "Average R2 of Y" et "R2 of X" donnent le cumul de l'accroissement du R2 en fonction du nombre de composantes retenues. Nous observons que pour Y, le coefficient de détermination commence à se stabiliser autour de 0,99 à partir de 8 composantes, tandis que pour X, cela se produit à partir de 10 composantes avec une valeur maximale de R2 de 0,991 et 1 respectivement.

Figure 22: Predictor Weights

Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"
Compo 1	-0,157641	0,085681	-0,169313	0,121527	0,071134	0,065188
Compo 2	-0,193810	0,021349	-0,404829	0,229796	0,071941	0,250337
Compo 3	0,130046	-0,113962	0,001702	-0,125617	-0,246332	0,283583
Compo 4	0,118332	0,025651	0,263189	-0,108027	-0,094759	0,006084
Compo 5	-0,045868	0,228215	0,310581	0,008716	0,204099	-0,399491
Compo 6	0,061752	0,080174	0,209562	-0,288848	-0,115836	-0,263064
Compo 7	0,110030	-0,026349	0,087751	0,049820	0,035832	0,013176
Compo 8	-0,108092	0,122918	-0,069392	0,000680	-0,042247	0,086976
Compo 9	0,106628	-0,146492	-0,178732	0,057439	0,059039	-0,020814
Compo 10	0,312921	0,378214	0,489695	0,231033	-0,035346	0,470374
Compo 11	0,679530	0,665916	-0,307054	0,004732	0,001234	0,005658

Figure 23: X Weights vs column numbers



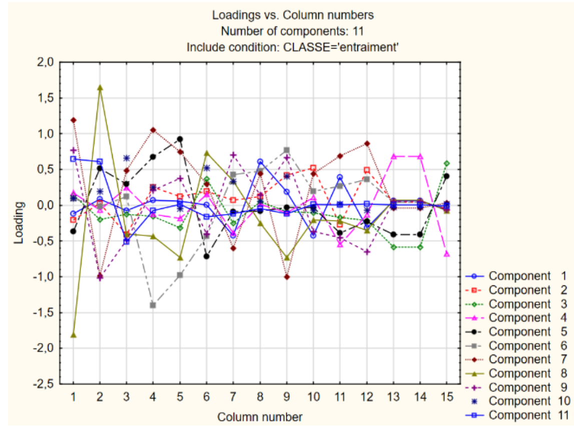
Sur le graphique de la Figure 23 "X Weights vs column numbers", on peut observer le poids des variables indépendantes pour chacune des composantes du modèle M1. Chaque couleur représente une composante particulière et le tableau de la Figure 22 donne les valeurs de ces poids, qui représentent l'impact de chaque variable explicative sur la réponse Y_LogRay. On peut remarquer, par exemple, que la colonne 8, ou L3, a une influence significative sur la prédiction de la variable dépendante. Les colonnes L1 et S1 ont également des poids importants et influencent grandement Y.

Figure 24: X loadings

X loadings (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2
Comp 1	-0,11633	0,08572	-0,072268	0,06815	0,056647
Comp 2	-0,20525	0,02768	-0,408255	0,25417	0,119262
Comp 3	0,14501	-0,20086	-0,123460	-0,15533	-0,314092
Comp 4	0,17703	-0,06847	0,243298	-0,12433	-0,181340
Comp 5	-0,36769	0,51481	0,294229	0,67401	0,926073
Comp 6	0,09541	-0,01646	0,123799	-1,40122	-0,978174
Comp 7	1,19269	-0,98727	0,487268	1,05068	0,746460
Comp 8	-1,80463	1,65360	-0,403619	-0,43346	-0,729090
Comp 9	0,77094	-1,01152	-0,491114	0,21237	0,377241
Comp 10	0,10335	0,19323	0,663150	0,25309	-0,042701
Comp 11	0,64526	0,60703	-0,514752	-0,07644	0,013872

Figure 25: loadings vs column numbers



L'analyse de ce graphique Figure 25 permet de déterminer les variables les plus importantes pour chaque composante. les variables avec les Loadings les plus importants pour chaque composante sont situées en haut sur le graphique. Les variables situées plus bas ont une influence plus faible sur la composante correspondante.

