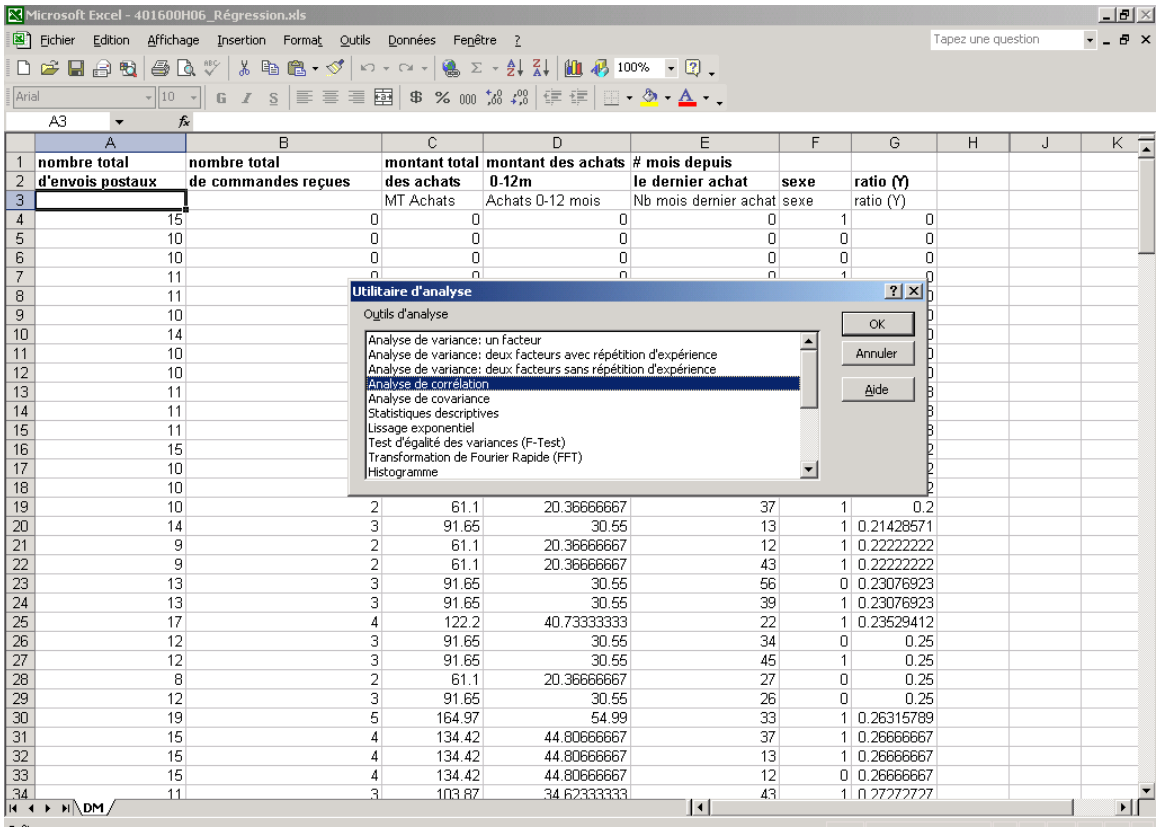
Comment effectuer une régression linéaire avec Excel

L'analyse de régression linéaire utilise la méthode des « moindres carrés » pour tracer une droite sur l'ensemble des observations, et analyse l'incidence des variables indépendantes sur la variable dépendante unique. (Par exemple, vous voulez savoir si la probabilité d'achat du catalogue par le client varie en fonction du montant total des achats, du sexe, etc…). Excel peut prendre en compte jusqu'à 16 variables indépendantes.

Prenons l'exemple du fichier Régression.xls et effectuons une analyse de régression linéaire afin d'expliquer le ratio Y (probabilité d'achat).

