POLYTECHNIQUE MONTRÉAL

MTH8302
Modèles de régression et d'analyse de la variance
Devoir 3
distribution : 13 juin 2018
remise: 23 juin 2018 (au plus tard 23h55)

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.

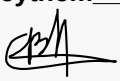
Information concernant le plagiat à Polytechnique : http://www.polymtl.ca/etudes/ppp/index.php

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.
Une copie de cette page est disponible sur le site du cours.

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-Devoir3.pdf**

TABLEAU CORRECTION

	valeur	obtenu	
No 8 - BostonHousing		30	
No 9 - Amphétamine		30	
No 10 - Assurances		30	
Qualité générale		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/

No 8 Étude de modélisation avec MARS et réseaux de neurones

Données = BostonHousing.sta

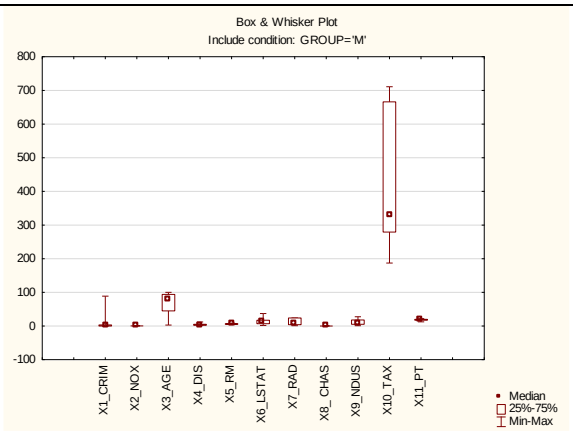
Réponse

8a)

Modèle de Régression MARS (MARS) sur l'ensemble M

Means and Standard Deviations (BostonHousing.sta in 2018-M
Include condition: GROUP='M'

Variable	Means	Std.Dev.	Min	Max	N
X1_CRIM	3,7042	9,1754	0,0063	88,9762	405
X2_NOX	0,5544	0,1176	0,389	0,871	405
X3_AGE	68,7998	28,0887	2,9	100	405
X4_DIS	3,8241	2,1210	1,1296	12,1265	405
X5_RM	6,2888	0,7046	3,561	8,78	405
X6_LSTAT	12,6965	7,1497	1,73	36,98	405
X7_RAD	9,5086	8,6424	1	24	405
X8_CHAS	0,0691	0,2540	0	1	405
X9_NDUS	11,0376	6,8866	0,74	27,74	405
X10_TAX	406,3580	168,1131	187	711	405
X11_PT	18,4111	2,1460	12,6	22	405



On remarque que la variable X10_Tax plus important par rapport les autres avec une avec une moyenne de 406,35 alors que les autres variables ont des moyennes variantes entre 0 et 69. Le graphique de Box&Whisker confirmer notre remarque.

0 cases with missing data were found.

MARSplines Results:
Dependent: Y_MV
Independents: X1_CRIM, X2_NOX, X3_AGE, X4_DIS, X5_RM, X6_LSTAT, X7_RAD, X8_ CHAS, X9_NDUS, X10_TAX, X11_PT
Number of terms = 16
Number of basis functions = 23
Order of interactions = 2
Penalty = 2,000000
Threshold = 0,000500
GCV error = 12,366354
Prune = Yes

Après la régression MARS, les variables retenues dans le modèle (7 sur 11) sont X1_CRIM, X2_NOX, X4_DIS, X5_RM, X6_LSTAT, X10_TAX et X11_PT.

Number of References to Each Predictor (BostonHousi
Number of times each predictor is referenced (used)
Include condition: GROUP='M'

Dependents	References (to Basis Functions)
X1_CRIM	2
X2_NOX	3
X3_AGE	0
X4_DIS	3
X5_RM	5
X6_LSTAT	7
X7_RAD	0
X8_CHAS	0
X9_NDUS	0
X10_TAX	2
X11_PT	1

Regression statistics (BostonHo
Include condition: GROUP='M'

Regression statistics	Y_MV
Mean (observed)	22,64568
Standard deviation (observed)	9,31405
Mean (predicted)	22,64568
Standard deviation (predicted)	8,72810
Mean (residual)	-0,00000
Standard deviation (residual)	3,25143
R-square	0,87814
R-square adjusted	0,87311

D'après le tableau on remarque que notre modèle prédit est très bon (R^2 = 87,8% et R^2ajusté = 87,3% les moyennes des valeurs observées et prédites

Ce tableau nous donne le nombre le nombre de fois que la variable est présente dans l'expression de l'équation prédictive. **D'après le tableau les variables X6_LSTAT (7 fois) et X5_RM (5 fois) sont les plus référencées. Et les significatives.**	**sont égales et leurs écarts-types proches (9,3 et 8,7)**

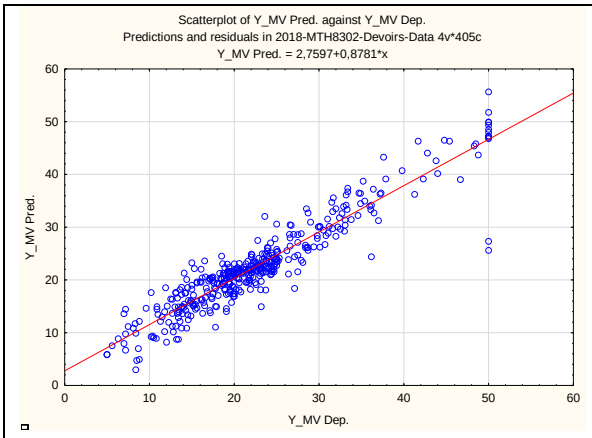
```
NOTE: The following model should be used directly, with categorical variables being
coded 0, 1.

Y_MV = 2,19770047540900e+001 - 5,53036467170238e-001*max(0; X6_LSTAT-
6,72000000000000e+000) + 1,15911619953539e+001*max(0; X5_RM-6,43400000000000e+000)
- 2,55229322508175e+000*max(0; 6,43400000000000e+000-X5_RM) - 1,96290138273353e+002
*max(0; X2_NOX-6,31000000000000e-001)*max(0; X5_RM-6,43400000000000e+000) +
1,57584987036513e-002*max(0; 6,72000000000000e+000-X6_LSTAT)*max(0; X10_TAX-
3,04000000000000e+002) + 2,50159573202584e-002*max(0; 6,72000000000000e+000-X6
_LSTAT)*max(0; 3,04000000000000e+002-X10_TAX) - 1,35694194791614e-001*max(0; X1
_CRIM-1,52880000000000e+001) + 1,62370332504544e-001*max(0; 1,52880000000000e+001-
X1_CRIM) + 7,60584716892318e+000*max(0; 1,85890000000000e+000-X4_DIS)*max(0;
6,43400000000000e+000-X5_RM) - 1,03828989230395e+000*max(0; X2_NOX-
6,14000000000000e-001)*max(0; X6_LSTAT-6,72000000000000e+000) + 2,51683385931161e+
000*max(0; 6,14000000000000e-001-X2_NOX)*max(0; X6_LSTAT-6,72000000000000e+000) -
1,73742309174300e-001*max(0; X6_LSTAT-6,72000000000000e+000)*max(0; X11_PT-
1,92000000000000e+001) + 6,87837946114155e-001*max(0; 6,43400000000000e+000-X5_RM)
*max(0; X6_LSTAT-2,68200000000000e+001) - 4,89877160805664e-001*max(0; X4_DIS-
3,26280000000000e+000) + 1,97656447281912e+000*max(0; 3,26280000000000e+000-X4_DIS)
```

L'équation prédictive du modèle avec les variables indépendantes retenues (16 termes différents pour 23 fonctions de base)

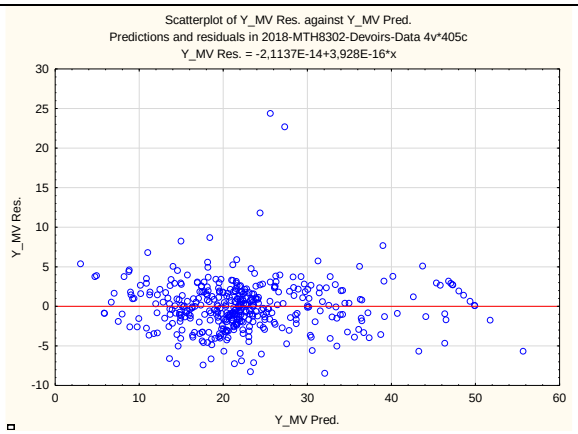
Model coefficients (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
NOTE: Highlighted cells indicate basis functions of type
max(0, independent-knot), otherwise max(0, knot-independent)
Include condition: GROUP='M'

Coefficients, knots and basis functions	Coefficients Y_MV	Knots X1_CRIM	Knots X2_NOX	Knots X3_AGE	Knots X4_DIS	Knots X5_RM	Knots X6_LSTAT	Knots X7_RAD	Knots X8_CHAS	Knots X9_NDUS	Knots X10_TAX	Knots X11_PT
Intercept	21,977											
Term.1	-0,553						6,72000					
Term.2	11,591					6,434000						
Term.3	-2,552					6,434000						
Term.4	-196,290		0,631000			6,434000						
Term.5	0,016						6,72000				304,0000	
Term.6	0,025						6,72000				304,0000	
Term.7	-0,136	15,28800										
Term.8	0,162	15,28800										
Term.9	7,606				1,858900	6,434000						
Term.10	-1,038		0,614000				6,72000					
Term.11	2,517		0,614000				6,72000					
Term.12	-0,174						6,72000					19,20000
Term.13	0,688					6,434000	26,82000					
Term.14	-0,490				3,262800							
Term.15	1,977				3,262800							

