TP de MOTUS – Génération

Nicolas Pelé
Maxim Peveri

Détermination de la loi de génération sur les données de 1995

Le motif sélectionné est le motif « enseignement université », en attraction et en émission.
Nous allons déterminer la meilleure loi de génération, en testant différent modèles de manière progressive.

Tout d'abord, le problème d'hétéroscédasticité est présent. Il peut être traité sur les variables de population (population totale, nombre d'étudiants par exemple) mais il ne peut être traité sur des variables d'activité (comme le nombre de place étudiants).

Afin de tester la qualité de chaque régression, nous avons utilisé différents critères :

- Le R^2 correspond au pourcentage de la variance initiale du nuage de points expliquée par la droite de régression. Plus le R^2 est important, meilleur est la régression car plus la régression explique un pourcentage élevé de la variance. Cependant, ce R^2 augmente automatiquement plus il y a de variables explicatives.
- Le R^2 corrigé permet de comparer des régressions qui ne possèdent pas le même nombre de variables explicatives.

- Le F de Fisher correspond au rapport entre les carrés expliqués et les carrés non expliqués. Si la valeur critique du F de Fisher est inférieure à la valeur critique de 5%, alors on considère que la régression est significative à 95%, c'est-à-dire qu'au moins une des variables possède un coefficient non nul.

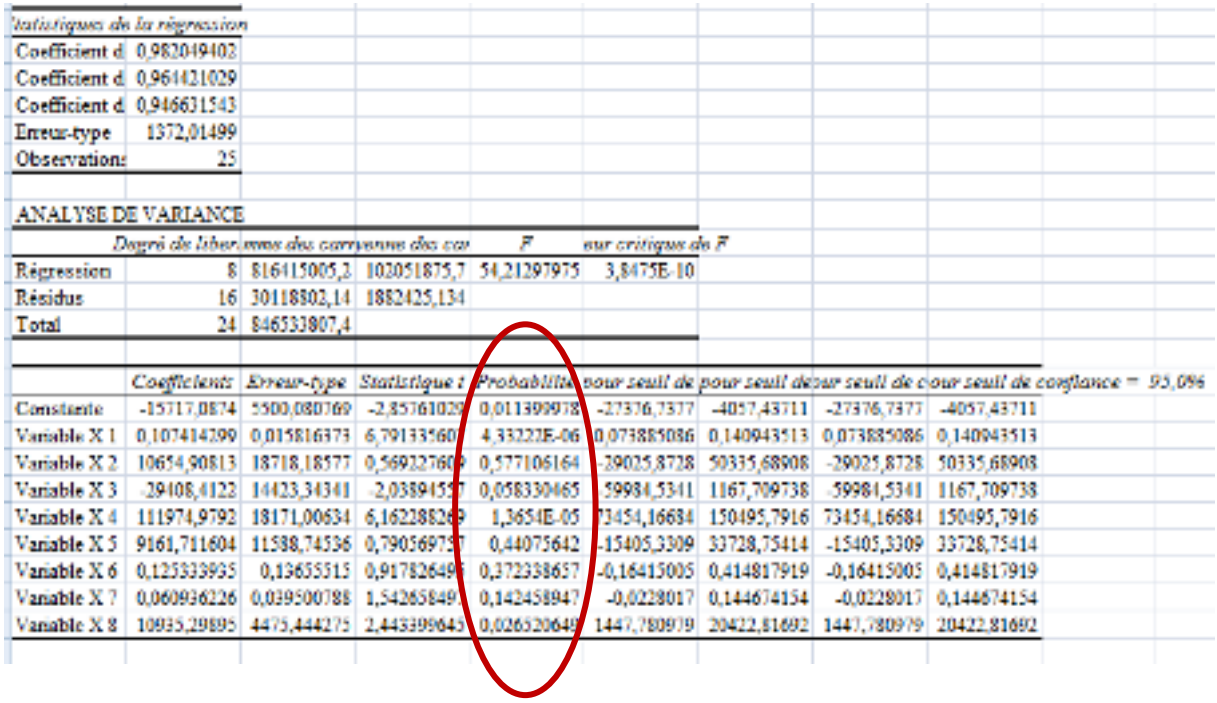
- Si le t de Student observé est supérieur au t théorique de 5% pris ici, on ne peut garder la variable concerné.