Figure 26: Coefficients de regression PLS

PLS regression coefficients (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Interc.	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
Y_logRAI	0,797596	0,126345	0,268079	-0,442360	0,103932	0,058552	0,080840	-0,049595	0,362006	0,117779	-0,031100	-0,057821	-0,328174	0,011786	0,090855	-0,008913

Figure 27: Scaled Coefficients de regression PLS

PLS scaled regression coefficients (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
Y_logRAI	0,208537	0,391487	-0,284554	0,088717	0,040368	0,071068	-0,131971	0,815799	0,060803	-0,028132	-0,060110	-0,062878	0,013625	0,013625	-0,013625

À l'aide de la figure 26 représentant les coefficients de régression, on peut déterminer l'équation du modèle M1, qui s'écrit comme suit : Y_logRAI_modele1= 0.798 + 0.126S1+0.268L1-0.443P1+0.104S2+0.059L2+0.081P2-0.05S3+0.36L3+0.12P3-0.031S4-0.06L4…-0.0089P5

Figure 28: Reg coeff vs col numbers

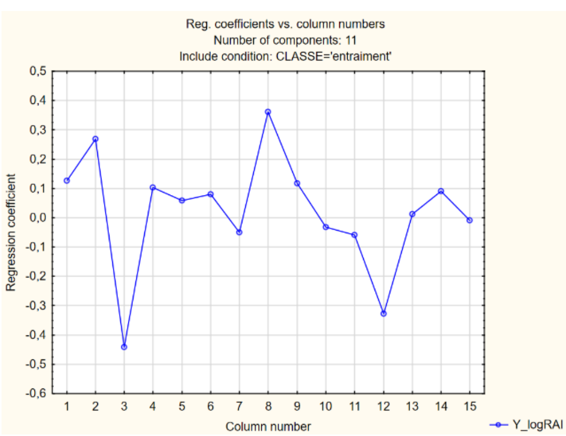
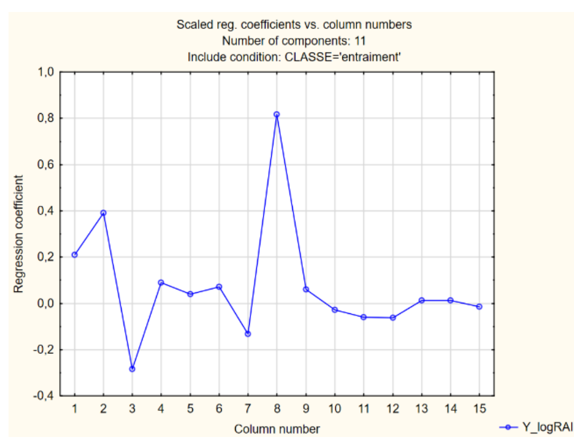


Figure 29: Scaled Reg Coefficients vs col numbers



Les figures 28 et 29 présentent les coefficients de régression et les coefficients de régression normalisés en fonction du numéro de colonne pour chaque composante du modèle PLS. La colonne correspond à une variable explicative dans le modèle. Nous pouvons observer que la variable L3 a le coefficient de régression le plus important, ce qui signifie qu'elle est la plus significative et a le plus grand impact sur la réponse Y. Ceci est cohérent avec ce que nous avons observé précédemment avec le graphique des X weights.

Figure 30: Change of regr Coefficients by number of components

Change of regr. coeffs (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	1 Comp Y_logRAI	2 Comp Y_logRAI	3 Comp Y_logRAI	4 Comp Y_logRAI	5 Comp Y_logRAI	6 Comp Y_logRAI	7 Comp Y_logRAI	8 Comp Y_logRAI	9 Comp Y_logRAI	10 Comp Y_logRAI	11 Comp Y_logRAI
"S1"	-0,058669	-0,083797	-0,079322	-0,075482	-0,076964	-0,072421	-0,063172	-0,071153	-0,067719	-0,067601	0,126345
"L1"	0,036041	0,039170	0,034737	0,035678	0,044016	0,050682	0,048179	0,058437	0,053104	0,053264	0,268079
"P1"	-0,161685	-0,296360	-0,296210	-0,274292	-0,248533	-0,208973	-0,190047	-0,203193	-0,217965	-0,217493	-0,442360
"S2"	0,087455	0,145064	0,136705	0,129926	0,130471	0,089380	0,097478	0,097575	0,101152	0,101320	0,103932
L2	0,063380	0,085710	0,065416	0,058054	0,073848	0,053446	0,060656	0,053189	0,057741	0,057710	0,058552
"P2"	0,045550	0,106487	0,124808	0,125179	0,100935	0,064598	0,066678	0,078735	0,077476	0,077808	0,080840
"S3"	-0,098066	-0,106782	-0,114041	-0,119853	-0,118108	-0,102999	-0,084254	-0,061293	-0,049004	-0,048928	-0,049595
"L3"	0,178190	0,209140	0,216353	0,225198	0,237959	0,276711	0,319586	0,352348	0,365546	0,365566	0,362006
"P3"	0,339176	0,515502	0,505293	0,486649	0,466035	0,416112	0,266446	0,141511	0,112378	0,112801	0,117779
"S4"	-0,199250	-0,101102	-0,095955	-0,086418	-0,076997	-0,051234	-0,028915	-0,024183	-0,027885	-0,027929	-0,031100
"L4"	0,176255	0,117977	0,087928	0,050902	0,025660	-0,000399	-0,018150	-0,045369	-0,061455	-0,061449	-0,057821
"P4"	-0,651143	-0,254619	-0,305488	-0,350220	-0,364320	-0,288810	-0,210585	-0,256265	-0,319765	-0,319997	-0,328174
"S5"	0,030239	0,028722	0,014028	0,021277	0,015938	0,010820	0,009082	0,010032	0,011040	0,011031	0,011786
"L5"	0,233097	0,221406	0,108133	0,164014	0,122856	0,083408	0,070007	0,077334	0,085105	0,085032	0,090855
"P5"	-0,022866	-0,021719	-0,010608	-0,016089	-0,012052	-0,008182	-0,006868	-0,007586	-0,008349	-0,008341	-0,008913