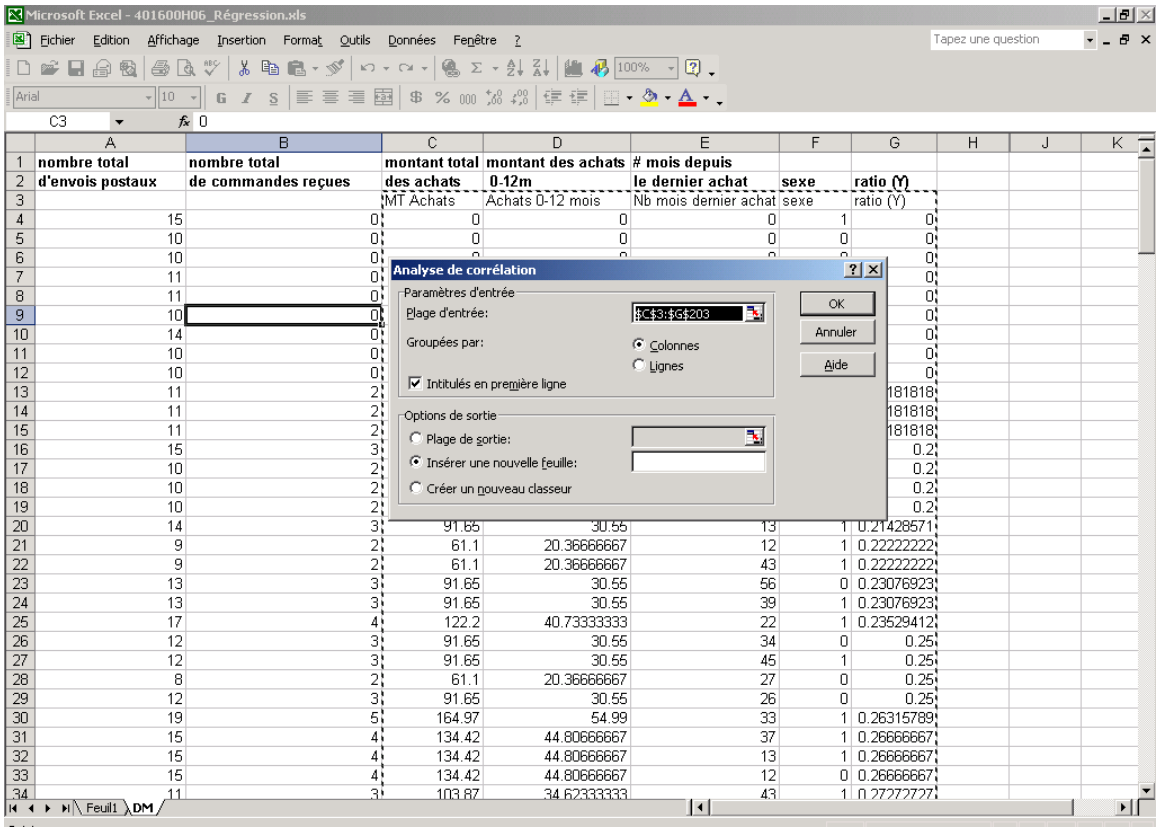
Effectuons d'abord une analyse de corrélation entre les variables.

Procédure : Outils → Utilitaire d'analyse → Analyse de corrélation



	A	B	C	D	E	F	G
1	**nombre total**	**nombre total**	**montant total**	**montant des achats**	**# mois depuis**		
2	**d'envois postaux**	**de commandes reçues**	**des achats**	**0-12m**	**le dernier achat**	**sexe**	**ratio (Y)**
3			MT Achats	Achats 0-12 mois	Nb mois dernier achat	sexe	ratio (Y)
4	15	0	0	0	0	1	0
5	10	0	0	0	0	0	0
6	10	0	0	0	0	0	0
19	10	2	61.1	20.36666667	37	1	0.2
20	14	3	91.65	30.55	13	1	0.21428571
21	9	2	61.1	20.36666667	12	1	0.22222222
22	9	2	61.1	20.36666667	43	1	0.22222222
23	13	3	91.65	30.55	56	0	0.23076923
24	13	3	91.65	30.55	39	1	0.23076923
25	17	4	122.2	40.73333333	22	1	0.23529412
26	12	3	91.65	30.55	34	0	0.25
27	12	3	91.65	30.55	45	1	0.25
28	8	2	61.1	20.36666667	27	0	0.25
29	12	3	91.65	30.55	26	0	0.25
30	19	5	164.97	54.99	33	1	0.26315789
31	15	4	134.42	44.80666667	37	1	0.26666667
32	15	4	134.42	44.80666667	13	1	0.26666667
33	15	4	134.42	44.80666667	12	0	0.26666667
34	11	3	103.87	34.62333333	43	1	0.27272727

Ensuite sélectionner les 4 dernières colonnes de la feuille de calcul dans la plage d'entrée (avec les intitulés pour pouvoir repérer les noms des variables après). Cocher la case : intitulés en première ligne. Cliquer OK.