- Le signe des coefficients est important et nécessite réflexion : si le coefficient est positif, cela signifie que la variable favorise le motif. On peut alors rejeter une variable si son coefficient n'est pas logique.

Emission :

1ère régression

Le premier modèle choisi consiste à prendre comme variables explicatives un maximum des variables. Nous avons donc commencé en prenant 8 variables : la "population totale", le "taux de scolaire dans le primaire", le "taux de scolaire dans le collège-lycée", le "taux d'étudiants", le "taux d'actifs ayant un emploi", le "nombre de place dans l'enseignement secondaire", le "nombre de places étudiantes" et la "motorisation".
Il faut commencer par ramener les variables de nombre de scolaire dans le primaire et dans les collèges-lycées, et d'étudiant en taux, en divisant ces chiffres par la population de la zone : cela permet de résoudre le problème l'effet de masse.



Statistiques de la régression	
Coefficient d	0,982049402
Coefficient d	0,964421029
Coefficient d	0,946631543
Erreur-type	1372,01499
Observations	25

ANALYSE DE VARIANCE

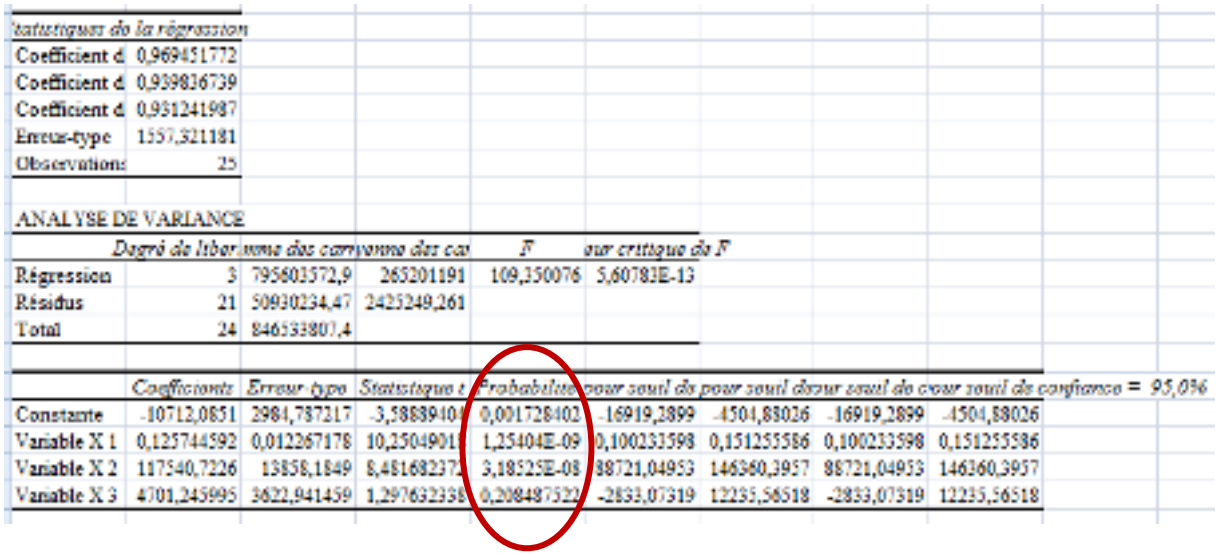
	Degré de liber	*mme des carrés*	*mme des car*	*F*	*aur critique de F*
Régression	8	816415005,2	102051875,7	54,21297975	3,8475E-10
Résidus	16	30118802,14	1882425,134		
Total	24	846533807,4			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*	*pour seuil de*	*pour seuil de*	*our seuil de*	*our seuil de confiance = 95,0%*
Constante	-15717,0874	5500,080769	-2,85761029	0,011399978	-27376,7377	-4057,43711	-27376,7377	-4057,43711
Variable X 1	0,107414299	0,015816373	6,791335601	4,33223E-06	0,073885086	0,140943513	0,073885086	0,140943513
Variable X 2	10654,90813	18718,18577	0,569227609	0,577106164	-29025,8728	50335,68908	-29025,8728	50335,68908
Variable X 3	-29408,4122	14423,34341	-2,038945127	0,058330465	-59984,5341	1167,709738	-59984,5341	1167,709738
Variable X 4	111974,9792	18171,00634	6,162288269	1,3654E-05	73454,16684	150495,7916	73454,16684	150495,7916
Variable X 5	9161,711604	11588,74536	0,790569737	0,44075642	-15405,3309	33728,75414	-15405,3309	33728,75414
Variable X 6	0,125333935	0,13655515	0,917826496	0,372338657	-0,16415005	0,414817919	-0,16415005	0,414817919
Variable X 7	0,060936226	0,039500788	1,542658497	0,142458947	-0,0228017	0,144674154	-0,0228017	0,144674154
Variable X 8	10935,29895	4475,444275	2,443399643	0,026520649	1447,780979	20422,81692	1447,780979	20422,81692