Ce tableau nous donne les nœuds employés dans le modèle

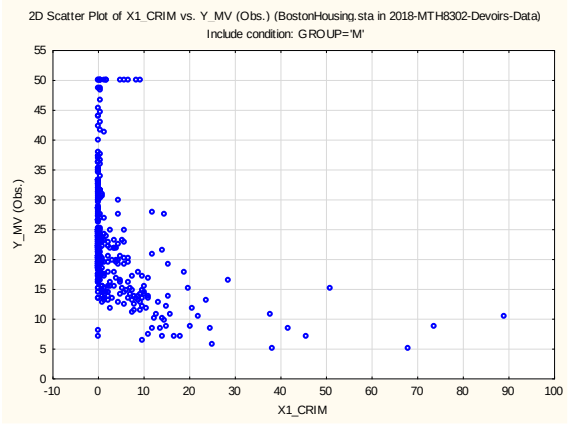


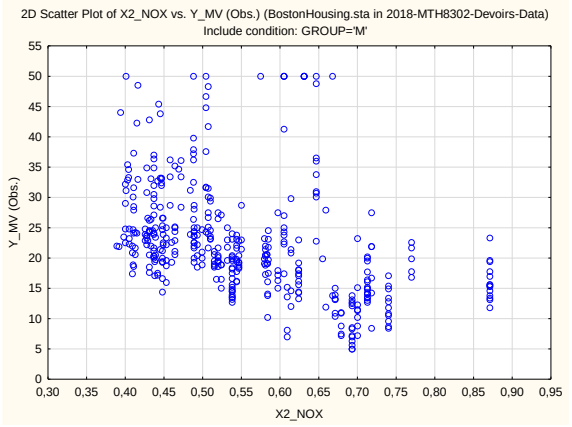
On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l'équation prédictive linéaire du modèle avec les valeurs observées)

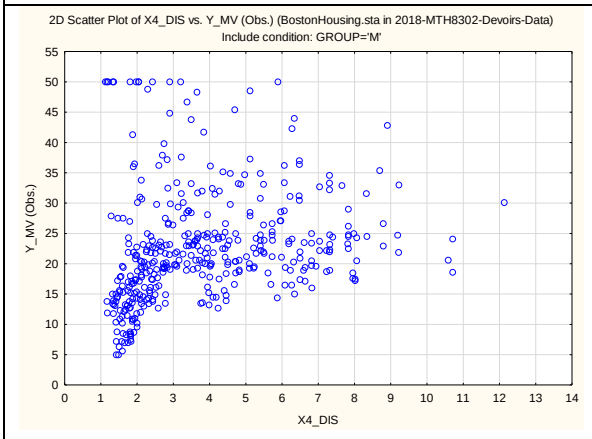


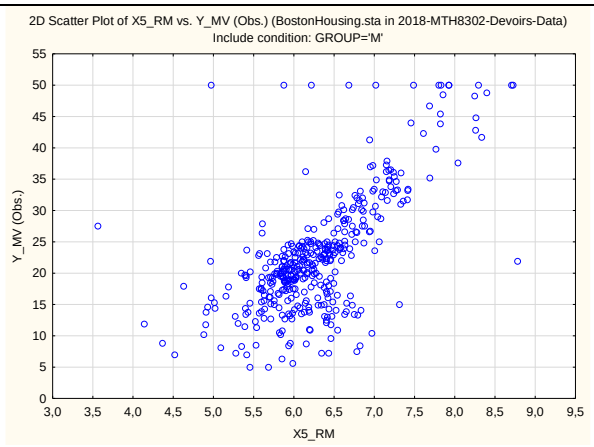
D'après le graphe de l'analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon.

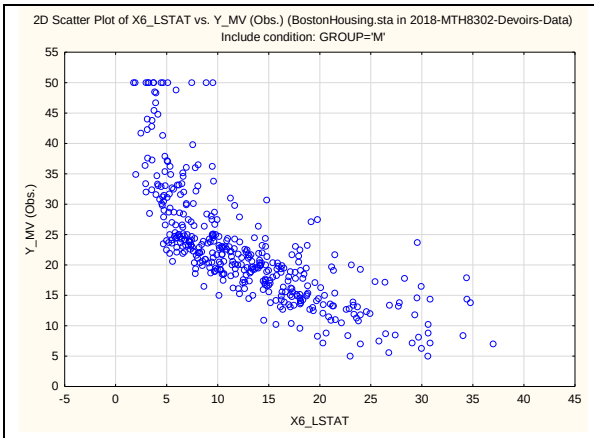
Les graphiques qui identifient les nœuds employés dans le modèle :

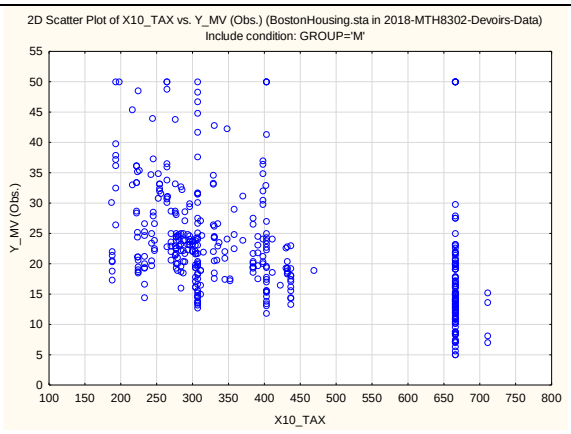


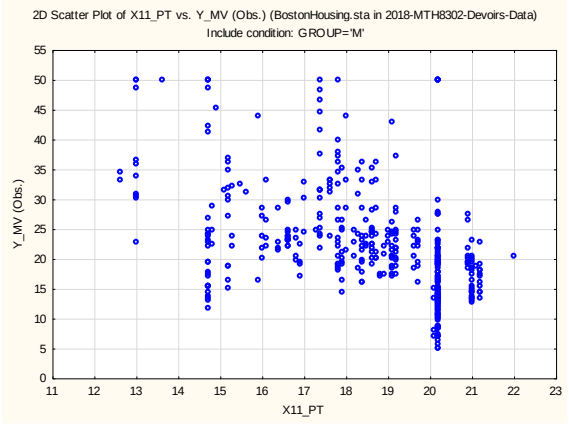






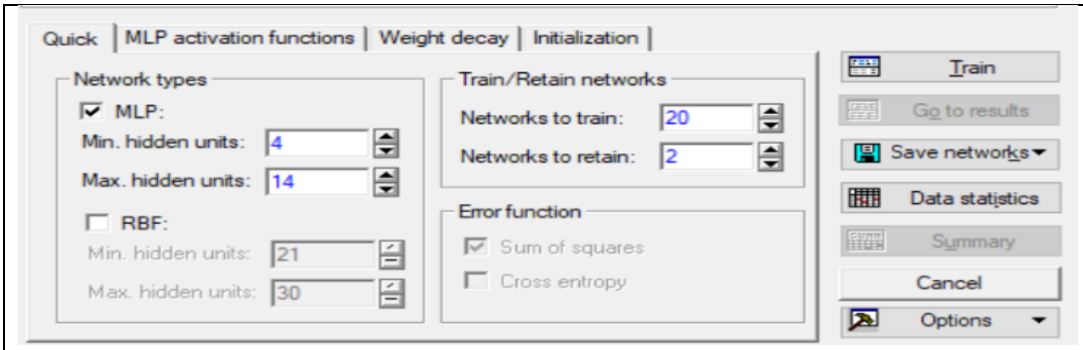






8b)

Réseaux de neurones sur l'ensemble M

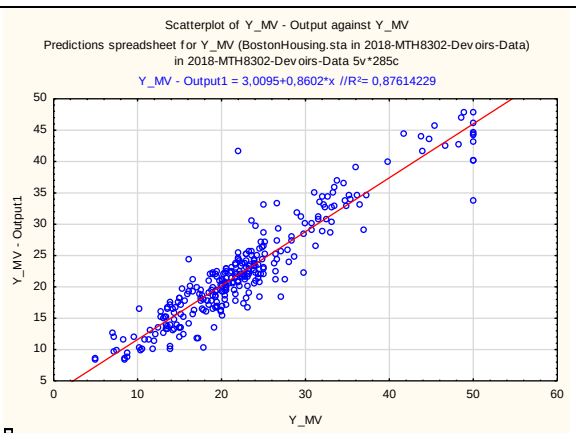


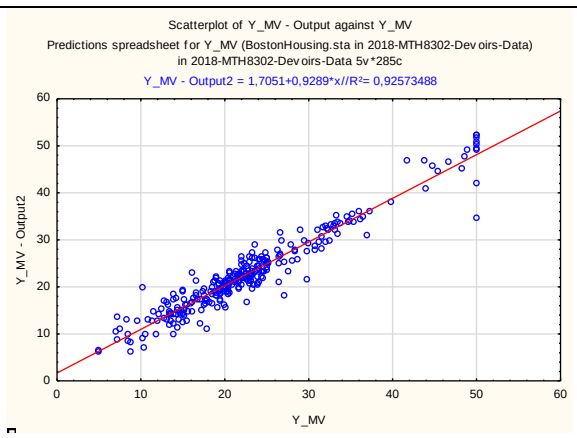
Nous avons développé 20 réseaux de neurones et nous avons retenu les 2 meilleurs

Summary of active networks (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Include condition: GROUP='M'

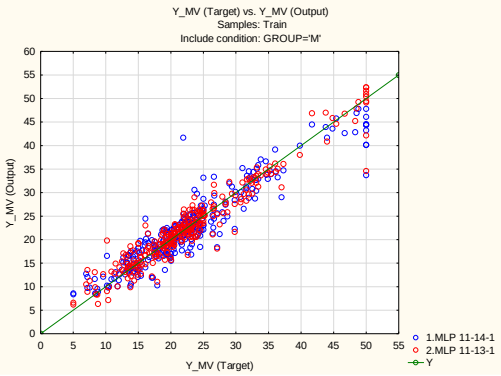
Index	Net. name	Training perf.	Test perf.	Validation perf.	Training error	Test error	Validation error	Training algorithm	Error function	Hidden activation	Output activation
1	MLP 11-11-1	0,960126	0,953138	0,907097	3,32030	4,665149	7,115698	BFGS 48	SOS	Exponential	Identity
2	MLP 11-8-1	0,859212	0,934607	0,894741	11,18472	6,973176	7,922635	BFGS 14	SOS	Identity	Exponential

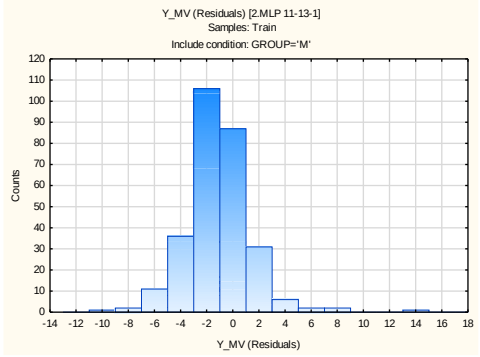
Ce tableau nous donne les informations sur les deux meilleurs réseaux retenus dans le modéle