	A	B	C	D	E	F	G
1	nombre total	nombre total	montant total	montant des achats	# mois depuis		
2	d'envois postaux	de commandes reçues	des achats	0-12m	le dernier achat	sexe	ratio (Y)
3			MT Achats	Achats 0-12 mois	Nb mois dernier achat	sexe	ratio (Y)
4	15	0	0	0	0	1	0
5	10	0	0	0	0	0	0
6	10	0					0
7	11	0					0
8	11	0					0
9	10	0					0
10	14	0					0
11	10	0					0
12	10	0					0
13	11	2					181818
14	11	2					181818
15	11	2					181818
16	15	3					0.2
17	10	2					0.2
18	10	2					0.2
19	10	2					0.2
20	14	3	91.65	30.55	13	1	0.21428571
21	9	2	61.1	20.36666667	12	1	0.22222222
22	9	2	61.1	20.36666667	43	1	0.22222222
23	13	3	91.65	30.55	56	0	0.23076923
24	13	3	91.65	30.55	39	1	0.23076923
25	17	4	122.2	40.73333333	22	1	0.23529412
26	12	3	91.65	30.55	34	0	0.25
27	12	3	91.65	30.55	45	1	0.25
28	8	2	61.1	20.36666667	27	0	0.25
29	12	3	91.65	30.55	26	0	0.25
30	19	5	164.97	54.99	33	1	0.26315789
31	15	4	134.42	44.80666667	37	1	0.26666667
32	15	4	134.42	44.80666667	13	1	0.26666667
33	15	4	134.42	44.80666667	12	0	0.26666667
34	11	3	103.87	34.62333333	43	1	0.27272727

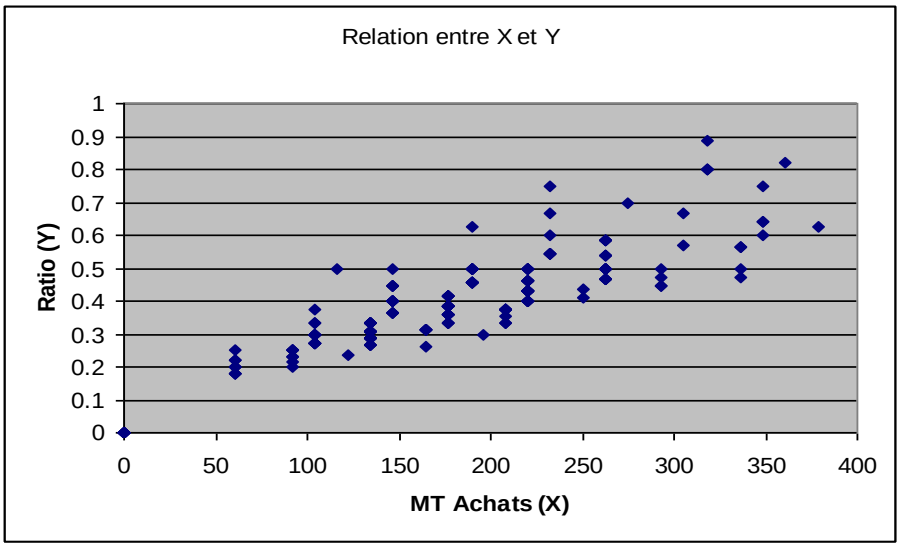
Une nouvelle feuille est créée (Feuil1). C'est la matrice de corrélation des variables.