Le R^2 de la régression est bon (0.9644), ce qui s'explique par le grand nombre de variables.
La valeur critique de F est excellente (3.8475E-10), ce qui signifie qu'au moins une variable possède un coefficient non nul.
Cependant les probabilités t des variables "taux de scolaire dans le primaire", "taux de scolaire dans le collège-lycée", "taux d'actifs ayant un emploi", "nombre de place dans l'enseignement secondaire" et "nombre de places étudiantes" sont supérieurs à 5%, ce qui signifie que ce ne sont pas des variables explicatives.

Par conséquent nous gardons les variables "population totale", "taux d'étudiants" et "motorisation" pour effectuer la deuxième régression.

2ème régression



Statistiques de la régression	
Coefficient d	0,969451772
Coefficient d	0,939836739
Coefficient d	0,931241987
Erreur-type	1557,321181
Observations	25

ANALYSE DE VARIANCE

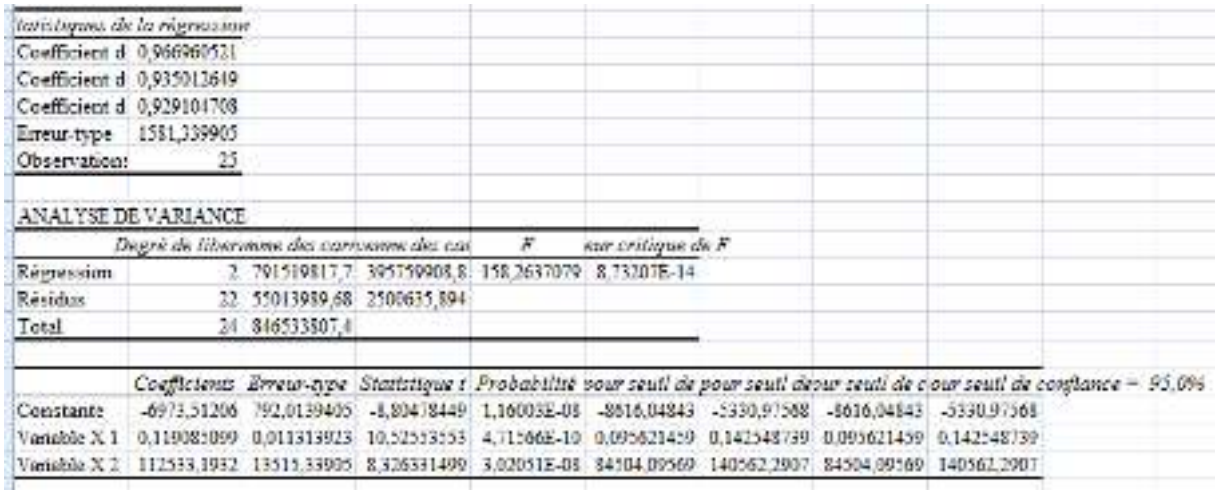
	Degré de liber	somme des carr	somme des car	F	eur critique de F
Régression	3	795603572,9	265201191	109,350076	5,60783E-13
Résidus	21	50930234,47	2425249,261		
Total	24	846533807,4			

