POLYTECHNIQUE MONTRÉAL

MTH8302

Modèles de régression et d'analyse de la variance

Devoir 1

distribution : 14 mai 2018

remise : 27 mai 2018 à 23h59 (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.

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 

Transmettez votre rapport par courriel à bernard.clement@polymtl.ca

Nom suggéré pour le fichier à transmettre : aaaa_mmm_2018_MTH8302_devoirN.pdf

aaaa = nom de famille mmm = matricule N = numéro du devoir (1, 2, 3, 4)

TABLEAU CORRECTION

	valeur	obtenu
No 1-Anscombe	30	
No 2-Vaccins	30	
No 3-Croissance	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/

No 1 **Analyse diagnostique / graphique dans les modèles statistiques**

Données = Anscombe.sta

Réponse

1a)

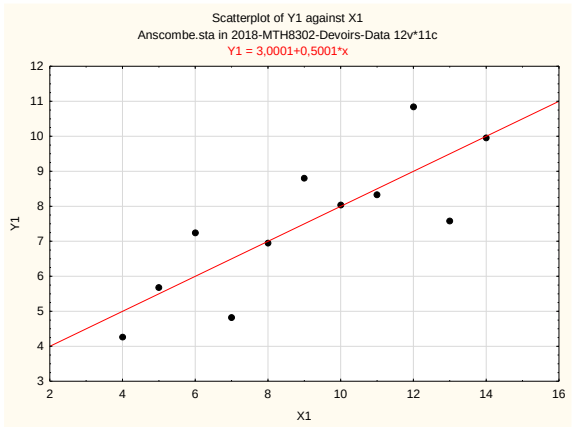
modèle couple		β0	β1	R^2	SSreg	SSresid	SStot
1	(X1, Y1)	3,0000909090	0,50009090909	0,66654246	27,5100009090909	13,76269	41,2726909090909
2	(X2, Y2)	3,0009090909	0,5	0,66624203	27,5	13,7762909090909	41,2762909090909
3	(X3, Y3)	3,0024545454	0,49972727272	0 ,66632404	27,4700081818182	13,7561918181818	41,2262
4	(X4, Y4)	3,0017272727	0,49990909090	0 ,66670726	27,4900009090909	13,74249	41,2324909090909

Ce tableau contient les coefficients (β0 , β1) des modèles de régression linéaire simple pour prédire Y en fonction de X pour chacun des 4 couples (X, Y) [Y= β0 + β1 X + ε]

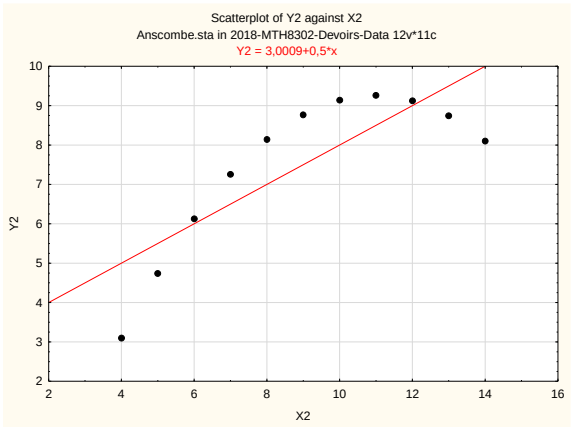
D'après le coefficient de détermination $R^2 > 0,5$ $(R > 0,7)$, on remarque que on a une bonne corrélation entre X et Y pour les 4 couples. Par contre on remarque que la somme de carrés résiduelle presque 32%, il y a grand partie de la variabilité de X qui n'est pas expliqué par notre modèle (pour les 4 couples). Ce que qui montre que nos modèles n'est pas bon.

On conclure que R^2 ne peut pas être suffisant pour juger la pertinence de la qualité de notre modèle

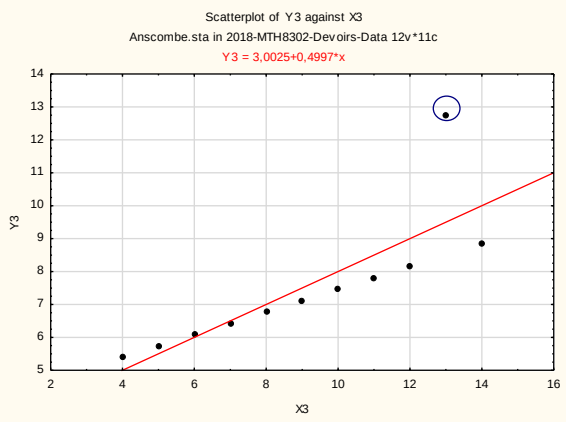
1b)