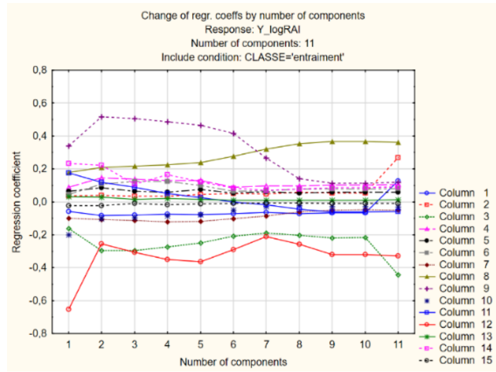


Figure 31: Normal Probability Plot

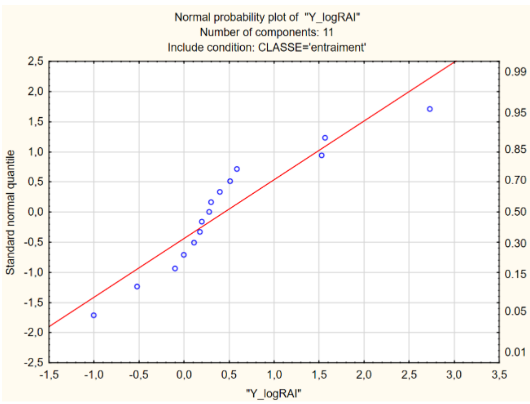
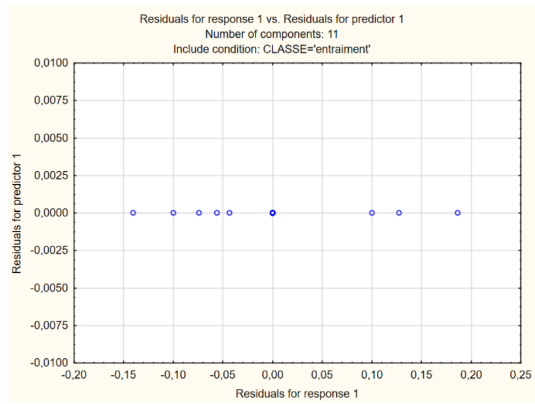


Figure 32: Residuals vs Predictor

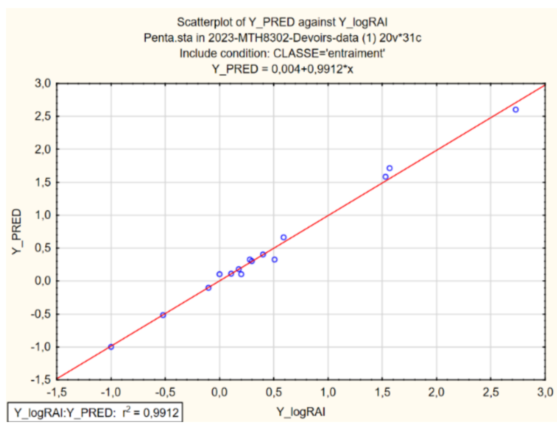


Les observations sont a peut pret normale et les points dans le graphique "residuals vs predictors" sont alignés sur la ligne y=0, cela signifie que les résidus (différence entre les valeurs prédites et les valeurs réelles) sont en moyenne égaux à zéro pour chaque niveau de la variable indépendante. Cela suggère que le modèle est bien ajusté ce qui peut etre attendu avec R2 de 99.12%.

Figure 33: Tableau de Y_obs et Ypred

18 Y_logRAI	19 CLASSE	20 Y_PRED
0,00	entraiment	0,10001
0,28	entraiment	0,32348
0,20	entraiment	0,09999
0,51	entraiment	0,32345
0,11	entraiment	0,11000
2,73	entraiment	2,60227
0,18	entraiment	0,18000
1,53	entraiment	1,58613
-0,10	entraiment	-0,10000
-0,52	entraiment	-0,52000
0,40	entraiment	0,40000
0,30	entraiment	0,30000
-1,00	entraiment	-1,00000
1,57	entraiment	1,71053
0,59	entraiment	0,66413

Figure 34: Scatterplot de Y_Pred vs Y_LogRai



On peut observer sur la figure 34, un scatterplot de Y_Pred vs Y_LogRai, que ce modèle présente un coefficient de détermination R2 de 99,1%, ce qui indique que 99.1% de la variance est expliquée par le modèle. En effet, on peut observer que les observations sont quasi alignées avec la droite, ce qui confirme une bonne performance du modèle.