	Coefficients	Erreur-type	Statistique t	Probabilité	pour seuil de	pour seuil de	our seuil de c	our seuil de confiance = 95,0%
Constante	-10712,0851	2984,787217	-3,5888940	0,001728402	-16919,2899	-4504,88026	-16919,2899	-4504,88026
Variable X 1	0,125744592	0,012267178	10,25049019	1,25404E-09	0,100233598	0,151255586	0,100233598	0,151255586
Variable X 2	117540,7226	13858,1849	8,481682373	3,18525E-08	88721,04953	146360,3957	88721,04953	146360,3957
Variable X 3	4701,245995	3622,941459	1,297632338	0,208487522	-2833,07319	12235,56518	-2833,07319	12235,56518

Le R^2 est toujours satisfaisant (0.9398) ainsi que la valeur critique de F (5.6078E-13).
Cependant la probabilité t de la variable "motorisation" est supérieure à 5% : cette variable n'est pas significative.

Nous gardons donc les variables "population totale" et "taux d'étudiant" pour la régression suivante.

3ème régression



tatistiques de la régression	
Coefficient d	0,966960521
Coefficient d	0,935013649
Coefficient d	0,929104708
Erreur-type	1581,339905
Observations	25

ANALYSE DE VARIANCE

	Degré de libe	somme des carr	somme des car	F	eur critique de F
Régression	2	791519817,7	395759908,8	158,3637079	8,73207E-14
Résidus	22	55013989,68	2500635,894		
Total	24	846533807,4			

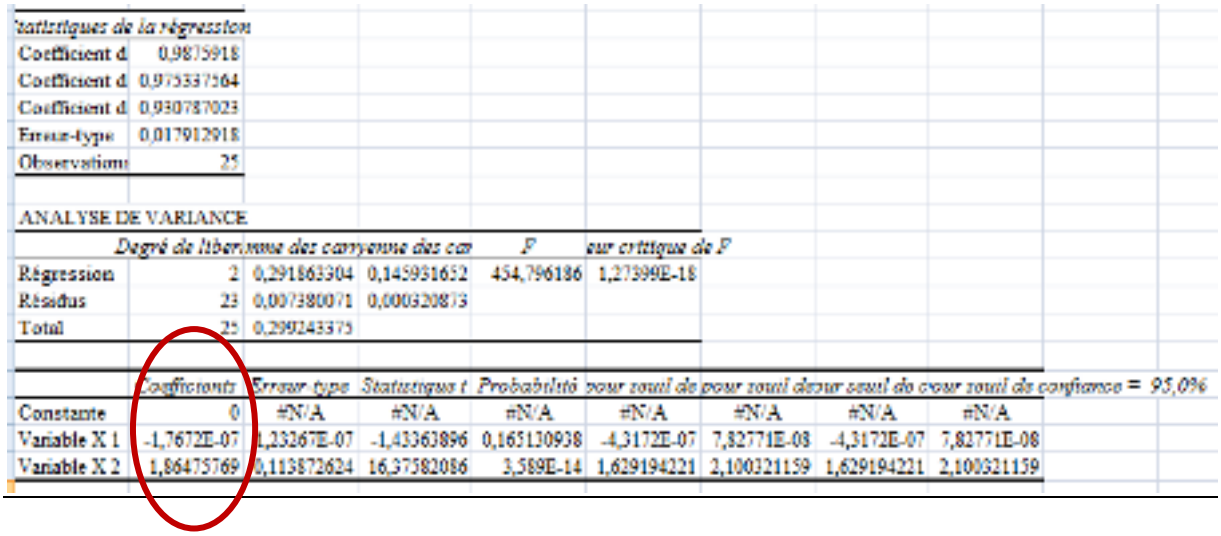
	Coefficients	Erreur-type	Statistique t	Probabilité	our seuil de	pour seuil de	our seuil de c	our seuil de confiance = 95,0%
Constante	-6973,51206	792,0139405	-8,804178449	1,16003E-08	-8616,04843	-5330,97568	-8616,04843	-5330,97568
Variable X 1	0,119083099	0,011313923	10,52553553	4,71566E-10	0,095621459	0,142544739	0,095621459	0,142544739
Variable X 2	112533,3932	13515,33905	8,326331499	3,02051E-08	84504,09569	140562,2907	84504,09569	140562,2907