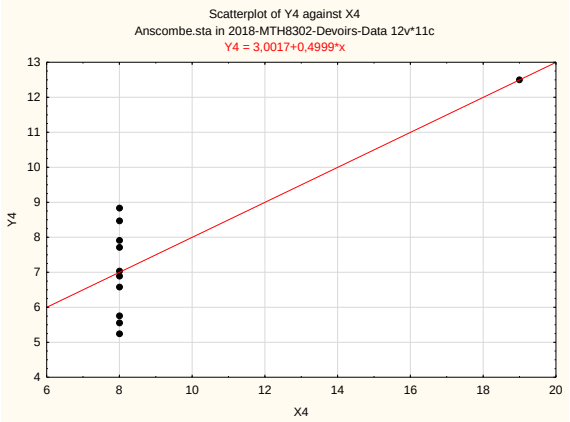
On remarque qu'il n'y a pas une corrélation linéaire, qui montre que notre modèle de prédiction n'est pas bon



On remarque qu'il n'y a pas une corrélation entre les 2 modèle réel et prédit, a cause de notre modèle réel qui n'est pas linéaire



On remarque qu'il y a un point aberrant qui empêche notre modèle prédit a suivi notre modèle réel

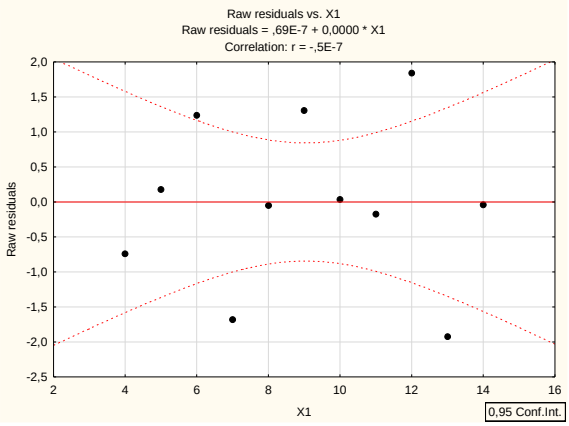


On remarque que X est constant et on a aucune corrélation entre le modèle réel et prédit

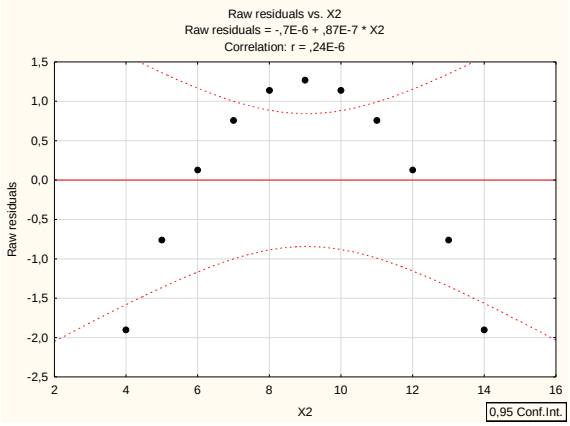
Nous concluons que les valeurs aberrantes ont eu une grande influence sur les modèles prédits

Et aussi d'après les remarques des graphes (Y en fonction de X), que nos modèles prédits ne sont pas bon, et on remarque que c'est le même résultat de la question 1a)

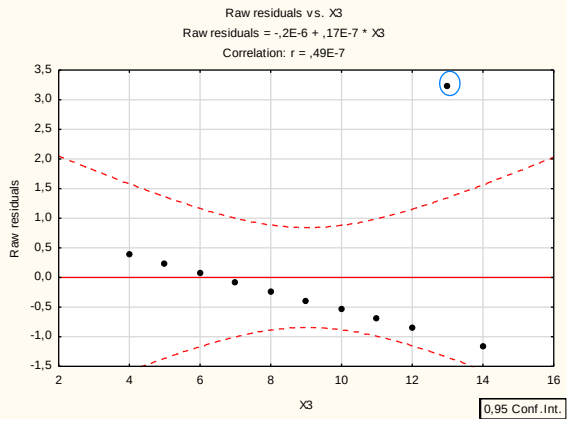
1c)



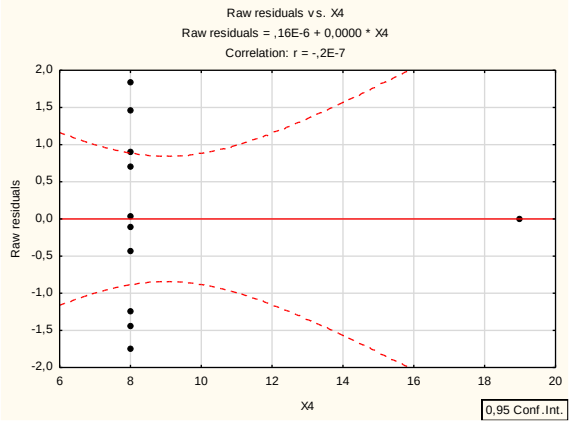
On remarque que la plupart des points sont loin de l'axe de 0, qui montre que notre modèle de prédiction n'est pas bon



On remarque que la majorité des points sont loin de l'axe de 0, et aussi qu'on peut pas linéariser notre modèle réel (modèle réel n'est pas linéaire)



On remarque que la plupart des points sont loin de l'axe de 0, à cause de le point aberrant