7b)

Nous developpons maintenant un deuxième modèle PLS (M2) est basé sur les 2 premières composantes seulement.

Figure 35: Summary of PLS

Summary of PLS (Penta.sta in 2023-MTH8302-Dev
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

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

Figure 36: Predictor weights responses

Predictor weights (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

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

Figure 37: R^2 vs number of components

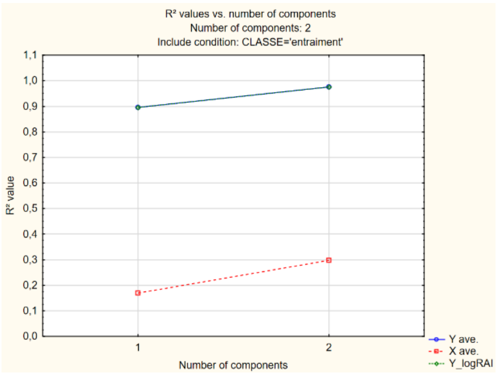
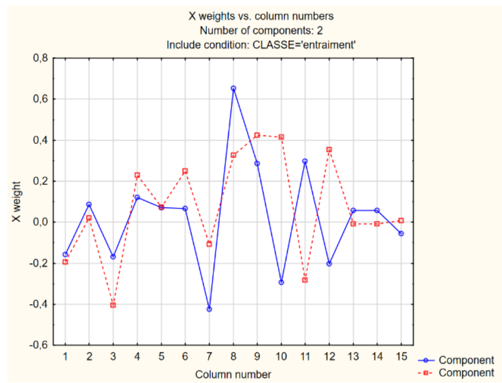


Figure 38: X weights vs column numbers



Le deuxième modèle PLS (M2) est basé sur les 2 premières composantes seulement, contrairement au modèle M1 qui utilise toutes les composantes. L'abandon des composantes au-delà des 2 premières est justifié par une observation du graphique X weights vs column number du modèle M1, où l'on voit que les poids des variables indépendantes diminuent fortement à partir de la troisième composante. Cela indique que les variables indépendantes sont moins importantes pour expliquer la variation de T_logRAI à partir de la troisième composante.

Figure 39: X loadings

X loadings (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
Comp 1	-0,116329	0,085722	-0,072268	0,068149	0,056647	0,002522	-0,424336	0,610345	0,190277	-0,424245	0,394162	-0,311831	0,062925	0,062925	-0,062925
Comp 2	-0,205250	0,027681	-0,408255	0,254169	0,119262	0,196278	0,072225	0,122679	0,421627	0,523812	-0,267828	0,492092	0,045734	0,045734	-0,045734

Figure 40: Coefficients de regression PLS

PLS regression coefficients (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	Interc.	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
Y_logRAI	-0,437289	-0,083797	0,039170	-0,296360	0,145064	0,085710	0,106487	-0,106782	0,209140	0,515502	-0,101102	0,117977	-0,254619	0,028722	0,221406	-0,021719

Figure 41: Scaled Coefficients de regression PLS

PLS scaled regression coefficients (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"L1"	"P1"	"S2"	L2	"P2"	"S3"	"L3"	"P3"	"S4"	"L4"	"P4"	"S5"	"L5"	"P5"
Y_logRAI	-0,138311	0,057201	-0,190637	0,123827	0,059091	0,093615	-0,284145	0,471307	0,266127	-0,091454	0,122649	-0,048785	0,033203	0,033203	-0,033203

Nous pouvons observer que la variable L3 a le coefficient de régression le plus important, ce qui signifie qu'elle est la plus significative et a le plus grand impact sur la réponse Y_LogRai.