	MT Achats	*Achats 0-12 mois*	*Nb mois dernier achat*	*sexe*	*ratio (Y)*
MT Achats	1				
Achats 0-12 mois	1	1			
Nb mois dernier achat	-0.339315952	-0.339315952	1		
sexe	-0.329743615	-0.329743615	0.224882033	1	
ratio (Y)	0.882789174	0.882789174	-0.31314819	-0.34869836	1

D'après le tableau, la variable la plus corrélée avec la variable dépendante Y est le montant total des achats (MT Achats). La variable Achats 0-12 mois est aussi corrélée avec Y au même degré que MT Achats mais cela provient d'un problème de ***multicollinéarité*** entre MT Achats et Achats 0-12 mois qu'on va essayer de résoudre par la suite. (Remarquer la corrélation entre MT Achats et Achats 0-12 mois qui est de 1.

Régression linéaire simple : (une seule variable explicative)

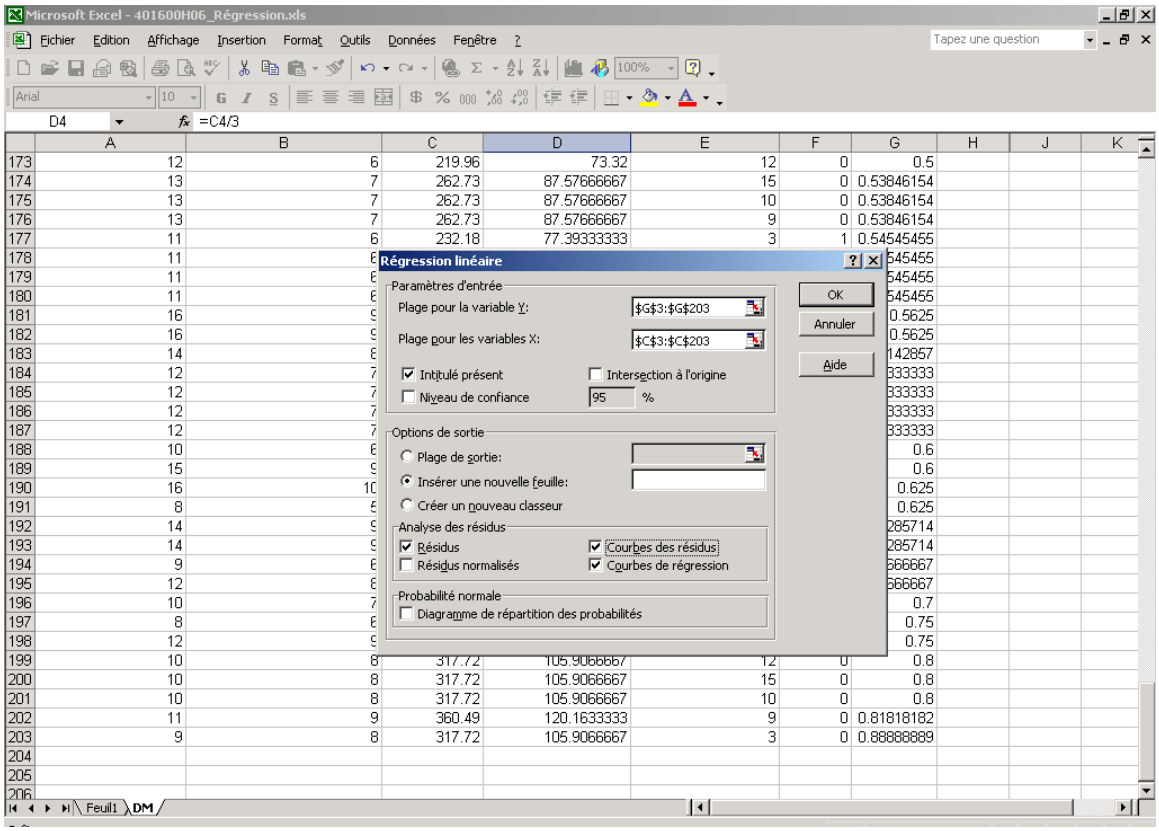
La prochaine étape avant de faire la régression est de tracer un graphique de la relation entre X et Y. (Je vous laisse le soin de le faire sur Excel). On voit que la relation entre X et Y est bien de forme linéaire (d'après le graphique ci-dessous).



Effectuons maintenant la régression entre MT Achats et Y.

Procédure : Outils → Utilitaire d'analyse → Régression linéaire

Indiquez les données pour la variable Y, et pour la (ou les) variable(s) X. Cochez les cases : Intitulé présent, Résidus, Courbes des résidus et Courbes de régression et faites OK.