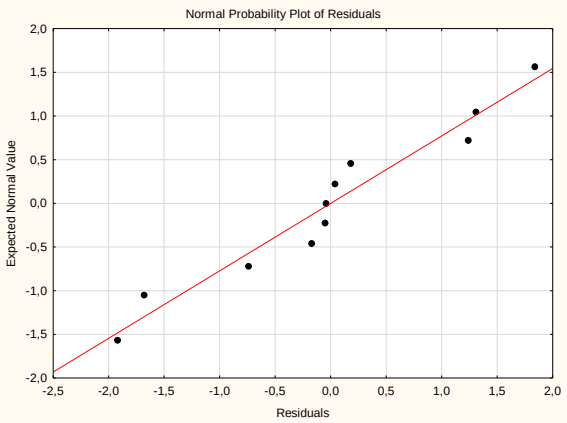


On remarque que X est constant et notre modèle n'est pas linéaire

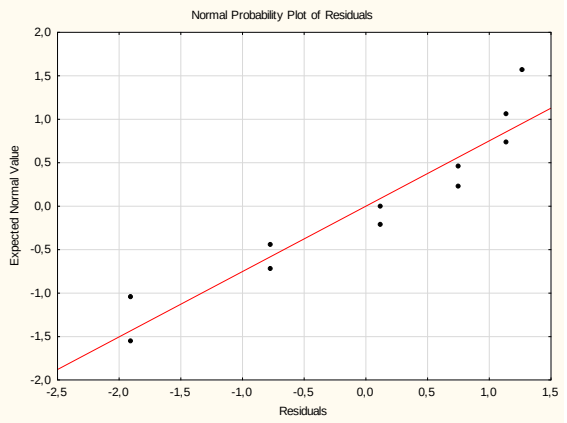
Nous concluons que les valeurs aberrantes ont eu une grande influence sur les modèles prédits

Et aussi d'après les remarques des graphes (résidus en fonction de la variable explicative X), que nos modèles prédits ne sont pas bon, et on remarque que c'est le même résultat de la question 1a)

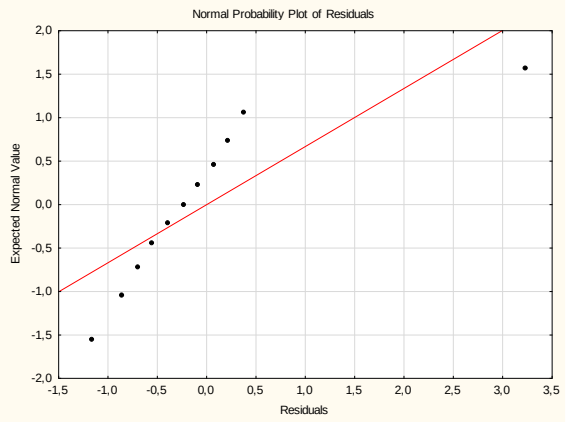
1d)



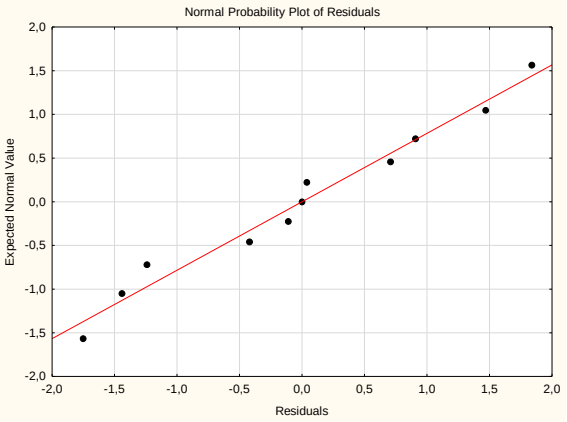
On remarque que les résidus respectent la loi de distribution normale