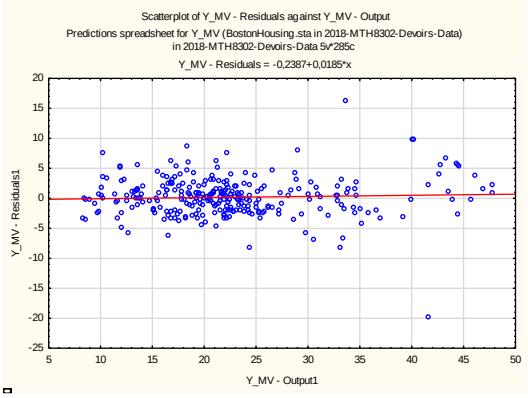


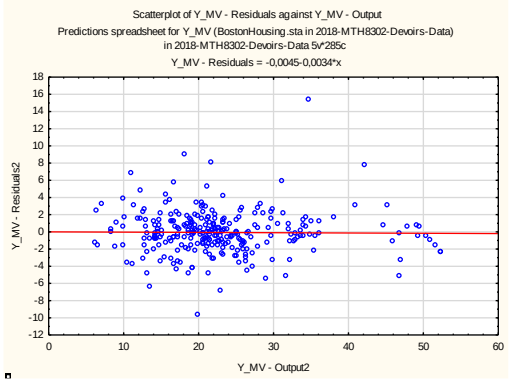


D'après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le deuxièmes réseau (R^2=0,9257) meilleur par rapport au premier (R^2=0,8761)









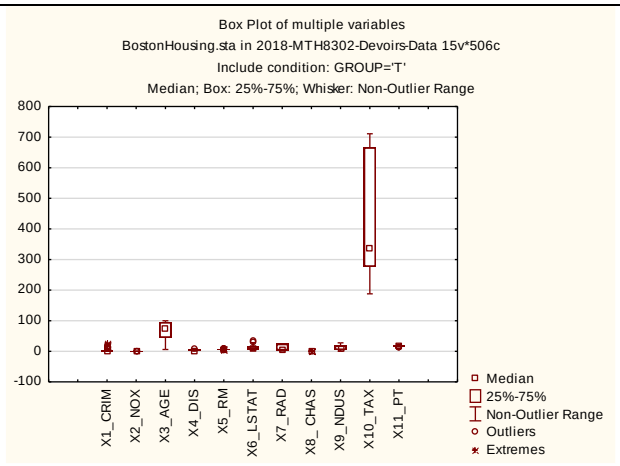
D'après les graphes nous avons confirmé notre remarque président (le second réseau est meilleure par rapport au premier)

8c) Régression MARS VS Régression ANN sur l'ensemble T

MARS sur l'ensemble T

Descriptive Statistics (BostonHousing
Include condition: GROUP='T'

Variable	N	Mean
X1_CRIM	101	3,2497
X2_NOX	101	0,5559
X3_AGE	101	67,6733
X4_DIS	101	3,6786
X5_RM	101	6,2680
X6_LSTAT	101	12,4788
X7_RAD	101	9,7129
X8_CHAS	101	0,0693
X9_NDUS	101	11,5345
X10_TAX	101	415,7723
X11_PT	101	18,6337



On remarque que la variable X10_Tax plus important par rapport les autres avec une avec une moyenne de 415,77 alors que les autres variables ont des moyennes variantes entre 0 et 68. Le graphique de Box&Whisker confirmer notre remarque

0 cases with missing data were found.

MARSplines Results:
Dependent: Y_MV
Independents: X1_CRIM, X2_NOX, X3_AGE, X4_DIS, X5_RM, X6_LSTAT, X7_RAD, X8_ CHAS, X9_NDUS, X10_TAX, X11_PT
Number of terms = 13
Number of basis functions = 18
Order of interactions = 2
Penalty = 2,000000
Threshold = 0,000500
GCV error = 7,510701
Prune = Yes

Après la régression MARS, les variables retenues dans le modèle (7 sur 11) sont X1_CRIM, X2_NOX, X4_DIS, X5_RM, X6_LSTAT, X10_TAX et X11_PT.

Number of References to Each Predictor (B
Number of times each predictor is reference
Include condition: GROUP='T'

Dependents	References (to Basis Functions)
X1_CRIM	1
X2_NOX	0
X3_AGE	1
X4_DIS	2
X5_RM	4
X6_LSTAT	5
X7_RAD	0
X8_CHAS	0
X9_NDUS	0
X10_TAX	2
X11_PT	3

Ce tableau nous donne le nombre le nombre de fois que la variable est présente dans l'expression de l'équation prédictive.

D'après le tableau les variables X6_LSTAT (5 fois) et X5_RM (4 fois) sont les plus référencées. Et les significatives.

Regression statistics (BostonHous
Include condition: GROUP='T'

Regression statistics	Y_MV
Mean (observed)	22,08020
Standard deviation (observed)	8,74232
Mean (predicted)	22,08020
Standard deviation (predicted)	8,49311
Mean (residual)	-0,00000
Standard deviation (residual)	2,07249
R-square	0,94380
R-square adjusted	0,93540

D'après le tableau on remarque que notre modèle prédit est très bon (R^2 = 94,38% et R^2ajusté = 93,5%), les moyennes des valeurs observées et prédites sont égales

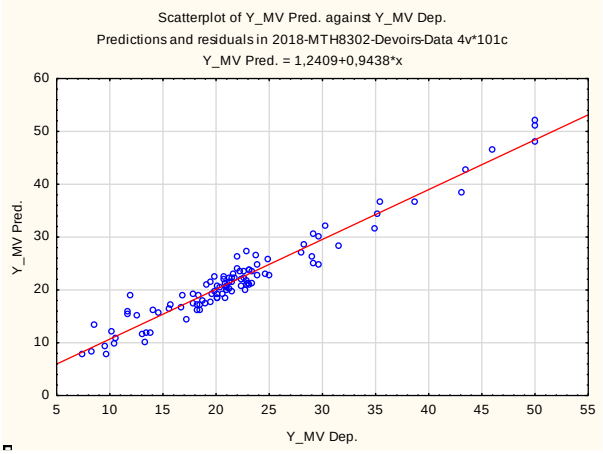
```
NOTE: The following model should be used directly, with categorical variables being
coded 0, 1.

Y_MV = 2,54037256676244e+001 - 7,95506249974916e-001*max(0; X6_LSTAT-
6,36000000000000e+000) + 1,42300380669044e+000*max(0; 6,36000000000000e+000-X6
_LSTAT) + 1,15336409271671e+001*max(0; X5_RM-6,43100000000000e+000) -
2,85795858334762e+000*max(0; 6,43100000000000e+000-X5_RM) - 9,34107769184320e-001
*max(0; X11_PT-1,52000000000000e+001) - 2,23974693394170e+000*max(0; X5_RM-
6,75000000000000e+000)*max(0; X11_PT-1,52000000000000e+001) + 5,72219536073686e-002
*max(0; 7,67202000000000e+000-X1_CRIM)*max(0; X6_LSTAT-6,36000000000000e+000) +
9,87470916512248e+000*max(0; 2,02180000000000e+000-X4_DIS)*max(0;
6,43100000000000e+000-X5_RM) - 1,86957957517049e-001*max(0; X4_DIS-
3,10250000000000e+000)*max(0; X6_LSTAT-6,36000000000000e+000) + 3,57070414142405e-
002*max(0; 2,87000000000000e+002-X10_TAX) + 1,11264870553806e-002*max(0;
9,54000000000000e+001-X3_AGE)*max(0; X6_LSTAT-6,36000000000000e+000) +
7,77464868808302e-003*max(0; X10_TAX-2,87000000000000e+002)*max(0; X11_PT-
1,92000000000000e+001)
```

L'équation prédictive du modèle avec les variables indépendantes retenues

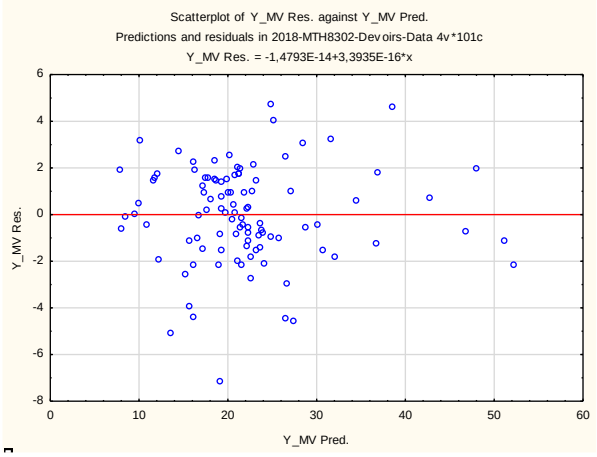
Model coefficients (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
NOTE: Highlighted cells indicate basis functions of type
max(0, independent-knot), otherwise max(0, knot-independent)
Include condition: GROUP='T'

Coefficients, knots and basis functions	Coefficients Y_MV	Knots X1_CRIM	Knots X2_NOX	Knots X3_AGE	Knots X4_DIS	Knots X5_RM	Knots X6_LSTAT	Knots X7_RAD	Knots X8_CHAS	Knots X9_NDUS	Knots X10_TAX	Knots X11_PT
Intercept	25,40373											
Term.1	-0,79551						6,360000					
Term.2	1,42300						6,360000					
Term.3	11,53364					6,431000						
Term.4	-2,85796					6,431000						
Term.5	-0,93411											15,20000
Term.6	-2,23975					6,750000						15,20000
Term.7	0,05722	7,672020					6,360000					
Term.8	9,87471				2,021800	6,431000						
Term.9	-0,18696				3,102500		6,360000					
Term.10	0,03571										287,0000	
Term.11	0,01113			95,40000			6,360000					
Term.12	0,00777										287,0000	19,20000

Ce tableau nous donne les nœuds employés dans le modèle



On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l'équation prédictive linéaire du modèle avec les valeurs observées)



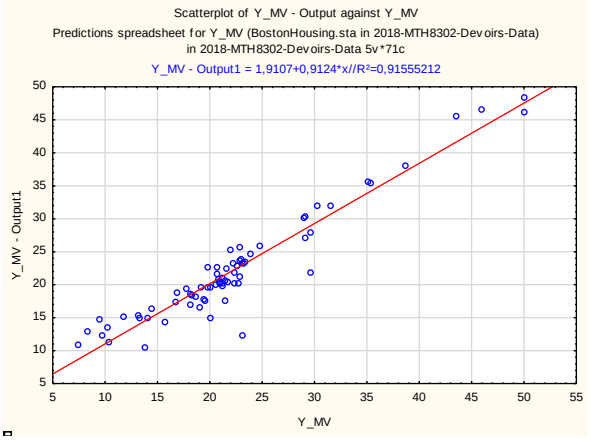
D'après le graphe de l'analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MARS est globalement bon.