Figure 42: Loading vs Column numbers

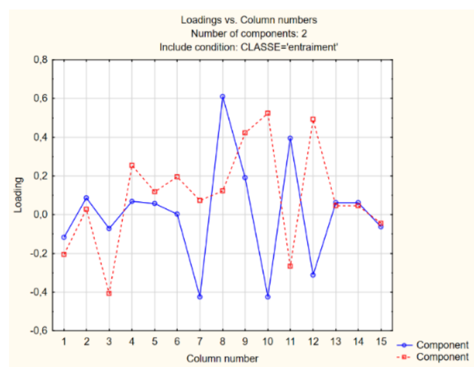


Figure 43: Change of regr Coeffs

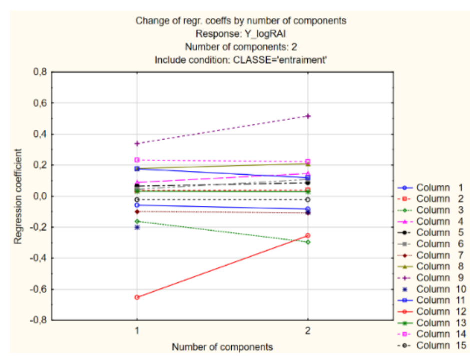


Figure 44: Normal Probability Plot

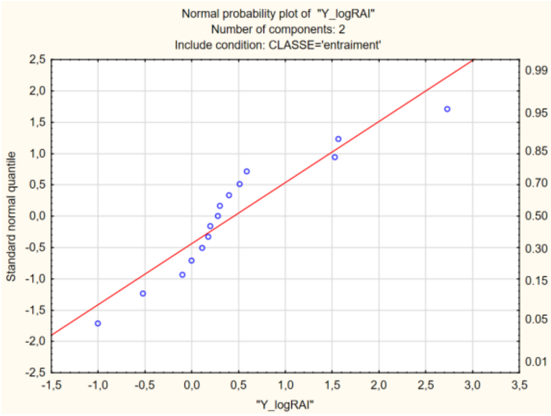
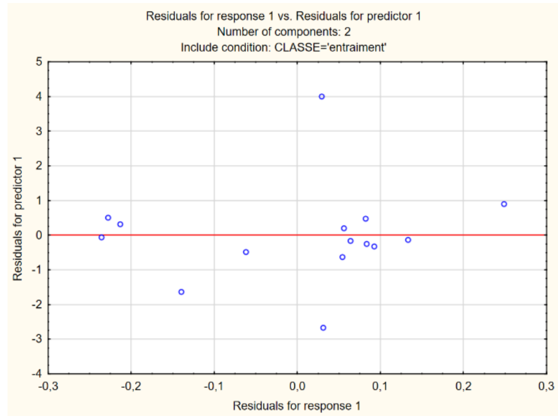
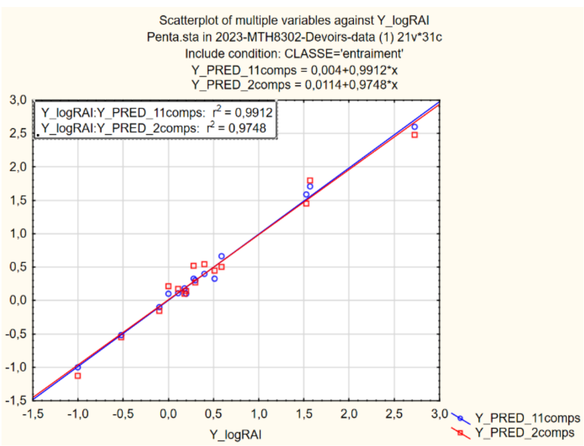


Figure 43: Residuals vs Predictor



Les observations suivent un distribution normale et les points dans le graphique "residuals vs predictors" sont distribue aléatoirement autour de 0,cela signifie que les erreurs de prédiction sont réparties de manière égale autour de zéro pour toutes les valeurs prédites, ce qui suggère que le modèle ne présente pas de biais important dans ses prédictions. Ce qui indique un bon ajustement.

Figure 45: Scatterplot Y_PRED_2 vs Y_PRED_11 against Y_LogRAI



Le Scatterplots suivants comparent les droites de prédiction des deux modèles, celui avec toutes les composantes (M1) et celui avec seulement 2 composantes (M2). On remarque que les deux modèles ont un bon ajustement, avec un R2 de 0.99 pour le modèle M1 et de 0.975 pour le modèle M2, ce qui représente une différence négligeable de 1.5%. On peut donc dire que le modèle M2 est préférable, car l'abandon des composantes au-delà des 2 premières est justifié. En effet, on avait remarqué que les poids des variables indépendantes diminuaient fortement à partir de la troisième composante dans le modèle M1. La réduction du nombre de composantes peut aider à éviter le surajustement et à améliorer la généralisation aux nouvelles données, car cela réduit la complexité du modèle et élimine les composantes qui ne contribuent pas de manière significative à la prédiction de la variable dépendante. Cela peut également rendre le modèle plus simple à interpréter et plus facile à utiliser dans la pratique.

7c)

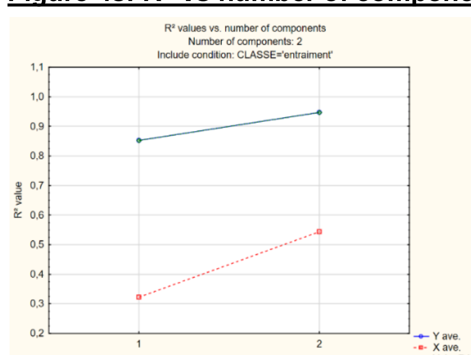
Nous allons maintenant développer modèle M3 basé sur les 2 premières composantes et sur les régresseurs S1 P1 S3 P3 L3 S4 L4 P4.