On remarque que les résidus respectent la loi de distribution normale

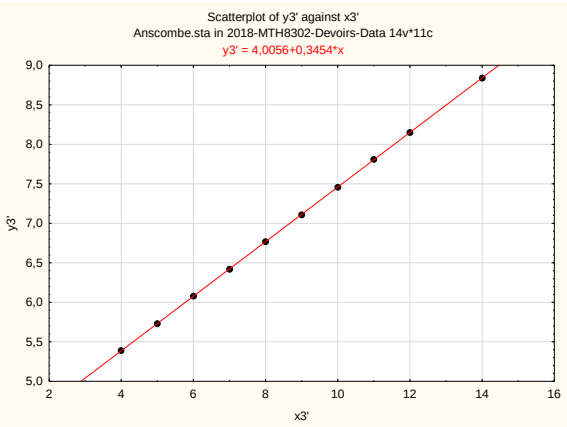


On remarque que les résidus ne respectent pas la loi de distribution normale

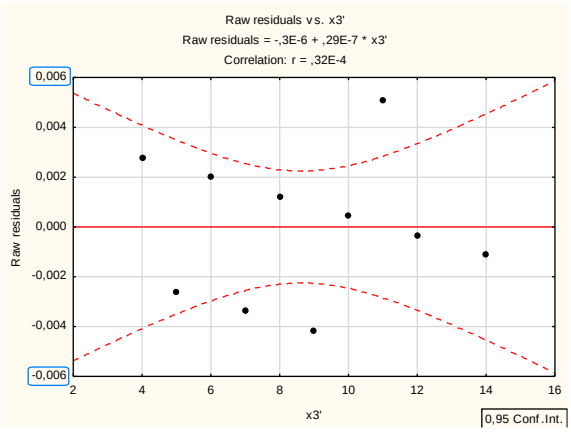


On remarque que les résidus respectent la loi de distribution normale

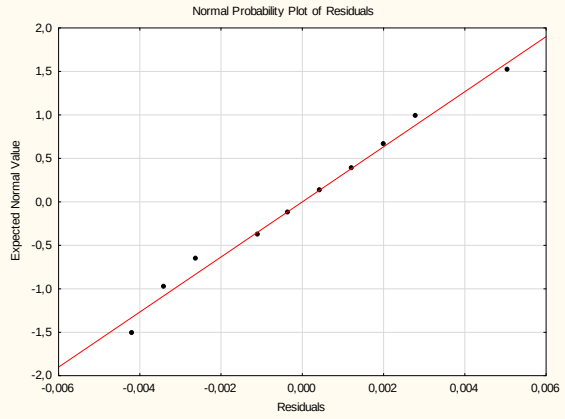
1e) Refaire les calculs sans l'observation (X3 =13 Y3 = 12,74).



Le modèle prédit a suivi notre modèle réel



On remarque que la plupart des points sont proche de l'axe de 0 qui prouve que le modèle prédit a suivi notre modèle réel



On remarque que les résidus respectent la loi de distribution normale

modèle couple	β0	β1	R^2	SSreg	SSresid	SStot
(X33, Y33)	4,0056493506	0,3453896103	0,99999311	11,022764025974	0,00007597402597 1940	11,02284

D'après coefficient de détermination $R^2 = 0,9999931 > 0,5$, $(R > 0,7)$, on remarque que on a une excellente corrélation entre X3' et Y3'

Et aussi on remarque que la somme de carrés résiduelle presque 0%. Ce que qui montre que nos modèles prédit sont bien suivis le modèle réel.

Cette observation est adéquate, avec cette observation nous avons obtenir un meilleur résultat.

Nous concluons que les points aberrants ont eu une grande influence sur les modèles prédits

1f) <u>Conclusion Générale</u>

les points aberrantes ont eu une grande influence sur les modèles prédits

On conclure que R^2 ne peut pas suffisant pour juger la pertinence de la qualité de notre modèle

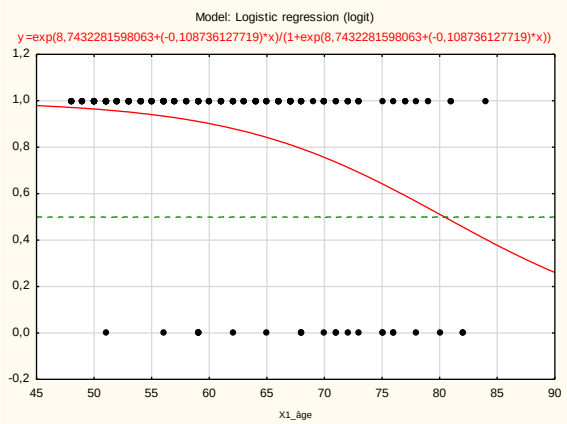
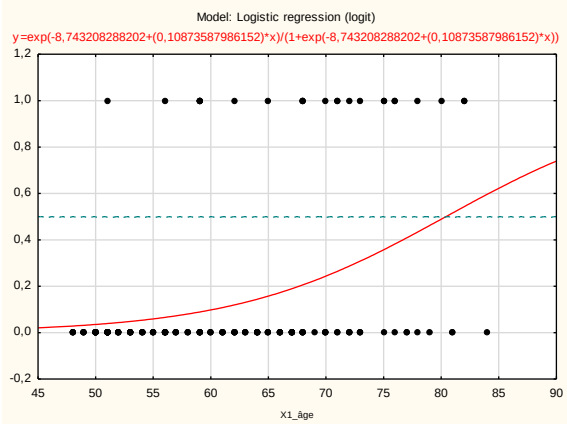
No 2 Régression logistique – programme de sensibilisation