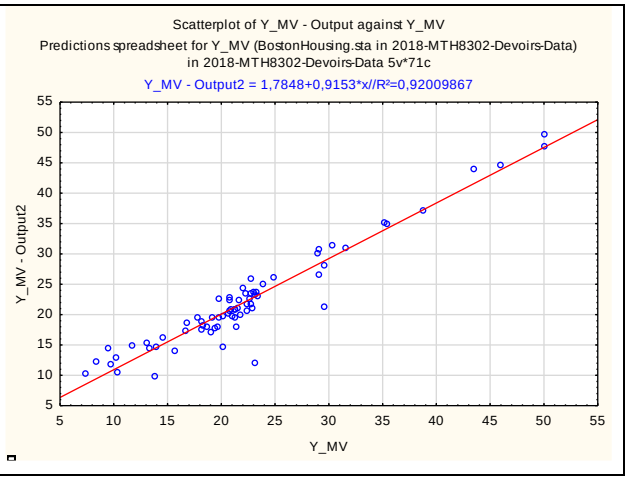
Réseaux de neurones sur l'ensemble T

Summary of active networks (BostonHousing.sta in 2018-MTH8302-Devoirs-Data)
Include condition: GROUP='T'

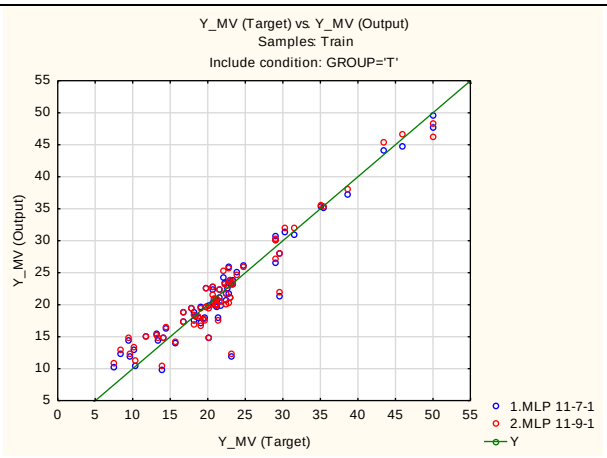
Index	Net. name	Training perf.	Test perf.	Validation perf.	Training error	Test error	Validation error	Training algorithm	Error function	Hidden activation	Output activation
1	MLP 11-7-1	0,959218	0,784520	0,976138	3,006607	4,151159	6,212299	BFGS 24	SOS	Logistic	Exponential
2	MLP 11-9-1	0,956845	0,767139	0,970682	3,172268	4,355951	4,984129	BFGS 16	SOS	Tanh	Exponential

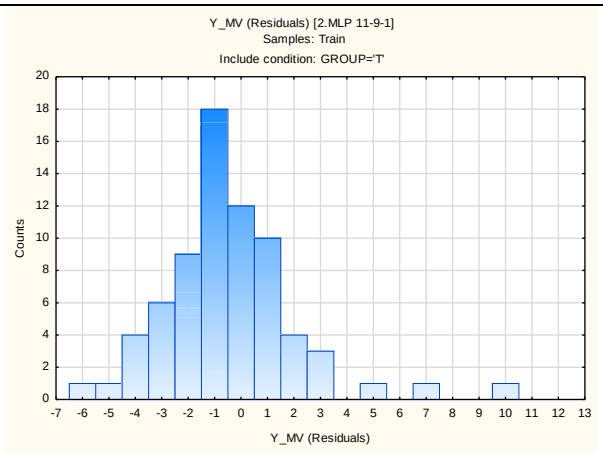
Ce tableau nous donne les informations sur les deux meilleurs réseaux retenus dans le modéle

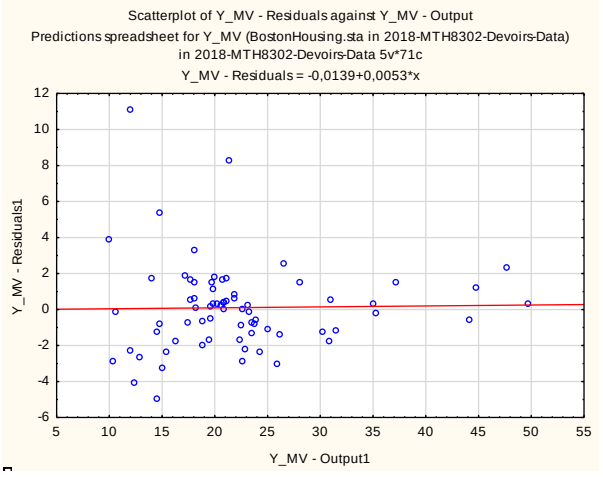


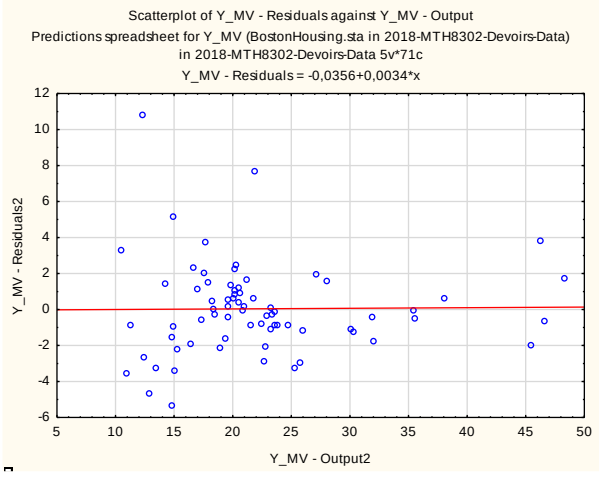


D'après les deux graphes (les équations prédictives linéaires de deux modèles retenus avec les valeurs observées) on remarque que le deuxièmes réseau (R^2=0,9200) meilleur par rapport au premier (R^2=0,9155)









D'après les graphes nous avons confirmé notre remarque président (le second réseau est meilleure par rapport au premier)

Comparaison MARS et Réseaux de neurones

MARS : pouvoir prédictif excellent (R^2 = 94,38% et $R^2_{ajusté}$ = 93,5%) et analyse des résidus excellente sur les variables prédites (aucun résidu supérieur à 6 en valeur absolue, 93 sur 101 ont des résidus inférieurs à 4 en valeur absolue.

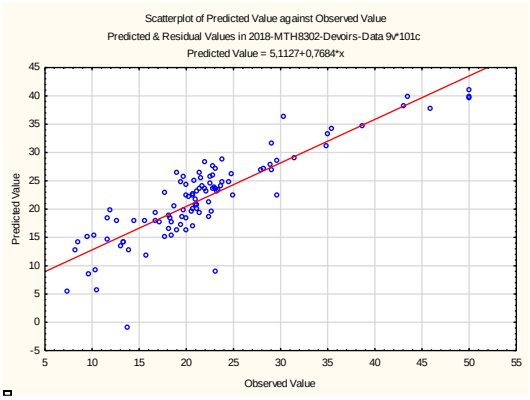
Réseaux de neurones avec le meilleur réseau : pouvoir prédictif excellent (R^2 = 92%) et analyse des résidus très bonne sur les variables prédites (aucun résidu supérieur à 12 en valeur absolue, 99 sur 101 ont des résidus inférieurs à 6 en valeur absolue.

➔D'après les deux on conclure le modèle de MARS est le meilleur choix par rapport au Réseaux de neurones pour ces données

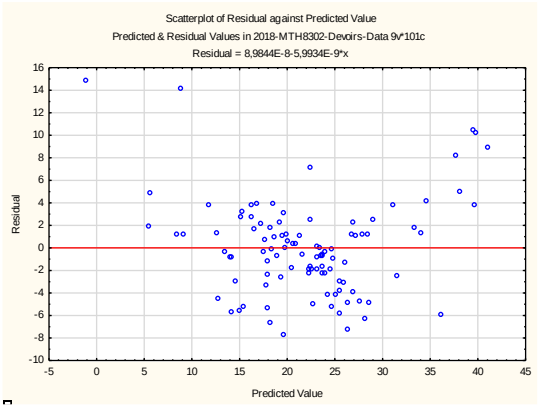
(*Backward Stepwise*) (MRB) vs MARS et ANN

Var	Nom	coefficient	MRB sélection arrière
X0	GENÉRAL intercepte	b0	27,50599
X1	CRIM	b1	
X2	NOX	b2	
X3	AGE	b3	
X4	DIS	b4	
X5	RM	b5	4,35412
X6	LSTAT	b6	-0,45228
X7	RAD	b7	
X8	CHAS	b8	
X9	INDUS	b9	
X10	TAX	b10	
X11	PT	b11	-1,45293
		SS resid résiduelle	1769,708
		MSE = sigma^2 (ANOVA)	18,244
		R^2	0,76844831

On remarque que la méthode de Régression avec Backward Stepwise a éliminé les variables qui ne sont pas significatives



On remarque que la plupart des valeurs prédites sont proches de la droite obtenue (l'équation prédictive linéaire du modèle avec les valeurs observées)



D'après le graphe de l'analyse des résidus avec les valeurs prédites on remarque que le modèle obtenu avec la régression MRB est globalement bon.

		R^2ajusté	0, 76128691
		F	107,30

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

MRB : pouvoir prédictif bon (R^2 = 76,8% et R^2ajusté = 76,1%) et analyse des résidus quand même bonne sur les variables prédites,97 sur 101 ont des résidus inférieurs à 10 en valeur absolue.