Figure 47: Summary PLS

Summary of PLS (Penta.sta in 2023-MTH8302-Dev
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

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

Figure 48: R^2 vs number of components



Le modèle M3 à deux composantes a un coefficient de détermination de 0,948, ce qui signifie qu'il explique 94,8 % de la variance des données. Bien que ce soit inférieur aux deux premiers modèles, cela reste tout à fait raisonnable et est considéré comme un bon ajustement. Comme on peut le voir sur la Figure 48, le passage de 1 à 2 composantes

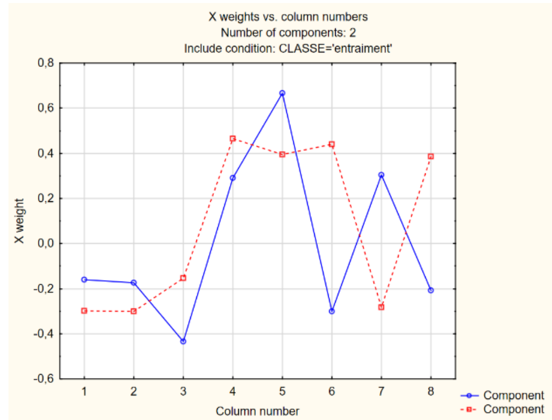
augmente significativement le R2, en passant de 0,85 à 0,95, ce qui représente un gain de 10 % en précision.

Figure 49: Predictor weights

Predictor weights (Penta.sta in 2023-MTH8302-Devoirs-data (1))
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

Figure 50: X weights vs column numbers



On voit sur la Figure 50 que le poids de la variable explicative column number 5 ou "L3" est le plus elevé ce qui etait également le cas pour les modèles précedent.

Figure 51: Loadings vs column numbers

X loadings (Penta.sta in 2023-MTH8302-Devoirs-data (1))
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='entraiment'

	"S1"	"P1"	"S3"	"P3"	"L3"	"S4"	"L4"	"P4"
Comp 1	-0,087420	-0,099271	-0,422937	0,179278	0,605520	-0,449743	0,409280	-0,334766
Comp 2	-0,273234	-0,463735	0,063804	0,473267	0,161839	0,533399	-0,250389	0,512029

Figure 52: Reg coeff va column numbers

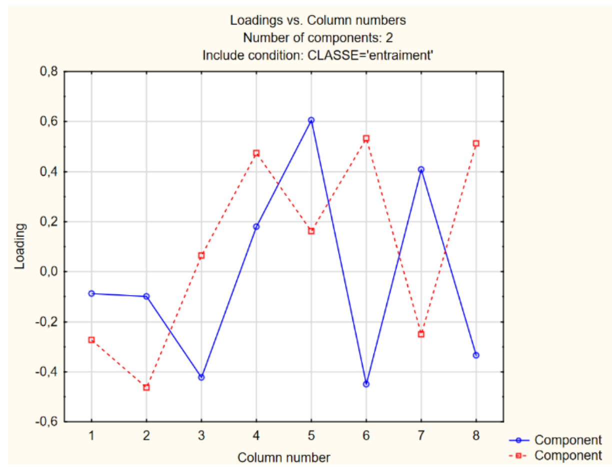
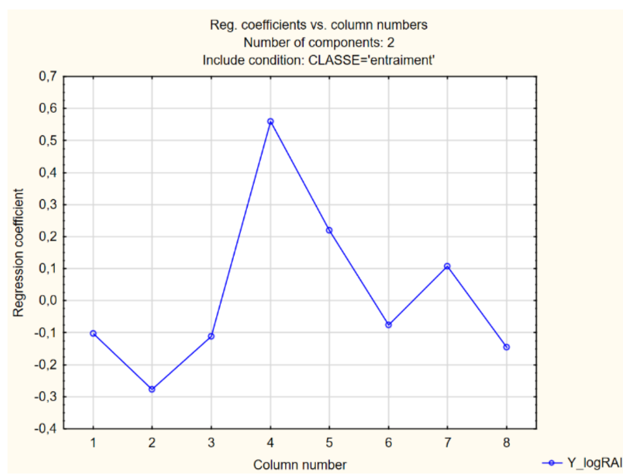


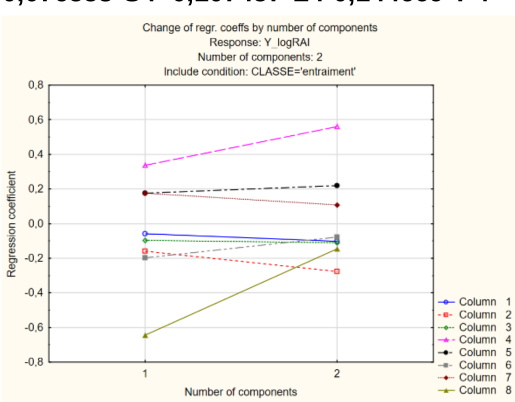
Figure 53: regr.coefficients

PLS regression coefficients (Penta.sta in 2023-MTH8302-Devoirs-data (1))
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