Données = Vaccins.sta

Réponse

2a) modèle de régression logistique entre Y et X1_âge

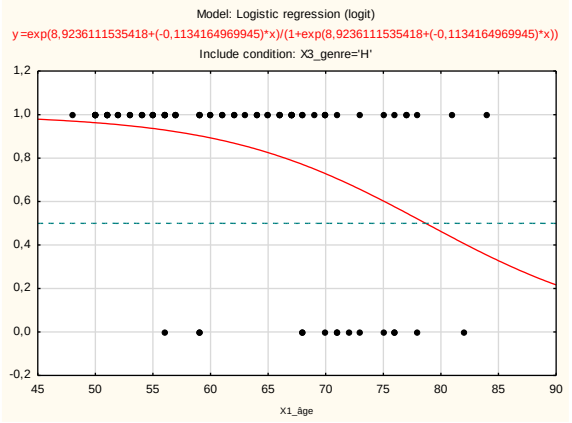
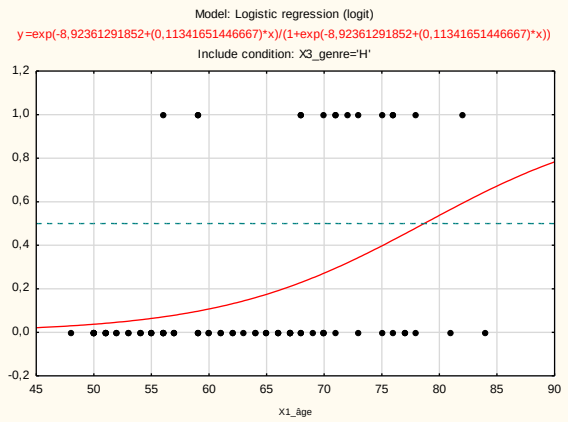
π(x) = exp(α+ βx)/(1 + exp(α+ βx))

Pour Y	Pour Yc=1-Y
	
α= 8,7432281598063	**α=-8,7432281598063**
β=-0,108736127719	**β=0,108736127719**

Ce graphique nous montre la probabilité prédite pour les personnes qui reçut le vaccin en fonction de l'Age. Par exemple la probabilité pour une personne de l'Age 67ans a reçu le vaccin est 80% (0,8), et d'autre de l'Age 84ans a reçu le vaccin est 40% (0,4).

Pour trouver la valeur prédite de Y on a choisi la probabilité 50% (0,5) comme un seuil de probabilité. Donc, il est possible de dire que toutes les personnes de 81 ans et moins sont vaccinées, par contre toutes les personnes de 81 ans et plus ne sont pas reçu le vaccin.

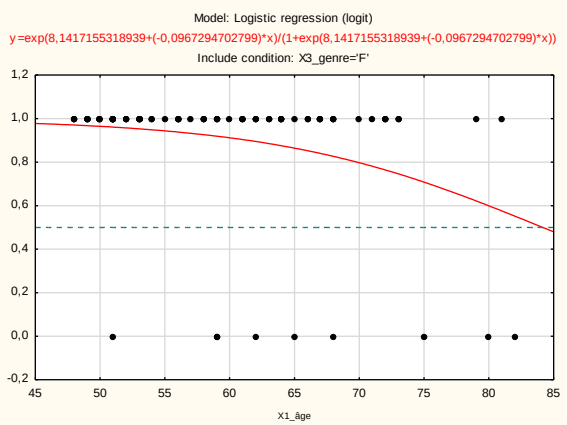
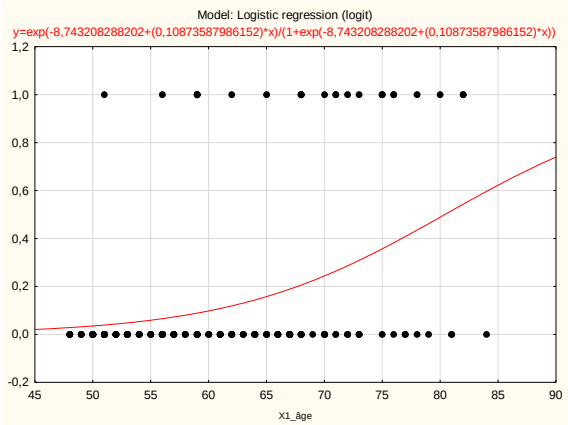
2b) pour les hommes seulement

Pour Y	Pour Yc=1-Y
	
α= 8,9236111535418 **β=-0,1134164969945**	**α= -8,9236111535418** **β=0,1134164969945**

Ce graphique nous montre la probabilité prédite pour les hommes qui reçut le vaccin en fonction de l'Age. Par exemple la probabilité pour un homme de l'Age 70ans a reçu le vaccin est 73% (0,73), et d'autre de l'Age 85ans a reçu le vaccin est 33% (0,33).

Pour trouver la valeur prédite de Y on a choisi la probabilité 50% (0,5) comme un seuil de probabilité. Donc, il est possible de dire que tous les hommes de 79 ans et moins sont vaccinées, par contre tous les hommes de 79 ans et plus ne sont pas reçu le vaccin

2c) pour les femmes seulement

Pour Y	Pour Yc=1-Y
	
α= 8,1417155318939 **β=-0,0967294702799**	**α=-8,1417155318939** **β=0,0967294702799**

Ce graphique nous montre la probabilité prédite pour les femmes qui reçut le vaccin en fonction de l'Age. Par exemple la probabilité pour une femme de l'Age 80ans a reçu le vaccin est 60% (0,60), et d'autre de l'Age 70ans a reçu le vaccin est 80% (0,80).