Une portion de la feuille des résultats (Feuil2) est affichée ci-dessous :

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	**0.88278917**
Coefficient de détermination R^2	**0.77931673**
Coefficient de détermination R^2	0.77820216
Erreur-type	0.0732685
Observations	200

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*
Régression	1	3.75356875	3.75356875	699.213441
Résidus	198	1.06291808	0.00536827	
Total	199	4.81648683		

	Coefficients	*Erreur-type*	*Statistique t*	***Probabilité***
Constante	**0.08329775**	0.01287962	6.46740804	**0.00**
MT Achats	**0.00172084**	6.5078E-05	26.4426444	**0.00**

Les résultats affichés sont :
- le ***coefficient de détermination multiple*** (dans le cas à deux variables, cela correspond simplement au coefficient r de corrélation de Pearson)
- le ***coefficient de détermination*** $\mathbf{R^2}$: il donne une idée du % de variabilité de la variable à modéliser, et plus le coefficient R^2 est proche de 1, plus il y a une corrélation et meilleur est le modèle. Dans notre exemple, 77.9 % de la variabilité de Y est expliquée par MT Achats.
- les ***coefficients*** : sont les coefficients de la droite de régression

$$\hat{Y} = b_0 + b_1 * X$$

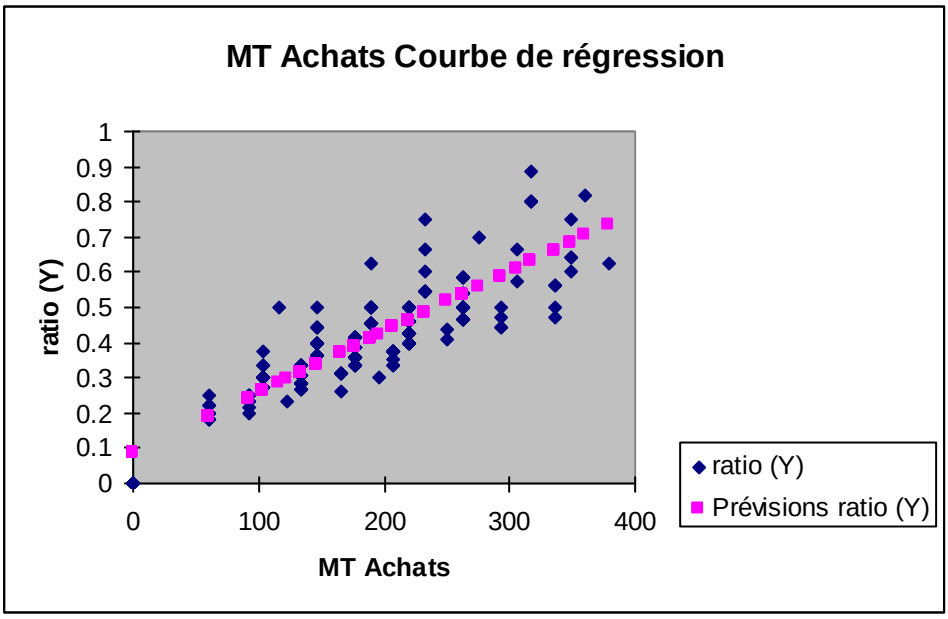
$b_0 = 0.08$ (ordonnée à l'origine)
$b_1 = 0.001$ (pente de la droite)