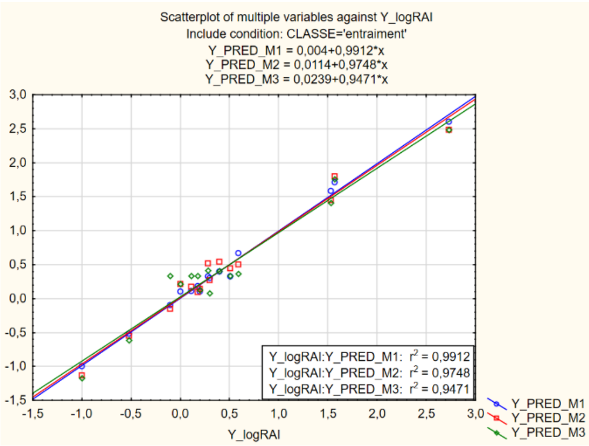


À l'aide de la figure 53 représentant les coefficients de régression, on peut déterminer l'équation du modèle M1, qui s'écrit comme suit : Y_logRAI_modele1=
: Y_logRAI = -0,241943-0,10355*S1-0,276803*P1 -0,111584*S3+0,220074*L3+0,560382*P3-0,076553*S4+0,107457*L4-0,144669"P4



Le graphique suivant montre le changement des coefficients de regression lorsqu'on passe de une à la deux composantes, les couleurs representent les variables explicatives retenue dans le modele M3.

Figure 54: Scatterplot Y_Obs vs Y_pred



Le modèle M3 est un bon ajustement pour la prédiction des valeurs Y_LogRai avec un R2 de 0,948. Comme on peut le voir sur la figure 54, qui représente une comparaison des 3 modèles et leur capacité à prédire les valeurs de Y, la différence entre les 3 modèles est négligeable. L'abondan des variables L1, S2, L2, P2, S5, L5 et P5 permet de simplifier le modèle en ne gardant que les variables significatives dont le poids est le plus élevé et qui ont la plus grande influence sur la variable expliquée Y_LogRai. En conclusion, l'élimination des variables non significatives permet de diminuer la complexité du modèle. Cela peut également rendre le modèle plus simple à interpréter et plus facile à utiliser dans la pratique.

7d) On emploie le modèle M3 pour prédire l'activité brakinine pour les données de la deuxième étude:

Figure 55 : Summary of PLS for test

Summary of PLS (Penta.sta in 2023-MTH8302-Dev
Responses: Y_logRAI
Options: NO-INTERCEPT AUTOSCALE
Include condition: CLASSE='test'

	Increase R^2 of Y	Average R^2 of Y	Increase R^2 of X	Average R^2 of X
Comp 1	0,549403	0,549403	0,402486	0,402486
Comp 2	0,121264	0,670667	0,204661	0,607147

Figure 56 : R^2 vs number of component

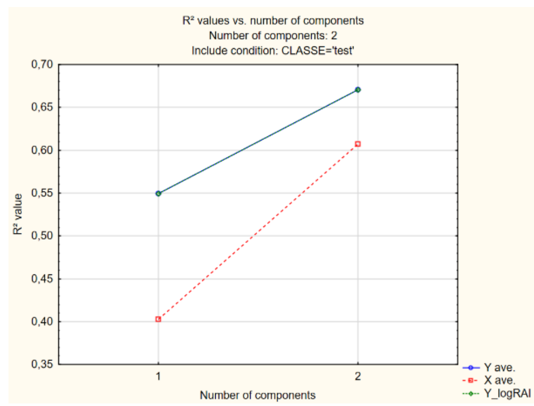


Figure 57 : Predictor weights

Include condition: CLASSE='test'

	"L1"	"S2"	L2	"P2"	"S5"	"L5"	"P5"
Compo 1	-0,033319	0,599793	0,154412	-0,608282	-0,314527	0,153465	0,350440
Compo 2	-0,515569	0,450272	0,205986	-0,263164	0,358537	0,445553	-0,304454