Pour trouver la valeur prédite de Y on a choisi la probabilité 50% (0,5) comme un seuil de probabilité. Donc, il est possible de dire que toutes les femmes de 84 ans et moins sont vaccinées, par contre toutes les femmes de 84 ans et plus ne sont pas reçu le vaccin.

2d) effet de l'âge et du sexe sur Y

Pour trouver la valeur prédite de Y on a choisi la probabilité 50% (0,5) comme un seuil de probabilité.

D'après les graphes on remarque que toutes les personnes (159 personnes) de 81 ans et moins sont vaccinées ($\widehat{Y} = 1\ ou\ \widehat{Y}_c = 0$).

Et par sexe, on a trouvé que les hommes (78 hommes) de 79 ans et moins sont vaccinées ($\widehat{Y} = 1\ ou\ \widehat{Y}_c = 0$), et 84 ans et moins sont vaccinées ($\widehat{Y} = 1\ ou\ \widehat{Y}_c = 0$) pour les femmes (81 femmes).

On remarque que le sexe joux un rôle sur la variation de Y ou on a trouvé que l'intervalle de l'Age des femmes qui sont vaccinées est supérieur par rapport l'intervalle de l'Age des hommes.

Et à partir de ces remarques et les remarques précédents, on peut déduire les valeurs prédites de Y.

2e)

Les équations utilisées pour trouver le tableau :
Valeurs observer :

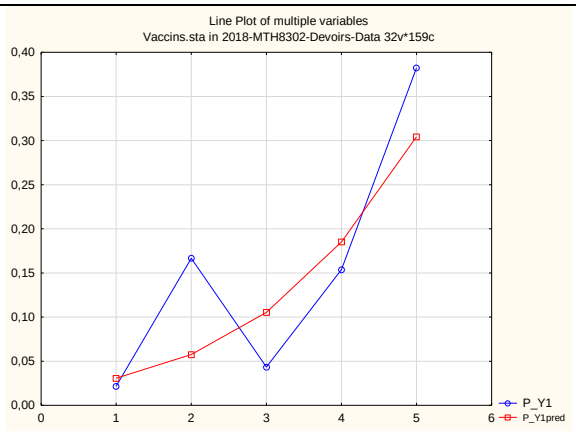
P_Y1= (Y=1) / n_total
P_Y0 = (Y=0) / n_total

logit(p) = ln[p/(1-p)]

Et pour trouver les valeurs de probabilité prédit on a fait la régression de logit(p)(obs) sur X1_catAge2 ,et à la fin on a utilisé l'équation p_pred = exp (logit(p)(pred)) / (1 + exp(logit(p)(pred))) pour trouver la probabilité prédit.

Pour Y=1

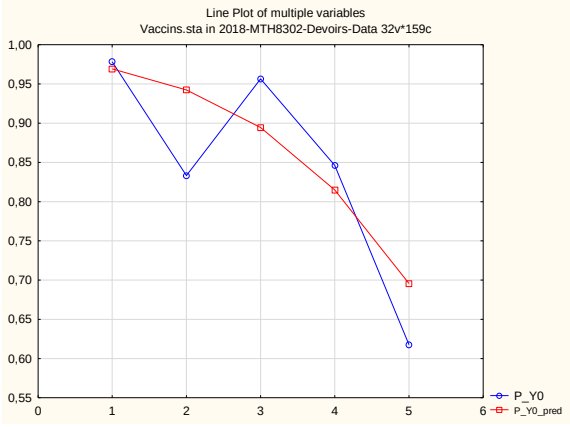
19 X1_catAge2.	20 Y=1.	21 P_Y1	22 logit(P_Y1) obs	23 logit(P_Y1) pred	24 P_Y1pred
52	1	0,021739	-3,80666	-3,450067	0,030767
57	5	0,166667	-1,60944	-2,794180	0,05764
62	1	0,043478	-3,09104	-2,138293	0,10543
67	4	0,153846	-1,70475	-1,482406	0,185064
72	13	0,382353	-0,47957	-0,826519	0,304382



Courbe qui montre la variabilité de probabilité observe et prédit pour Y=1

Pour Y=0

10 X1_catAge2	11 Y=0	12 Y=1	13 n_total	14 P_Y0	15 logit(P_Y0) obs	16 logit(P_Y0) pred	17 P_Y0_pred
52	45	1	46	0,978261	3,806662	3,450067	0,969233
57	25	5	30	0,833333	1,609438	2,794180	0,94236
62	22	1	23	0,956522	3,091042	2,138293	0,89457
67	22	4	26	0,846154	1,704748	1,482406	0,814936
72	21	13	34	0,617647	0,479573	0,826519	0,695618