$$\text{Probabilité d'achat} = 0.08 + 0.001 * \text{MT Achats}$$

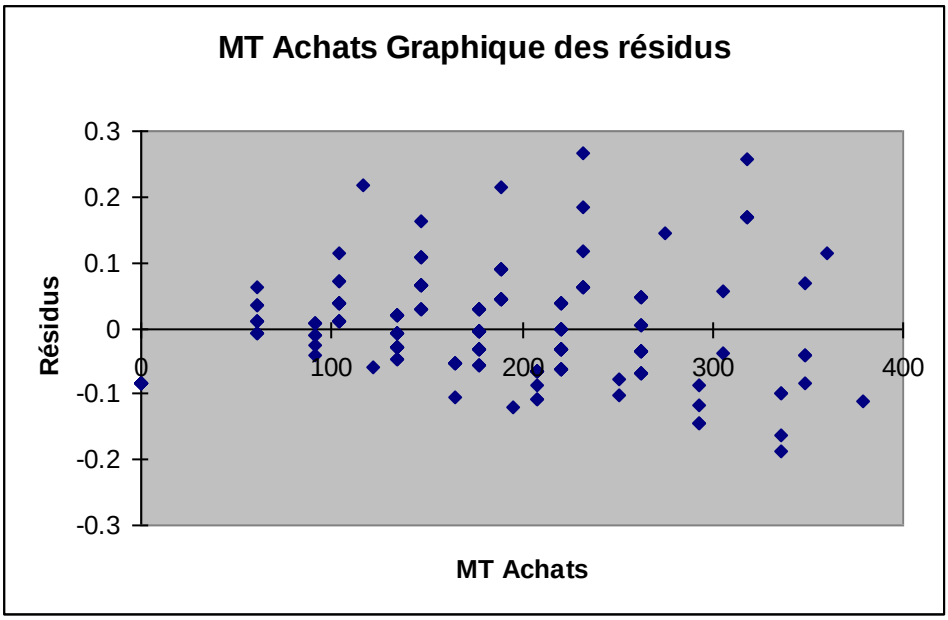
En d'autres termes, si MT Achats est nul, alors la probabilité Y sera de 0.08 et si MT Achats augmente d'une unité (1 dollar), Y augmente de 0.001.

Il ne faut pas oublier de prêter attention à la colonne des ***probabilités***. Surtout celle de la variable X. Si cette dernière est < 0.05, alors la variable X est significative. Dans notre cas, P=0.00 donc MT Achats est significative dans le modèle de régression. Cela veut dire aussi que la pente de la droite de régression diffère de 0, et donc nous admettons qu'il existe une relation linéaire significative entre la probabilité d'achat Y et le montant total des achats X.

En effet, d'après le graphique de la courbe de régression, il est clair qu'il y a une relation significative entre MT Achats et ratio (Y).



On peut aussi remarquer à l'aide du graphique des résidus que l'hypothèse de constance de la variance n'est pas respectée (résidus sous dorme de cône). Cependant, celle de l'indépendance l'est puisque la majorité des résidus se répartissent de façon aléatoire et il y a autant de positifs que de négatifs.



Régression linéaire simple : (plusieurs variables explicatives)

Maintenant, essayons de construire un modèle de régression avec plusieurs variables explicatives et ce, afin d'expliquer encore plus la variable Y et réduire la variance résiduelle. On a 4 variables indépendantes au total mais rappelons que 2 variables parmi les 4 sont fortement corrélées (MT Achats et Achats 0-12 mois). On ne va donc introduire qu'une des 2 afin de ne pas avoir des problèmes de stabilité du modèle.

Revenons donc à l'utilitaire d'analyse → Régression linéaire, sélectionner ratio (Y) dans la plage des Y et MT Achats, Nb mois dernier achat et sexe dans la plage des X.

Les résultats sont affichés dans Feuil3 :

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0.88492112
Coefficient de détermination R^2	**0.78308539**
Coefficient de détermination R^2	0.77976527
Erreur-type	0.07300987
Observations	200

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*
Régression	3	3.77172047	1.25724016	235.860457
Résidus	196	1.04476636	0.00533044	
Total	199	4.81648683		

	Coefficients	*Erreur-type*	*Statistique t*	***Probabilité***
Constante	**0.10254739**	0.01989798	5.15365896	**0.00**
MT Achats	**0.0016751**	7.1735E-05	23.3511432	**0.00**
Nb mois dernier achat	**-9.7579E-05**	0.00048071	-0.20299152	**0.84**
sexe	**-0.01983856**	0.01106192	-1.79341005	**0.07**

À première vue, le modèle n'est pas aussi meilleur que l'on a prévu puisque le R^2 n'a augmenté que de 0.04 % ce qui est peu. Les 2 variables supplémentaires Nb mois dernier achat et sexe n'apportent donc que 0.04 % d'explication de Y. Les probabilités de ces 2 variables sont aussi > 0.05 donc elles ne sont pas significatives dans le modèle. Une solution est de les enlever et donc on revient à notre modèle de départ qui reste le meilleur modèle trouvé.