Figure 59 : Reg coeffs vs column number

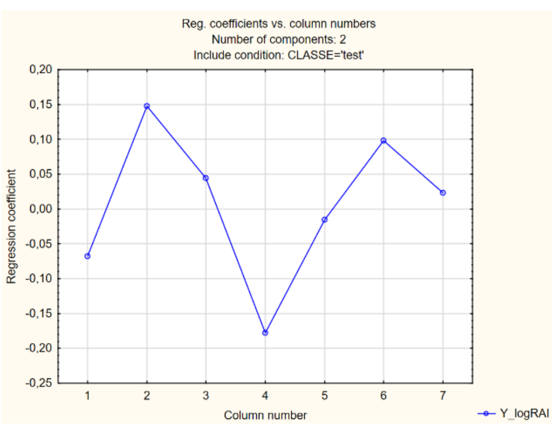
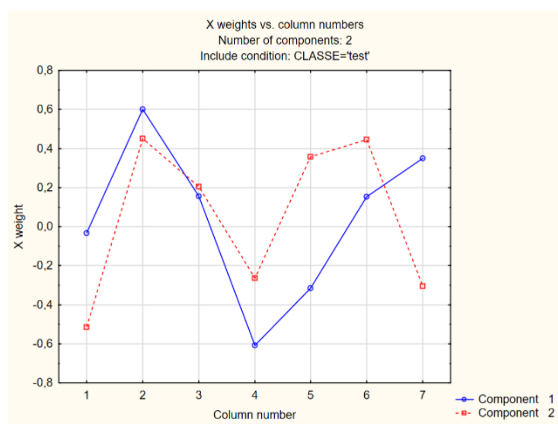


Figure 58 : X weights vs column numbers



On peut remarquer que les coefficients de régression ainsi que les poids diffèrent grandement entre les deux études. Tout d'abord, le poids de la variable 4 ou P2, qui était le plus élevé, est le plus faible pour la deuxième étude, et la variable S2, qui a le poids le plus élevé dans la deuxième étude, a le poids le plus faible dans la première étude. De plus, comme on peut le voir sur les figures 55 et 56, la valeur du coefficient de détermination est de 0,67, ce qui est beaucoup plus faible que celui obtenu dans la première étude, qui était de 0,95. Cela signifie que seulement 67 % de la variance totale est expliquée par le modèle M3 pour les données tests.

Figure 60 : Normal probabilty plot de Y pour les tests

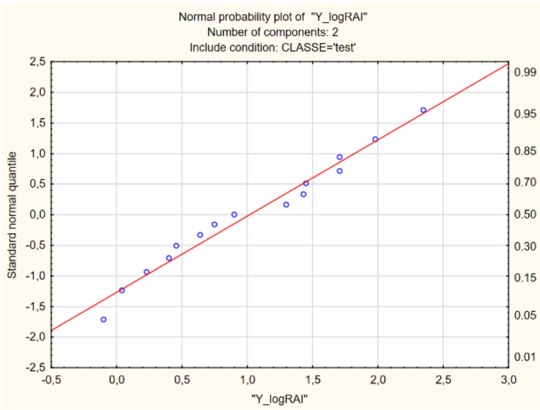
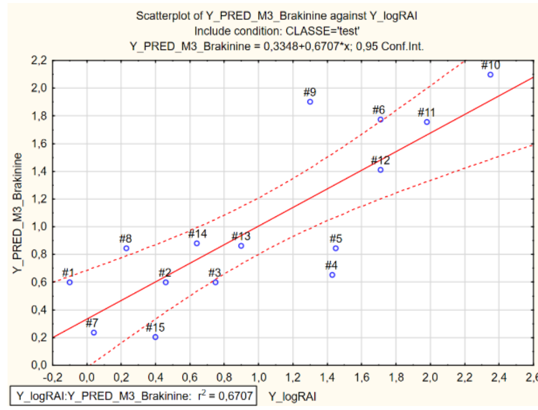


Figure 61 :Scatterplot Y_PredTest vs Y_LogObs



On peut voir sur la Figure 60 que le modèle suit une distribution normale, et que la normalité est mieux respectée que dans les modèles de la première étude. Cependant, Les résultats montrent que le modèle M3 ne prédit pas très bien les données de la deuxième étude. En effet, la valeur du R^2 est seulement de 0,67, ce qui signifie que le modèle explique seulement 67% de la variance totale. De plus, nous pouvons observer sur le scatterplot que plusieurs valeurs prédites de Y sont à l'extérieur de l'intervalle de confiance. Lorsque nous avons ré-analysé les données de la deuxième étude en incluant toutes les variables explicatives, nous avons obtenu des valeurs de R^2 beaucoup plus élevées. Pour 2 composantes, le R^2 était de 0,92 et pour 13 composantes, il était de 0,9999. Cela indique que l'élimination des variables explicatives L1 S2 L2 P2 S5 L5 P5 dans le modèle M3 peut être la cause de la faible performance de prédiction pour les données de la deuxième étude. Les variables qui n'étaient pas influentes dans la première étude semblent être importantes dans la deuxième étude et doivent être prises en compte pour améliorer la performance de prédiction et obtenir un meilleur ajustement. En conclusion, le modèle M3 basé sur les régresseurs S1 P1 S3 P3 L3 S4 L4 P4, bien qu'il explique une grande partie de la variance dans les données d'entraînement, ne parvient pas à prédire avec précision les données de la deuxième étude. Cela peut être dû aux différences entre les peptides et les bradykinines utilisées dans les deux études. Pour améliorer la performance de prédiction, il est recommandé de réexaminer les variables explicatives à inclure dans le modèle et d'inclure toutes les variables importantes pour les deux études.