Le R^2 est toujours satisfaisant (0.9350) ainsi que la valeur critique de F (8.7320E-14).
Les valeurs des probabilités t sont bons, ainsi que les valeurs seuils inférieures et supérieures (pas de changement de signe). Le signe des coefficients sont aussi logiques (plus la population est grande et plus le taux d'étudiants est grand, plus il y a de déplacements en émission pour le motif étudiant).

Nous pouvons maintenant forcer la constante à 0 (ce qui est logique car si la population est nulle ou si le nombre d'étudiants est nul, il n'y aura pas de déplacement en émission pour le motif étudiant.

Par ailleurs, il faut diviser le nombre de déplacements d'émission par la population totale pour avoir une valeur des coefficients cohérentes.

4ème régression



Statistiques de la régression	
Coefficient d	0,9875918
Coefficient d	0,975337564
Coefficient d	0,930787023
Erreur-type	0,017912918
Observations	25

ANALYSE DE VARIANCE

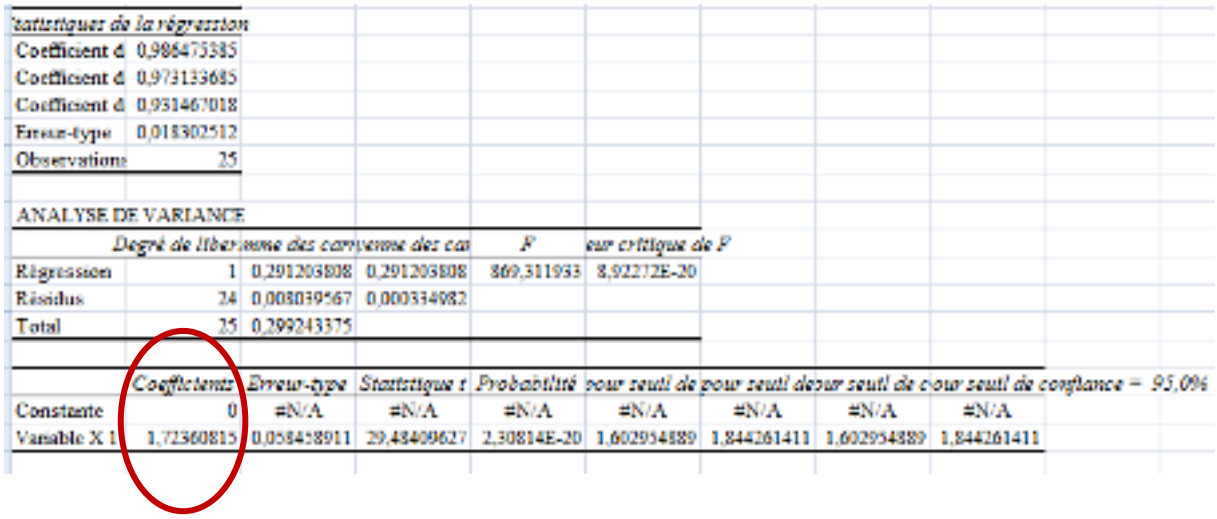
	Degré de liber	Somme des carr	Somme des car	F	ur critique de F
Régression	2	0,291863304	0,145931652	454,796186	1,27399E-18
Résidus	23	0,007380071	0,000320873		
Total	25	0,299243375			

	Coefficients	Erreur-type	Statistique t	Probabilité	pour seuil de	pour seuil de	our seuil de	our seuil de confiance = 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable X 1	-1,7672E-07	1,23267E-07	-1,43363896	0,165130938	-4,3172E-07	7,82771E-08	-4,3172E-07	7,82771E-08
Variable X 2	1,86475769	0,113872624	16,37582086	3,589E-14	1,629194221	2,100321159	1,629194221	2,100321159

Le R^2 est bon, ainsi que valeur seuil de F.
Cependant le signe du coefficient de la variable "population totale" n'est pas logique : en effet, plus la population est grande, plus le nombre de déplacement en émission doit être grand.