Courbe qui montre la variabilité de probabilité observe et prédit pour Y=0

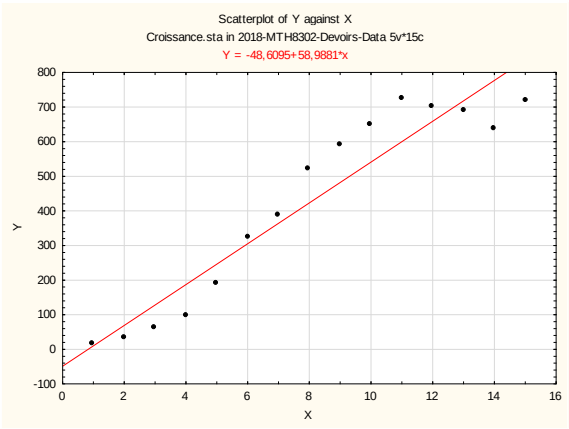
On remarque que le modèle 2a) plus précis par rapport le modèle 2e) puisque le modèle 2a) il nous donne toutes les probabilités pour chaque Age définie, par contre le 2eme modèle 2e) il nous donne la probabilité de chaque catégorie de l'Age mais ce modèle plus facile a étudié.

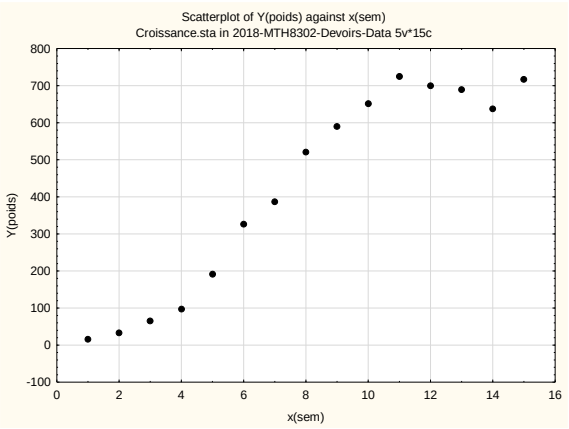
No 3 Régression non linéaire

Données = croissance.sta

Réponse

3a) le graphique des données.



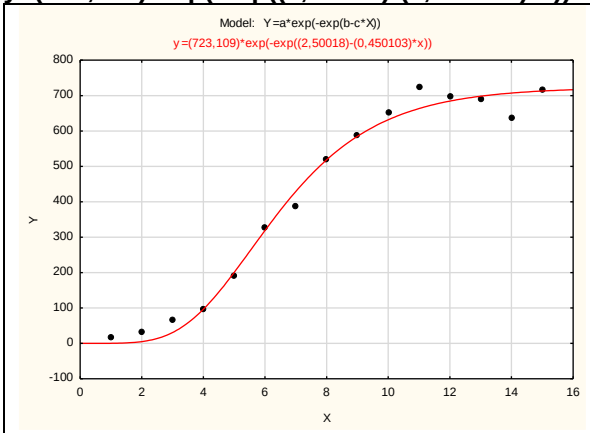


On remarque que la courbe est en forme de S (sigmoide)

3b) fonction de Gompertz

Y = a*exp [- exp (b – cx)]

y=(723,109)*exp(-exp((2,50018)-(0,450103)*x))



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

N=15 — Model: Y = a*exp(-exp(b-c*X)) (Croissance.sta in 2018-MTH8302-Devoirs-Data)
Dep. var: Y Loss: (OBS-PRED)**2
Final loss: 13606,142708 R= ,99366 Variance explained: 98,736%

	a	b	c
Estimate	723,1086	2,500185	0,450103
Std.Err.	22,6246	0,325911	0,057625
t(12)	31,9612	7,671375	7,810844
-95%CL	673,8139	1,790086	0,324548
+95%CL	772,4033	3,210283	0,575658
p-value	0,0000	0,000006	0,000005

D'après la Variance explained (98,736%) on conclure que on a une bonne corrélation entre le modèle prédit et réel

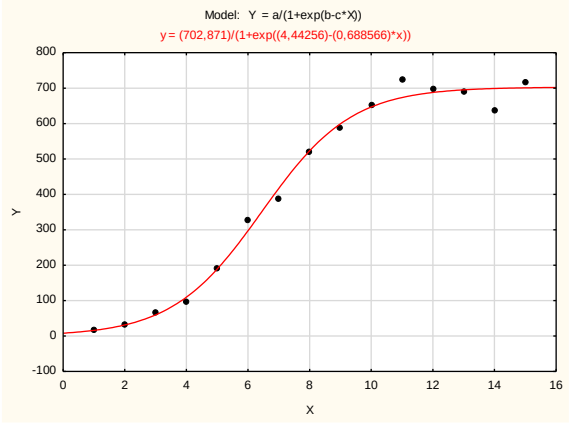
Model is: Y = a*exp(-exp(b-c*X)) (Croissance.sta in 2018-MTH8302-Devoirs-Data)
Dep. Var. : Y