8d)

Comparasion MRB(*Backward Stepwise*) MARS et Réseaux de neurones

Modèle	R^2	Prédiction	Analyse des résidus	Complexité du modéle
MRB	76,8%	Bon	Moyenne bon	Facile (les étapes et la compréhension de résultats de modèle)
MARS	94,38%	Excellent	Excellent	Moyenne (les étapes et la compréhension de résultats de modèle un peu facile)
Réseaux de neurones	92%	Excellent	Très bonne	Étapes moyenne et la compréhension de résultats est difficile

D'après le tableau on remarque que MARS est meilleur

8e)

	forces	faiblesses
MRB	-résous le problème de multicolinéarité (élimine les variables explicatives les moins importantes dans les données) -les résultats donnent directement l'imporance relative des variables d'entrée sur le modèle a partir de leurs effets significatifs -le meilleur choix par rapport MRO et MRF -Facile a applique	-Pouvoir prédictif et analyse des résidus faible par rapport MARS et Réseaux de neurones -résultat peut etre biaisés (résultats aberrants, manque de robustesse..) pour une très grande quantité des données (manque des données…). -Incapable de découvrir la structure locale des données
MARS	-Facilement a utilisé pour exploiter une très grande quantité des données complexe -Capable de découvrir la structure locale des données -Pouvoir prédictif et analyse des résidus très bon par rapport à la régression multiple	-les étapes et la compréhension de résultats de modèle sont un peu difficile par rapport MRB -méthode un peu complexe par rapport MRB

Réseaux de neurones	**-Facilement a utilisé pour exploiter une très grande quantité des données complexe -Pouvoir prédictif et analyse des résidus très bon par rapport à la régression multiple**	**-les étapes et la compréhension de résultats de modèle sont difficile par rapport MRB et MARS -méthode un peu complexe par rapport MRB**

8f) conclusion

Le processus de modélisation statistique à l'aide de modèles de régression incluant la méthode les réseaux de neurones serait définir le problème et identifier et préparer es données, et nous permit de construire, tester et évaluer le modèle en choisissant la technique la plus optimale et trouver le meilleur réseau. Et pour aide au traitement la méthode de MARS nous permit a exploiter une très grande quantité des données complexe et utilise l'algorithme itératif adaptif utile pour aider le réseau à effectuer le traitement.

No 9 Modèles d'analyse de la variance

Données = Amphétamine.sta **voir 2018-MTH8302-Devoirs-data.stw**

Réponse

9a)

	nature	rôle
Étude 1	**XB_vitesse : Qualitatif** **XC_dose :Quantitatif (fixe)/ catégorique**	**XB_vitesse : Facteur Inter** **XC_dose : Facteur Intra**
Étude1 et Étude2 combinées	**XA_levier : Quantitatif (fixe)/ catégorique** **XB_vitesse : Qualitatif** **XC_dose : Quantitatif (fixe)/ catégorique**	**XA_levier : Facteur Inter** **XB_vitesse : Facteur Inter** **XC_dose : Facteur Intra**

9b)
les souris initialement sont classées en catégories de vitesse pour étudie l'évolution des souries en fonction de vitesse et pour faciliter le regroupement et les calcules des moyenne et aussi pour limiter les coûts (utiliser les statistiques pour prédit une résultat pour une expérience).

9c)
Les souris reçoit la dose d'amphétamine dans un ordre dicté par le hasard pour réalisé une expérience qui approcher a la réalité
Et on sait que dans l'expérimentation les modalités sont affectés au hasard aux unités

9d)

Pour faire l'analyse de l'étude 1 on utilise le modèles d'analyse de la variance avec la Méthode approche à mesures répétées
Au début pour faire l'analyse avec la méthode approche à mesures répétées on doit d'abord organiser nos données :

vitesse	Y_0,0	Y_0,5	Y_1,0	Y_1,8
lente	0,81	0,80	0,82	0,50
lente	0,77	0,78	0,79	0,51
lente	0,80	0,82	0,83	0,52
lente	0,95	0,95	0,91	0,60
moyenne	1,03	1,13	1,04	0,82
moyenne	0,96	0,93	1,02	0,63
moyenne	0,98	1,00	0,98	0,74
moyenne	1,17	1,20	1,18	0,91
vite	1,20	1,24	1,27	0,96
vite	1,25	1,23	1,30	1,01
vite	1,23	1,20	1,18	0,95
vite	1,31	1,42	1,41	1,08
lente	0,84	0,85	0,88	0,58
lente	0,72	0,73	0,74	0,42
lente	0,73	0,76	0,75	0,48
lente	0,89	0,90	0,97	0,67

moyenne	1,11	1,02	1,12	0,75
moyenne	1,01	1,05	0,95	0,72
moyenne	1,05	1,07	1,05	0,79
moyenne	1,12	1,13	1,11	0,83
vite	1,28	1,17	1,21	0,91
vite	1,21	1,31	1,22	0,93
vite	1,16	1,15	1,23	1,02
vite	1,40	1,33	1,35	1,20

Exécution de l'analyse et présention des résultats

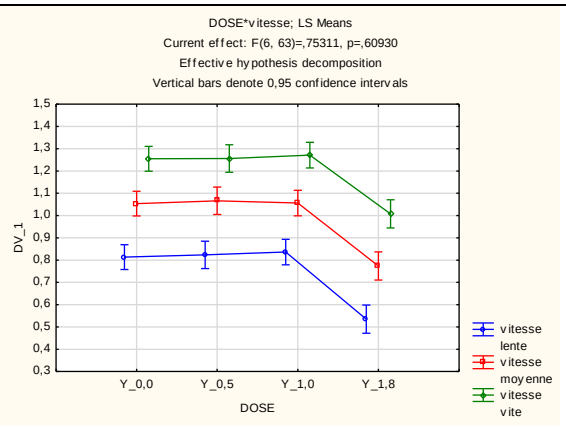
Repeated Measures Analysis of Variance (Amphetamine.sta in
Sigma-restricted parameterization
Effective hypothesis decomposition
Include condition: Étude='étude1'

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	92,02208	1	92,02208	4187,609	0,000000
vitesse	3,17627	2	1,58813	72,271	0,000000
Error	0,46147	21	0,02197		
DOSE	1,37318	3	0,45773	321,603	0,000000
DOSE*vitesse	0,00643	6	0,00107	0,753	0,609301
Error	0,08967	63	0,00142		

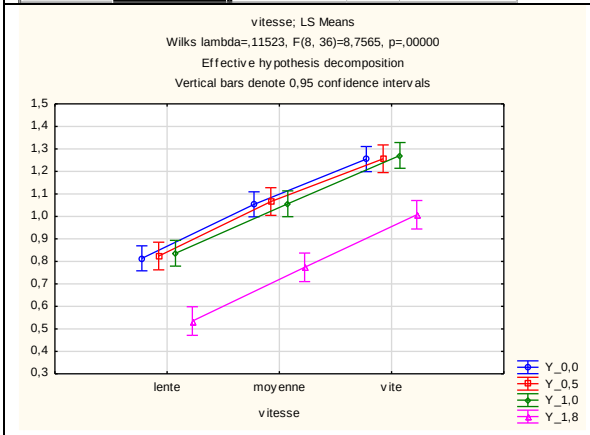
On remarque que le facteur vitesse et DOSE sont significatifs
Et l'interaction DOSE*vitesse est non significatif

Mauchley Sphericity Test (Amphetamine.sta
Sigma-restricted parameterization
Effective hypothesis decomposition

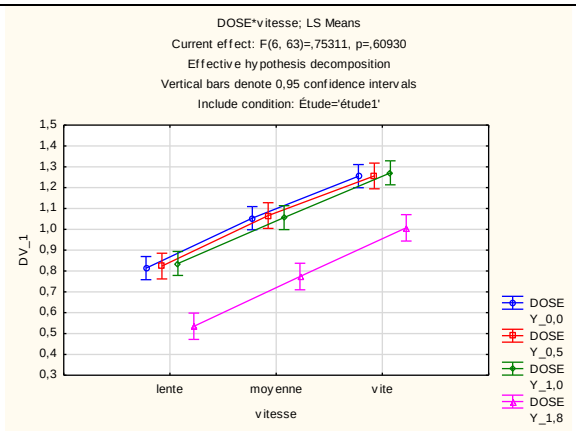
Effect	W	Chi-Sqr.	df	p
DOSE	0,948018	1,052806	5	0,958195



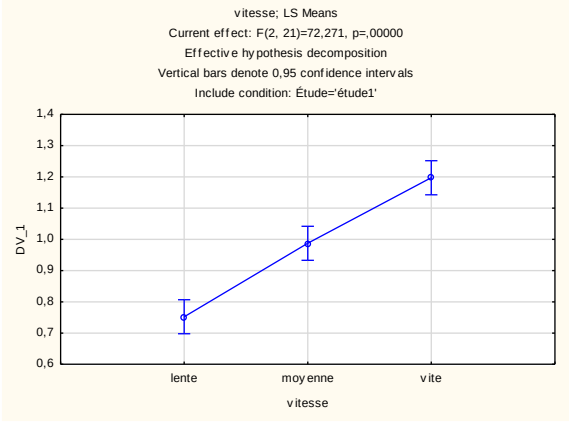
On remarque que l'interaction DOSE*vitesse n'est non significatif



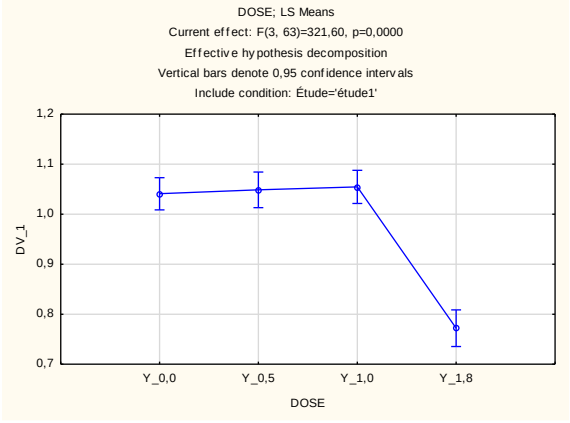
L'effet de significatif de vitesse



Pas l'effet de significatif de vitesse et DOSE



On remarque que la vitesse vite plus effet par rapport DOSE



L'effet de significatif de DOSE

Parameter Estimates (Amphetamine.sta in 2018-MTH8302-Devoirs-Data)
Sigma-restricted parameterization
Include condition: Étude='étude1'

Effect	Level of Effect	Column	Y_0,0 Param.	Y_0,0 Std.Err	Y_0,0 t	Y_0,0 p
Intercept		1	1,040833	0,015442	67,4046	0,000000
vitesse	lente	2	-0,227083	0,021838	-10,3987	0,000000
vitesse	moyenne	3	0,012917	0,021838	0,5915	0,560509