La régression finale comporte donc une unique variable : "le taux d'étudiant"

Régression finale :



Statistiques de la régression	
Coefficient d	0,986475385
Coefficient d	0,973133685
Coefficient d	0,931467018
Erreur-type	0,018302512
Observations	25

ANALYSE DE VARIANCE

	Degré de liber	Somme des carr	Somme des car	F	ur critique de F
Régression	1	0,291203808	0,291203808	869,311933	8,92272E-20
Résidus	24	0,008039567	0,000334982		
Total	25	0,299243375			

	Coefficients	Erreur-type	Statistique t	Probabilité	pour seuil de	pour seuil de	our seuil de	our seuil de confiance = 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable X 1	1,72360815	0,058458911	29,48409627	2,30814E-20	1,602954889	1,844261411	1,602954889	1,844261411

Le R^2 est excellent (0.9731), ainsi que le F de Fisher (8.922E-20).
La probabilité t de la variable est très inférieure à 5% et l'on trouve un coefficient de 1.72. Cela signifie qu'un étudiant réalise en moyenne 1.72 déplacements par jour pour le motif étudiant en émission. Cela semble cohérent, en comptant un aller-retour par jour par étudiant, moins les jour où les étudiants n'ont pas cours.

Attraction :

En attraction, nous allons réaliser le même type de raisonnement pour arriver à la meilleure régression.

1ère régression

La première régression utilise les même 8 variables qu'en émission.

tatistiques de la régression	
Coefficient	0,9608436
Coefficient	0,9232204
Coefficient	0,8848305
Erreur-type	3064,5317
Observatio	25

ANALYSE DE VARIANCE

	Degré de liberti	*nme des car*	*venne des ca*	*F*	*eur critique de F*
Régression	8	1,807E+09	225848648	24,04857	1,612E-07
Résidus	16	150261675	9391354,7		
Total	24	1,957E+09			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*	*pour seuil de*	*pour seuil de*	*our seuil de*	*our seuil de confiance =*	*95,0%*
Constante	-15185,606	12284,977	-1,236112	0,2342572	-41228,593	10857,38	-41228,593	10857,38	
Variable X	0,1167359	0,0353274	3,304397	0,0044776	0,0418451	0,1916267	0,0418451	0,1916267	
Variable X	34602,37	41808,927	0,827631	0,4200529	-54028,595	123233,33	-54028,595	123233,33	
Variable X	-46466,388	32215,97	-1,442340	0,168496	-114761,19	21828,416	-114761,19	21828,416	
Variable X	82554,461	40586,747	2,034025	0,0588754	-3485,5986	168594,52	-3485,5986	168594,52	
Variable X	12691,015	25884,614	0,490291	0,6305828	-42181,914	67563,945	-42181,914	67563,945	
Variable X	-0,0024494	0,3050095	-0,008030	0,993692	-0,6490406	0,6441419	-0,6490406	0,6441419	
Variable X	0,5106597	0,0882289	5,7878935	2,774E-05	0,3236227	0,6976966	0,3236227	0,6976966	
Variable X	8423,8307	9996,3492	0,8426907	0,411821	-12767,483	29615,144	-12767,483	29615,144	

Le R^2 est bon, ainsi que le coefficient F.

En enlevant les variables dont la probabilités t sont supérieures à 5%, il ne reste que les variables "population totale" et "nombre de places d'étudiants".