	Observed	Predicted	Residuals
1	16,0800	0,3058	15,7742
2	33,8300	5,1071	28,7229
3	65,8000	30,7461	35,0539
4	97,2000	96,5698	0,6302
5	191,5500	200,3287	-8,7787
6	326,2000	318,9979	7,2021
7	386,8700	429,1448	-42,2748
8	520,5300	518,4801	2,0499
9	590,0300	584,9167	5,1133
10	651,9200	631,6521	20,2679
11	724,9300	663,3799	61,5501
12	699,5600	684,4353	15,1247
13	689,9600	698,2070	-8,2470
14	637,5600	707,1316	-69,5716
15	717,4100	712,8811	4,5289

3c) Fonction Logistique 3P

$Y = a / [1 + \exp (b – cx)]$

y= (702,871)/(1+exp((4,44256)-(0,688566)*x))



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

Model: Y = a/(1+exp(b-c*X)) (Croissance.sta in 2018-MTH8302-Devoirs-Data)
Dep. var: Y Loss: (OBS-PRED)**2
Final loss: 8929,8829725 R= ,99584 Variance explained: 99,170%

N=15	a	b	c
Estimate	702,8714	4,442564	0,688566
Std.Err.	15,5738	0,455089	0,075819
t(12)	45,1317	9,761972	9,081712
-95%CL	668,9391	3,451011	0,523371
+95%CL	736,8038	5,434117	0,853761
p-value	0,0000	0,000000	0,000001

D'après la Variance explained (99,170%) on conclure que on a une bonne corrélation entre le modèle prédit et réel

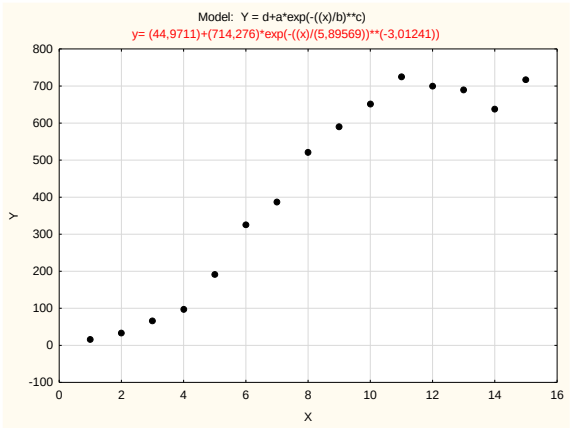
Model is: Y = a/(1+exp(b-c*X)) (Croissance.sta in 2018-MTH8302-Devoirs-Data)
Dep. Var. : Y

	Observed	Predicted	Residuals
1	16,0800	16,0872	-0,0072
2	33,8300	31,3171	2,5129
3	65,8000	59,7116	6,0884
4	97,2000	109,6476	-12,4476
5	191,5500	189,0680	2,4820
6	326,2000	297,1948	29,0052
7	386,8700	416,9752	-30,1052
8	520,5300	522,8158	-2,2858
9	590,0300	599,2142	-9,1842
10	651,9200	646,6804	5,2396
11	724,9300	673,4774	51,4526
12	699,5600	687,7931	11,7669
13	689,9600	695,2159	-5,2559
14	637,5600	699,0051	-61,4451
15	717,4100	700,9241	16,4859

3d) Fonction Weibull.

Y= d + a*e{ - [(x) / b]c }**

y= (44,9711)+(714,276)*exp(-((x)/(5,89569))(-3,01241))**



	Model: Y = d+a*exp(-((x)/b)**c) (Croissance.sta in 2018-MTH8302-Devoirs-Data) Dep. var: Y Loss: (OBS-PRED)**2 Final loss: 15227,380167 R= ,99290 Variance explained: 98,585%			
N=15	d	a	b	c
Estimate	44,97109	714,2764	5,895690	-3,01241

	Model is: Y = d+a*exp(-((x)/b)**c) (Croissance.sta in 2018-MTH8302-Devoirs-Data) Dep. Var. : Y		
	Observed	Predicted	Residuals
1	16,0800	44,9711	-28,8911
2	33,8300	44,9711	-11,1411
3	65,8000	45,3098	20,4902
4	97,2000	73,5821	23,6179
5	191,5500	183,1403	8,4097
6	326,2000	321,6147	4,5853
7	386,8700	438,4709	-51,6009
8	520,5300	524,3691	-3,8391
9	590,0300	585,0051	5,0249
10	651,9200	627,6757	24,2443
11	724,9300	658,0484	66,8816
12	699,5600	680,0304	19,5296
13	689,9600	696,2280	-6,2680
14	637,5600	708,3755	-70,8155
15	717,4100	717,6378	-0,2278

3e)

Pour trouver le meilleur choix de fonction pour modéliser les données on a utilisé le R-Square : proportion de la variation de la variable de réponse expliquée par le modèle.

	Gompertz	Logistique 3P	Weibull
Variance explained	**98,736%**	**99,170%**	**98,585%**
R-Square	**0,993659**	**0,995841**	**0,992899**

Nous avons trouvez que le meilleur choix de fonction pour modéliser les données est la fonction Logistique

Remarque : dans ce cas on peut aussi utilisé le critère SE : Sum of Square of Errors =différences entre les valeurs observées et les valeurs prédites.