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization
Include condition: Étude='étude1'

Effect	Y_0,5 Param.	Y_0,5 Std.Err	Y_0,5 t	Y_0,5 p
Intercept	1,048750	0,017088	61,37460	0,000000
vitesse	-0,225000	0,024166	-9,31074	0,000000
vitesse	0,017500	0,024166	0,72417	0,476953

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization
Include condition: Étude='étude1'

Effect	Y_1,0 Param.	Y_1,0 Std.Err	Y_1,0 t	Y_1,0 p
Intercept	1,054583	0,015940	66,15843	0,000000
vitesse	-0,218333	0,022543	-9,68522	0,000000
vitesse	0,001667	0,022543	0,07393	0,941763

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization
Include condition: Étude='étude1'

Effect	Y_1,8 Param.	Y_1,8 Std.Err	Y_1,8 t	Y_1,8 p
Intercept	0,772083	0,017578	43,92207	0,000000
vitesse	-0,237083	0,024860	-9,53684	0,000000
vitesse	0,001667	0,024860	0,06704	0,947182

D'après ce tableau on peut trouver nous Modéles :

Y0,0=1,04-0,22*lente+0,0129*moyenne	**Y_0,5=1,048-0,22*lente+0,017* moyenne**

Y1,0=1,05-0,218*lente+0,0016*moyenne	Y1,0=0,77-0,23*lente+0,0016*moyenne

On remarque que la variable moyenne de vitesse est non significative et on remarque aussi que la variable vite de vitesse n'apparaître pas dans notre modéle

Test of SS Whole Model vs. SS Residual (Amphetamine.sta in 2018-MTH8302-Devoirs-Data)

Dependent Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Y_0,0	0,930923	0,866618	0,853915	0,780808	2	0,390404	0,120175	21	0,005723	68,22124	0,000000
Y_0,5	0,914503	0,836316	0,820727	0,751900	2	0,375950	0,147163	21	0,007008	53,64784	0,000000
Y_1,0	0,924822	0,855296	0,841515	0,756933	2	0,378467	0,128062	21	0,006098	62,06188	0,000000
Y_1,8	0,922772	0,851508	0,837366	0,893058	2	0,446529	0,155737	21	0,007416	60,21101	0,000000

D'après les tableaux et les graphes on remarque que le facteur vitesse et DOSE sont significatifs
Et l'interaction DOSE*vitesse est non significatif (p_valeu), et on remarque que la variable vite de vitesse n'apparaître pas dans notre modèle
Les données répétées sont considérées comme plusieurs variables de réponse dépendantes ; dans ce cas il est préférable d'organiser les données avec autant de variables de réponse Y qu'il y a de mesures répétées.
La variabilité inter sujet est exclue de l'erreur expérimentale, donc il est plus facile de comparer les traitements entre eux
Chaque sujet sert comme son propre contrôle
Économie du nombre de sujets

9e) Répondez aux mêmes questions que 9d) pour les 2 études combinées.
au début pour faire l'analyse avec la méthode approche à mesures répétées on doit d'abord organiser nos données :

XB_vitesse	XA_levier	Y_0,0	Y_0,5	Y_1,0	Y_1,8
lente	2	0,81	0,80	0,82	0,50
lente	2	0,77	0,78	0,79	0,51
lente	2	0,80	0,82	0,83	0,52
lente	2	0,95	0,95	0,91	0,60
moyenne	2	1,03	1,13	1,04	0,82
moyenne	2	0,96	0,93	1,02	0,63
moyenne	2	0,98	1,00	0,98	0,74
moyenne	2	1,17	1,20	1,18	0,91
vite	2	1,20	1,24	1,27	0,96
vite	2	1,25	1,23	1,30	1,01
vite	2	1,23	1,20	1,18	0,95
vite	2	1,31	1,42	1,41	1,08
lente	2	0,84	0,85	0,88	0,58
lente	2	0,72	0,73	0,74	0,42
lente	2	0,73	0,76	0,75	0,48
lente	2	0,89	0,90	0,97	0,67
moyenne	2	1,11	1,02	1,12	0,75
moyenne	2	1,01	1,05	0,95	0,72
moyenne	2	1,05	1,07	1,05	0,79
moyenne	2	1,12	1,13	1,11	0,83
vite	2	1,28	1,17	1,21	0,91
vite	2	1,21	1,31	1,22	0,93
vite	2	1,16	1,15	1,23	1,02
vite	2	1,40	1,33	1,35	1,20

lente	5	2,18	2,44	1,92	0,92
lente	5	2,02	2,20	1,75	0,82
lente	5	2,06	2,28	1,86	0,80
lente	5	2,28	2,46	1,90	0,90
moyenne	5	2,62	2,58	2,21	1,03
moyenne	5	2,60	2,60	2,34	1,14
moyenne	5	2,39	2,41	2,09	0,90
moyenne	5	2,70	2,64	2,23	1,02
vite	5	2,98	2,64	2,34	1,28
vite	5	3,10	2,85	2,40	1,35
vite	5	2,80	2,48	2,16	1,01
vite	5	3,21	2,92	2,56	1,40
lente	5	2,26	2,40	1,99	0,99
lente	5	1,96	2,18	1,81	0,78
lente	5	2,10	2,24	1,92	0,88
lente	5	2,35	2,49	1,95	0,96
moyenne	5	2,68	2,64	2,17	0,96
moyenne	5	2,66	2,62	2,28	1,23
moyenne	5	2,43	2,48	2,16	0,84
moyenne	5	2,66	2,70	2,27	0,98
vite	5	2,94	2,70	2,44	1,33
vite	5	3,20	2,91	2,45	1,39
vite	5	2,84	2,53	2,23	1,07
vite	5	3,31	2,98	2,47	1,51

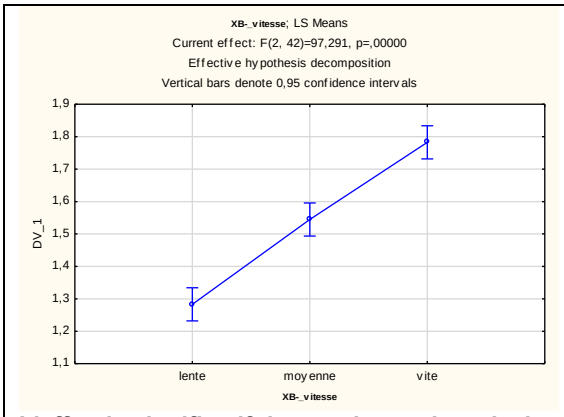
Repeated Measures Analysis of Variance (Amphetamine.sta in
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	453,5011	1	453,5011	11036,05	0,000000
XA-_levier	59,7417	1	59,7417	1453,83	0,000000
XB-_vitesse	7,9959	2	3,9979	97,29	0,000000
XA-_levier*XB-_vitesse	0,0946	2	0,0473	1,15	0,325890
Error	1,7259	42	0,0411		
DOSE_ET1	25,9021	3	8,6340	3844,04	0,000000
DOSE_ET1*XA-_levier	12,3610	3	4,1203	1834,46	0,000000
DOSE_ET1*XB-_vitesse	0,3517	6	0,0586	26,09	0,000000
DOSE_ET1*XA-_levier*XB-_vitesse	0,3704	6	0,0617	27,48	0,000000
Error	0,2830	126	0,0022		

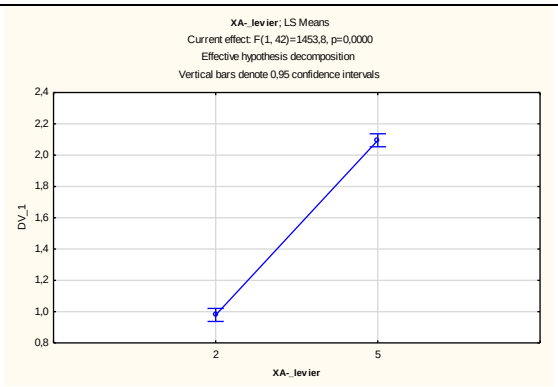
Mauchley Sphericity Test (Amphetamine.sta
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	W	Chi-Sqr.	df	p
DOSE	0,686060	15,34372	5	0,008990

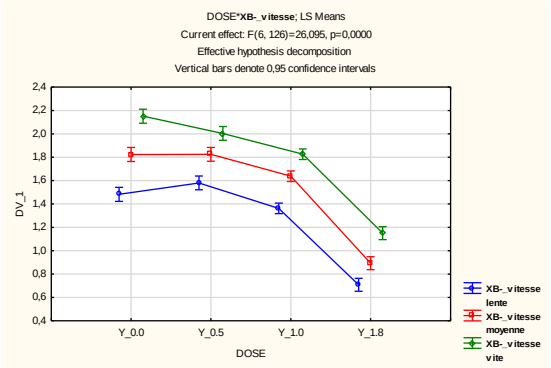
On remarque que les facteurs XB_vitesse, XA_levier, et DOSE et l'interaction DoSE*XA_levier , , DoSE*XB_vetesse, DoSE*XA_levier*XB_vetesse sont significatifs
Et l'interaction XA_levier*XB_vitesse sont non significatif



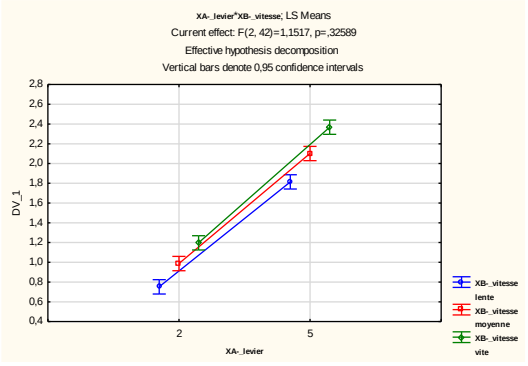
L'effet de significatif de XB_vitesse (p_valeu)



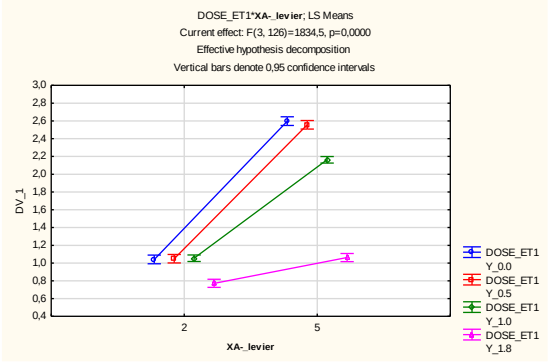
L'effet de significatif de XAlevier (p)