2ème régression

tatistiques de la régression	
Coefficient	0,9351107
Coefficient	0,874432
Coefficient	0,8254943
Erreur-type	3800,5482
Observatio	25

ANALYSE DE VARIANCE

	Degré de libert	nme des car	enne des ca	F	eur critique de F
Régression	2	2,313E+09	1,157E+09	80,083857	7,97E-11
Résidus	23	332215827	14444166		
Total	25	2,646E+09			

	Coefficients	Erreur-type	Statistique t	Probabilité	pour seuil de	pour seuil de	our seuil de c	our seuil de confiance = 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable X	0,0785164	0,0166299	4,7213879	9,322E-05	0,0441147	0,112918	0,0441147	0,112918
Variable X	0,6059961	0,0949142	6,3846708	1,627E-06	0,4096511	0,8023411	0,4096511	0,8023411

Régression finale

On peut se demander si en ne gardant que la variable "nombre de place étudiants", la régression sera meilleure.

Nous avons réalisé cette régression pour s'en assurer :

tatistiques de la régression	
Coefficient	0,8676013
Coefficient	0,7527321
Coefficient	0,7110654
Erreur-type	5220,9435
Observatio	25

ANALYSE DE VARIANCE

	Degré de libert	nme des car	enne des ca	F	eur critique de F
Régression	1	1,992E+09	1,992E+09	73,060701	1,352E-08
Résidus	24	654198016	27258251		
Total	25	2,646E+09			

	Coefficients	Erreur-type	Statistique t	Probabilité	pour seuil de	pour seuil de	our seuil de c	our seuil de confiance = 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable	0,8806437	0,1090287	8,5475553	9,587E-09	0,6580028	1,0932845	0,6580028	1,0932845

En ne gardant que la variable "nombre de place étudiants", on trouve un coefficient R^2 corrigé plus faible qu'avec les variables "nombre de places étudiants" et "population totale" (0.7110 contre 0.8254).

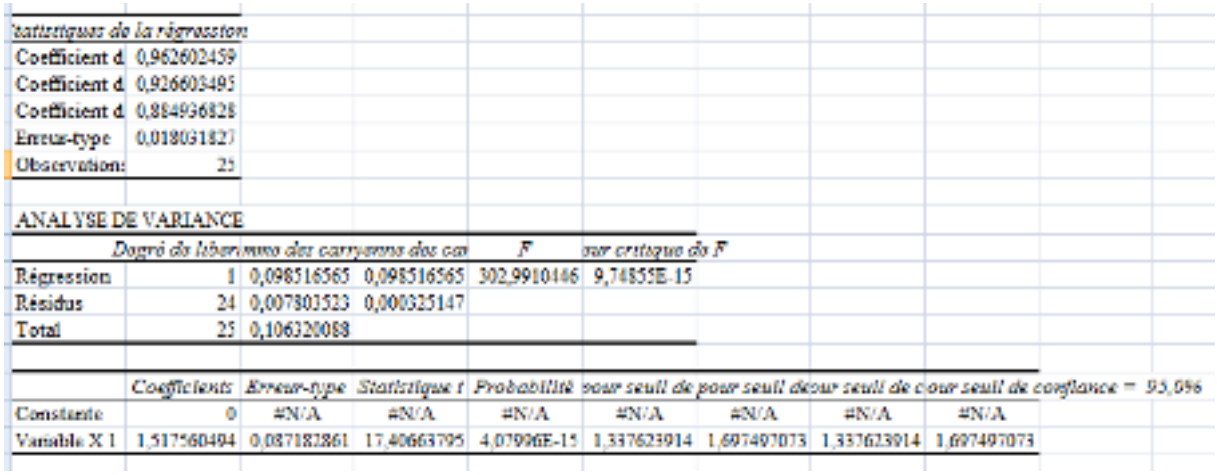
Cela signifie que la régression précédente est plus performante.

Cependant cette régression apporte plus d'information que la précédente : le coefficient apporte plus d'information, c'est-à-dire il y a 0.88 déplacement par jour par étudiant en attraction pour le motif étudiant.