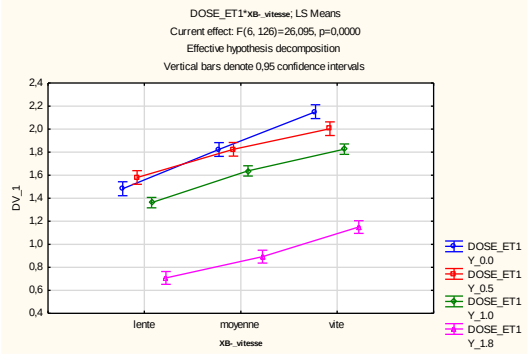
L'effet de significatif de DOSE*XB_vitesse (p_valeu)



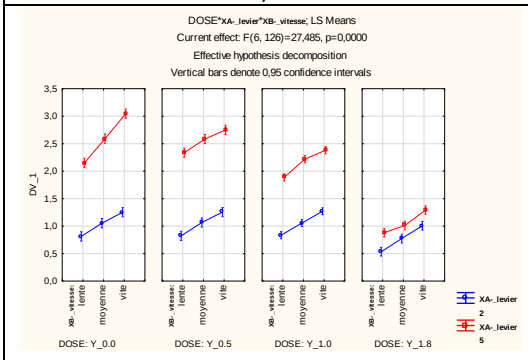
L'effet de significatif de XA_levier*XB_vitesse (p_valeu)

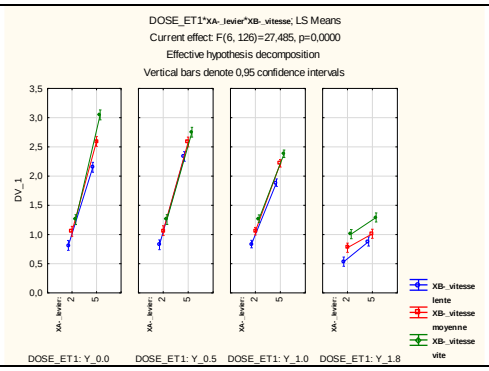


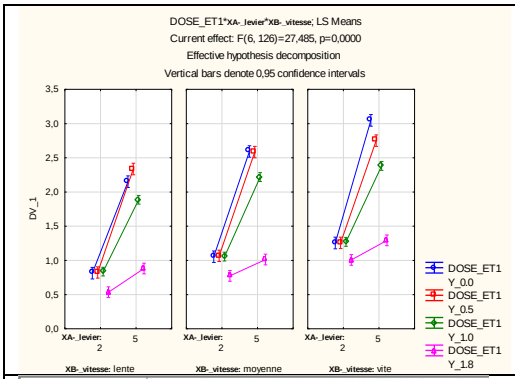
On remarque que la XA_levier moins important dans le modèle du DOSE 1,8



On remarque que la XA_vitesse moins important dans le modèle du DOSE 1,8







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

Effect	Level of Effect	Column	Y_0.0 Param.	Y_0.0 Std.Err	Y_0.0 t	Y_0.0 p
Intercept		1	1,818958	0,017136	106,1502	0,000000
XA-_levier	2	2	-0,778125	0,017136	-45,4096	0,000000
XB-_vitesse	lente	3	-0,336458	0,024234	-13,8840	0,000000
XB-_vitesse	moyenne	4	0,004167	0,024234	0,1719	0,864312
XA-_levier*XB-_vitesse	1	5	0,109375	0,024234	4,5134	0,000051
XA-_levier*XB-_vitesse	2	6	0,008750	0,024234	0,3611	0,719857

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization

Effect	Y_0.5 Param.	Y_0.5 Std.Err	Y_0.5 t	Y_0.5 p
Intercept	1,802917	0,016845	107,0319	0,000000
XA-_levier	-0,754167	0,016845	-44,7719	0,000000
XB-_vitesse	-0,222917	0,023822	-9,3576	0,000000
XB-_vitesse	0,022083	0,023822	0,9270	0,359215
XA-_levier*XB-_vitesse	-0,002083	0,023822	-0,0875	0,930726
XA-_levier*XB-_vitesse	-0,004583	0,023822	-0,1924	0,848356

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization

Effect	Y_1.0 Param.	Y_1.0 Std.Err	Y_1.0 t	Y_1.0 p
Intercept	1,608542	0,012910	124,5994	0,000000
XA-_levier	-0,553958	0,012910	-42,9102	0,000000
XB-_vitesse	-0,246667	0,018257	-13,5107	0,000000
XB-_vitesse	0,028958	0,018257	1,5861	0,120208
XA-_levier*XB-_vitesse	0,028333	0,018257	1,5519	0,128187
XA-_levier*XB-_vitesse	-0,027292	0,018257	-1,4949	0,142429

Parameter Estimates (Amphetamine.sta in 2018-M
Sigma-restricted parameterization

Effect	Y_1.8 Param.	Y_1.8 Std.Err	Y_1.8 t	Y_1.8 p
Intercept	0,917083	0,015888	57,72011	0,000000
XA-_levier	-0,145000	0,015888	-9,12612	0,000000
XB-_vitesse	-0,208958	0,022470	-9,29957	0,000000
XB-_vitesse	-0,023958	0,022470	-1,06625	0,292401
XA-_levier*XB-_vitesse	-0,028125	0,022470	-1,25169	0,217610
XA-_levier*XB-_vitesse	0,025625	0,022470	1,14043	0,260570

D'après ce tableau on peut trouver nous Modéles :

Y0,0=1,81-0,77*levier-0,33*vitesse(lente)+0,0041*vitesse(moyenne)+0,10* levier*vtesse+0,008* levier*vtesse	**Y0,5=1,80-0,75*levier-0,22*vitesse(lente)+0,022*vitesse(moyenne)-0,002* levier*vtesse-0,004* levier*vtesse**
Y1,0=1,60-0,55*levier-0,24*vitesse(lente)+0,028*vitesse(moyenne)+0,028* levier*vtesse-0,02* levier*vtesse	**Y1,8=0,91-0,14*levier-0,20*vitesse(lente)-0,02*vitesse(moyenne)-0,028* levier*vtesse-0,025* levier*vtesse**

No 10 Modélisation statistique

Données = Assurances.sta voir 2018-MTH8302-Devoirs-data.stw

Réponse

10a) L'analyse statistique des données peut se faire selon plusieurs modèles. Proposer 4 modèles statistiques sans effet d'interaction que l'on peut considérer pour faire l'analyse. Présenter vos modèles en complétant le tableau

Modèle	Nom statistique (*)	Définir le rôle de chacune des variables impliquées
M1	Multiple regression	-Continu : age / nombre interventions&procédures /nombre autres maladies/nombre médicaments prescrits/nombre visites unités soins intensifs/nombre complications/durée traitement (jr) -Dependent (continu) : Y_coûtTotal
M2	Factorial ANOVA	-catégorique (facteur) :genre/ v4_record / v6_record -Dependent (continu) : Y_coûtTotal
M3	One-way ANOVA	-catégorique (facteur) : genre / v4_record / v6_record -Dependent (continu) : Y_coûtTotal
M4	Analysis of covariance (ANCOVA)	-Qualitatif : genre/ v4_record / v6_record -Quantitatif (Continu) : age / nombre interventions&procédures /nombre autres maladies/nombre médicaments prescrits/nombre visites unités soins intensifs/nombre complications/durée traitement (jr) -Dependent (continu) : Y_coûtTotal

10b)

Multiple regression

Regression Summary for Dependent Variable: Y_coûtTotal (Assurances.sta i
R= ,74126362 R²= ,54947175 Adjusted R²= ,54542855
F(7,780)=135,90 p<0,0000 Std.Error of estimate: 4510,7

N=788	b*	Std.Err. of b*	b	Std.Err. of b	t(780)	p-value
Intercept			429,699	1423,149	0,30194	0,762782
age	-0,047086	0,024417	-46,640	24,186	-1,92843	0,054164
nombre interventions&procédures	0,669752	0,026740	800,909	31,976	25,04704	0,000000
nombre autres maladies	0,049824	0,027877	56,014	31,340	1,78726	0,074283
nombre médicaments prescrits	-0,055507	0,028509	-349,060	179,279	-1,94702	0,051890
nombre visites unités soins intensifs	0,158971	0,029808	403,248	75,613	5,33307	0,000000
nombre complications	0,031358	0,024728	845,703	666,900	1,26811	0,205137
durée traitement (jr)	-0,016018	0,028411	-0,886	1,572	-0,56382	0,573040

Analysis of Variance; DV: Y_coûtTotal (Assurances.sta in 2018-MT

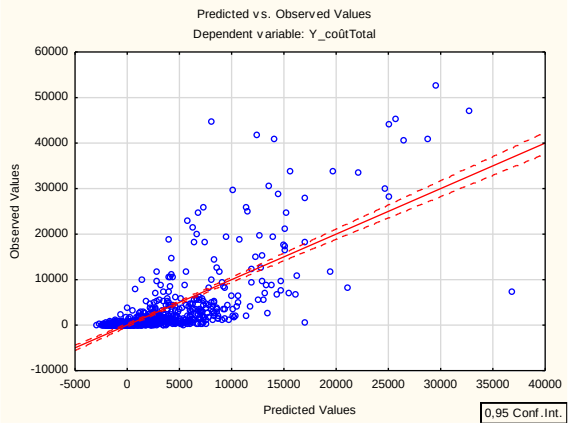
Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	1,935558E+10	7	2,765083E+09	135,9001	0,00
Residual	1,587022E+10	780	2,034643E+07		
Total	3,522579E+10				

D'après R^2(54,9%) on remarque que on n'a pas une bonne corrélation entre le modèle prédit et réel.

D'après les P-value on remarque que tous les variables Xi ne sont pas significatives sauf les variables Nombre intervention&procédures et nombre visites unités soins intensifs qui sont Significatives. (Variables Significatives p-level≤ 0,05)

Variables currently in the Equation; DV: Y_coûtTotal (Assurances.sta in 2018-MTH8302-Devoir

Variable	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(780)	p-value
age	-0,047086	-0,068885	-0,046347	0,968857	0,031143	-1,92843	0,054164
nombre interventions&procédures	0,669752	0,667659	0,601963	0,807815	0,192185	25,04704	0,000000
nombre autres maladies	0,049824	0,063864	0,042954	0,743235	0,256765	1,78726	0,074283
nombre médicaments prescrits	-0,055507	-0,069546	-0,046793	0,710671	0,289329	-1,94702	0,051890
nombre visites unités soins intensifs	0,158971	0,187565	0,128171	0,650052	0,349948	5,33307	0,000000
nombre complications	0,031358	0,045359	0,030477	0,944605	0,055395	1,26811	0,205137
durée traitement (jr)	-0,016018	-0,020184	-0,013550	0,715598	0,284402	-0,56382	0,573040



D'après le graphe on remarque que on n'a pas un bon modèle

Factorial ANOVA

Univariate Tests of Significance for Y_coûtTotal (Assurances.sta in 2018-M
Sigma-restricted parameterization
Effective hypothesis decomposition

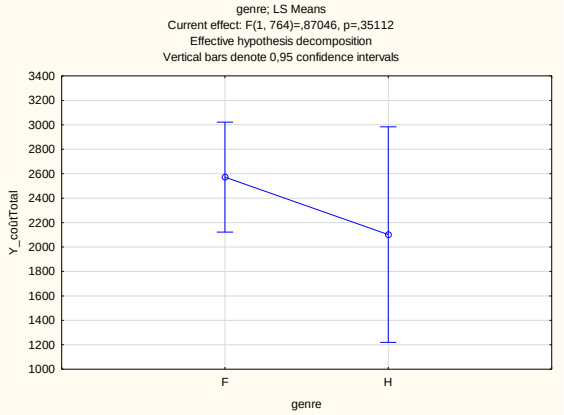
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	2,642963E+09	1	2,642963E+09	85,92807	0,000000
genre	2,677343E+07	1	2,677343E+07	0,87046	0,351123
v4_recod	6,234362E+09	5	1,246872E+09	40,53834	0,000000
v6_recod	1,472324E+08	1	1,472324E+08	4,78682	0,028981
genre*v4_recod	8,326301E+07	5	1,665260E+07	0,54141	0,744961
genre*v6_recod	8,948836E+06	1	8,948836E+06	0,29094	0,589773
v4_recod*v6_recod	3,243376E+08	5	6,486751E+07	2,10897	0,062406
genre*v4_recod*v6_recod	1,573121E+07	5	3,146242E+06	0,10229	0,991667
Error	2,349900E+10	764	3,075786E+07		

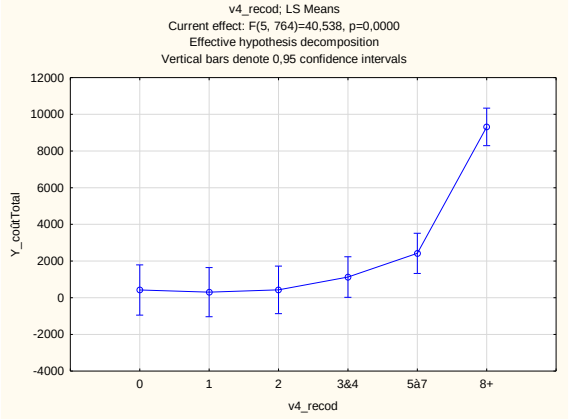
Test of SS Whole Model vs. SS Residual (Assurances.sta in 2018-MTH8302-Devoirs-Data)

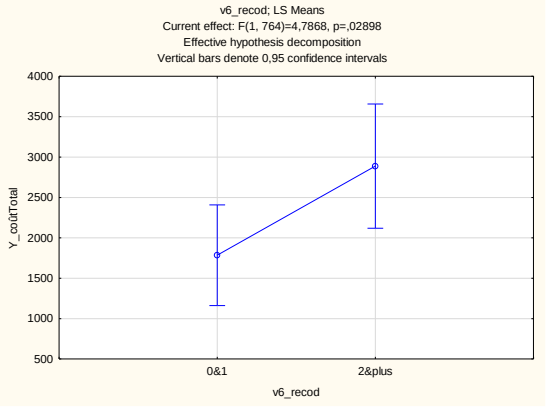
Dependent Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Y_coûtTotal	0,576978	0,332904	0,312821	1,172679E+10	23	509860513	2,349900E+10	764	30757856	16,57659	0,00

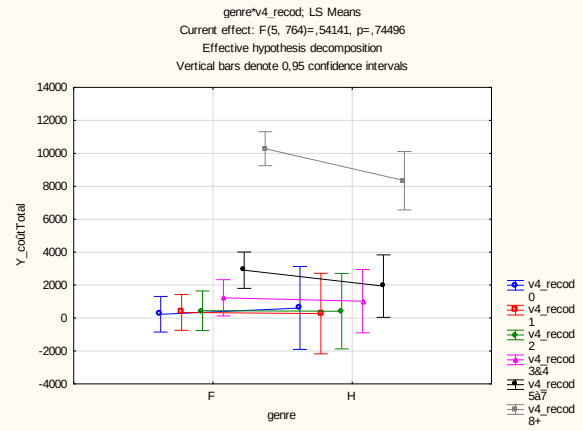
D'après R^2(33,29%) on remarque que on a une mauvaise corrélation entre le modèle prédit et réel.

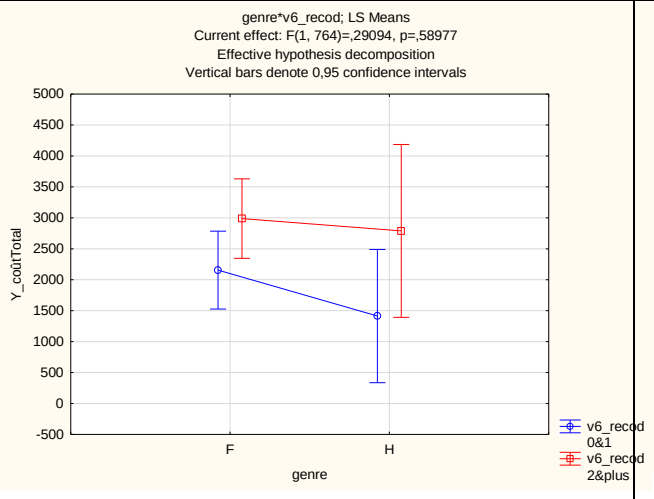
D'après les P-value on remarque que tous les variables Xi ne sont pas significatives sauf les variables v4_record et v6_record qui sont Significatives. (Variables Significatives p-level≤ 0,05)

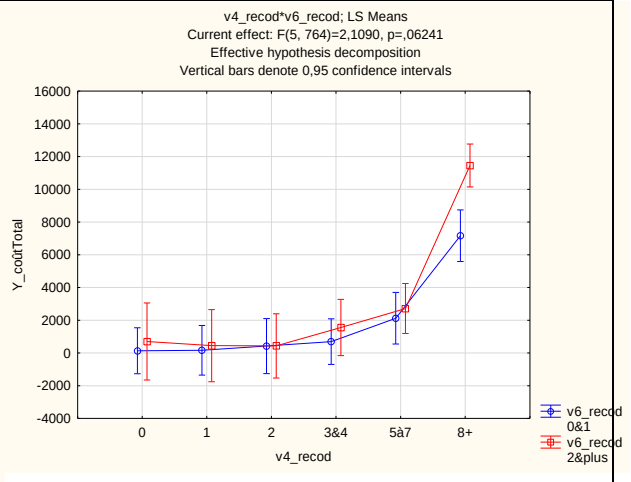


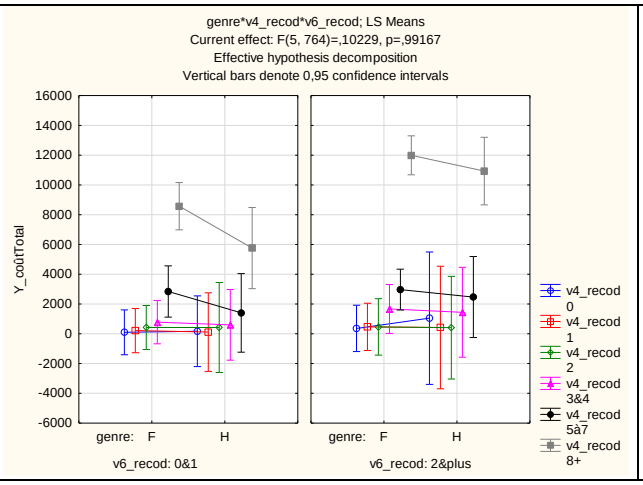












D'après les graphes on remarque que 8+ de la variable V4_recod prend les grandres valeurs de Y_coutTotal

Avec ces modéles on peut trouver les variables qui sont la plus effet sur notre réponse

<u>Remrque :</u> nous avons aussi éssayer d'analyse juste avec les facteurs qui sont significatifs et nous avons obtenu presque les mémes résultats (R^2)

10c) Comparer les résultats de 2 modèles. Y – a-t-il des différences d'interprétations?

	R^2	F
Multiple regression	**54,94%**	**135,9**
Factorial ANOVA	**33,29%**	**16,57**

D'après le tableau on remarque que la méthode Multiple regression ($R^2 = 54,94$) et F(135,9) meilleure par rapport Factorial ANOVA si on veut trouver un modèle prédit, mais pour analyser la signification des données la méthode ANOVA (Factorial ANOVA) est la meilleure

10d)
Oui, la variable de réponse devrait être transformée, puisque si on veux faciliter l'analyse et trouver un bon effet on transformée la réponse continu au variable catégorique (pour minimiser la taille des données et faciliter l'analyse).

<u>Conclusion générale</u>
Dans cette devoir nous avons appris et étudie la modélisation avec MARS et réseaux de neurones et nous avons fait la comparasion avec eux et MRB(*Backward Stepwise*), et on a trouvé que la méthode MARS est meilleur par rapport réseaux de neurones et MBR. Et nous avons étudié le modèles d'analyse de la variance avec la Méthode approche à mesures répétées pour traiter les données et nous aussi étudié l'analyse statistique des données avec plusieurs méthodes.
Dans notre étude on a montré que l'analyse de variance est peu sensible à la non-normalité des sujets et à l'inégalité des variances.
Et on a conclure que on peut transformée la variable de réponse au variable catégorique pour faciliter l'analyse.