Comparaison avec 1985

Emission

Nous avons repris la meilleure régression de 1995 et nous l'avons appliqué aux données de 1985.



Statistiques de la régression	
Coefficient d	0,962602459
Coefficient d	0,926603495
Coefficient d	0,884936828
Erreur-type	0,018031827
Observations	25

ANALYSE DE VARIANCE

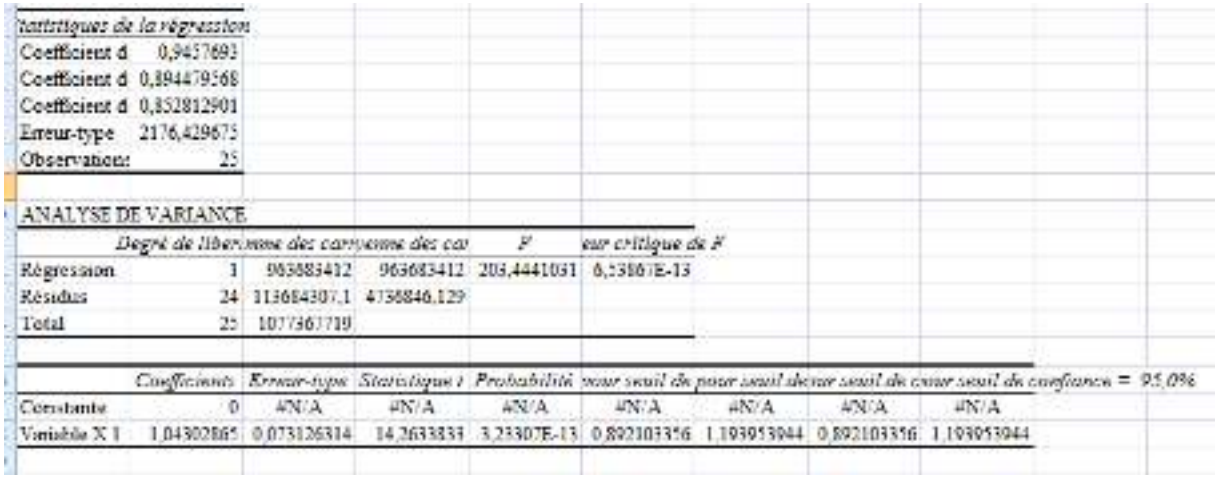
	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	1	0,098516565	0,098516565	302,9910446	9,74855E-15
Résidus	24	0,007803523	0,000325147		
Total	25	0,106320088			

	Coefficients	Erreur-type	Statistique t	Probabilité	Inférieur 95%	Supérieur 95%	Inférieur 95,0%	Supérieur 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable X 1	1,517560494	0,087182861	17,40663795	4,07996E-15	1,337623914	1,697497073	1,337623914	1,697497073

On trouve un R^2 de 0.9266 et une valeur critique bonne de 9.748E-15 avec une bonne probabilité inférieure à 5%.

Le coefficient est de 1.52, ce qui est à peu près similaire au coefficient trouvé en 1995 de 1.72. Cela signifie qu'entre 1985 et 1995, le nombre de déplacements de par jour a légèrement augmenté.

Attraction



Statistiques de la régression	
Coefficient d	0,9457693
Coefficient d	0,894479568
Coefficient d	0,852812901
Erreur-type	2176,429675
Observations	25

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	1	963683412	963683412	203,4441031	6,23867E-13
Résidus	24	113684307,1	4736846,129		
Total	25	1077367719			

	Coefficients	Erreur-type	Statistique t	Probabilité	Inférieur 95%	Supérieur 95%	Inférieur 95,0%	Supérieur 95,0%
Constante	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Variable X 1	1,04302865	0,073126314	14,2633833	3,23307E-13	0,892103356	1,193953944	0,892103356	1,193953944

On trouve un R^2 bon de 0.8944 et une bonne valeur critique de F.

Le coefficient est de 1.04, ce qui est légèrement plus important qu'en 1995. Cela signifie que le nombre de déplacement à diminué entre 1985 et 1995 pour le motif étudiant.