Multiple Linear Regression MLR

Startups Data

Pour notre data on aura : $Y= a0 + a1*x1 + a2*x2 + a3*x3 + a4*x4$

➔ Comme x4 est une variable qualitative, on la transforme en **DUMMY variables.**
Le DUMMY variables occasionnent la création d'autres variables ce qui rend nos prédictions difficiles à interpréter.

Manières de construire un modèle :

-**All in** : On intègre toutes les variables indépendantes dans un MLR.

Stepwise Regression (Bidirectional Elimination, Backward Elimination, Forward Selection)

-**Backward Elimination :**

Step1 choisir un seuil SL pour rester dans le modèle (e.g. SL= 0.05).
Step 2 Remplir le modèle de tous les prédicteurs possibles.
Step 3 Considérer le prédicteur ayant la plus grande p-value.
Si p-value > SL, aller au step4 sinon c'est FINI.
Step 4 Enlever le prédicteur.
Step 5 Ajuster le modèle sans cette variable.
Retourner au Step3 et recommencer jusqu'à ce que p-value < SL.

-**Forward Selection :**

Step1 choisir un seuil SL pour entrer dans le modèle (e.g. SL= 0.05).
Step 2 Ajuster tous les modèles simples de régression Y~Xn,
sélectionner celui avec la plus petite p -value.
Step 3 Garder cette variable et ajuster tous les modèles possibles avec un prédicteur en plus.
Step 4 Considérer le prédicteur ayant la plus petite p-value.
Si p-value < SL, aller au Step 3 sinon c'est FINI.
Recommencer jusqu'à ce que p-value > SL.

-**Bidirectional Elimination :**

Step1 choisir deux seuil pour entrer (SLenter = 0.05) et rester (SLstay= 0.05)dans le modèle.
Step 2 Effectuer le next step de la Forward Selection
(les nvlles variables doivent vérifier : p -value< SLenter pour entrer dans le modèle)
Step 3 Effectuer TOUS les steps de la Backward Elimination.
(les vielles variables doivent vérifier p-value< SLstay pour rester dans le modèle)
Recommencer au Step 2 jusqu'à ce que les conditions respectives sur p ne soient re.

Step 4 Aucune nouvelle variable peut entrer et aucune ancienne variable peut sortir.

-**Score Comparison :**

Step1 choisir un criterede qualité d'ajustement (ex : critère d'Akaike).
Step 2 Construire tous les modeles de regression possibles : 2^{N}-1 combinaisons au total.
Step 3 Choisir celui ayant le meilleur critère.

Remarques :

- ➔ Pour 10 colonnes le score Comparison donne 1023 alors ce modèle demande bcp de temps donc a éviter lorsqu'on a bcp de variables.
- ➔ Le modèle le plus utilisé est **Backward Elimination.**

Exemple Backward Elimination sur Startups Data

- ➔ Dummify la variable State, on considère SL=0.05, on prend comme variable dependante "Profit" et comme variables indépendantes: 'RDSpend', 'Administration ', 'MarketingSpend' et que 'New york' ou 'Californie' pour représenter 'State' afin d'éviter la multi colinéarité.

Model 1: OLS, using observations 1-50
Dependent variable: Profit

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	50416.5	6653.54	7.577	<0.0001	***
RDSpend	0.807956	0.0457466	17.66	<0.0001	***
Administration	−0.0236200	0.0518559	−0.4555	0.6509	
MarketingSpend	0.0263692	0.0166783	1.581	0.1209	
NewYork	−1332.09	2690.18	−0.4952	0.6229	

Mean dependent var	112012.6	S.D. dependent var	40306.18
Sum squared resid	3.90e+09	S.E. of regression	9309.026
R-squared	0.951013	Adjusted R-squared	0.946659
F(4, 45)	218.4023	P-value(F)	7.53e-29
Log-likelihood	−525.2499	Akaike criterion	1060.500
Schwarz criterion	1070.060	Hannan-Quinn	1064.140

Analyse : RDSpend est très significative pour le Profit.

- ➔ La p-value de 'Administration ' est la plus grande, on l'enlève du modèle.

Model 2: OLS, using observations 1-50
Dependent variable: Profit

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	47721.8	3018.34	15.81	<0.0001	***
RDSpend	0.800294	0.0421740	18.98	<0.0001	***
MarketingSpend	0.0285947	0.0158086	1.809	0.0770	*
NewYork	−1484.61	2646.17	−0.5610	0.5775	

Mean dependent var	112012.6	S.D. dependent var	40306.18
Sum squared resid	3.92e+09	S.E. of regression	9228.486
R-squared	0.950787	Adjusted R-squared	0.947578
F(3, 46)	296.2378	P-value(F)	4.44e-30
Log-likelihood	−525.3649	Akaike criterion	1058.730

Schwarz criterion	1066.378	Hannan-Quinn	1061.642

Analyse : RDSpend reste très significative pour le Profit.

➔ La p-value de 'New York ' est la plus grande, on l'enlève du modèle.

Model 3: OLS, using observations 1-50
Dependent variable: Profit

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	46975.9	2689.93	17.46	<0.0001	***
RDSpend	0.796584	0.0413476	19.27	<0.0001	***
MarketingSpend	0.0299079	0.0155200	1.927	0.0600	*

Mean dependent var	112012.6	S.D. dependent var	40306.18
Sum squared resid	3.94e+09	S.E. of regression	9160.966
R-squared	0.950450	Adjusted R-squared	0.948342
F(2, 47)	450.7713	P-value(F)	2.16e-31
Log-likelihood	−525.5354	Akaike criterion	1057.071
Schwarz criterion	1062.807	Hannan-Quinn	1059.255

Analyse : RDSpend reste très significative pour le Profit.

➔ Il nous reste que deux variables RDSpend et MarketingSpend , celle qui a la plus grande p-value est MarketingSpend mais sa p-value est > 0.05 alors on l'enlève.

Model 4: OLS, using observations 1-50
Dependent variable: Profit

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	49032.9	2537.90	19.32	<0.0001	***
RDSpend	0.854291	0.0293056	29.15	<0.0001	***

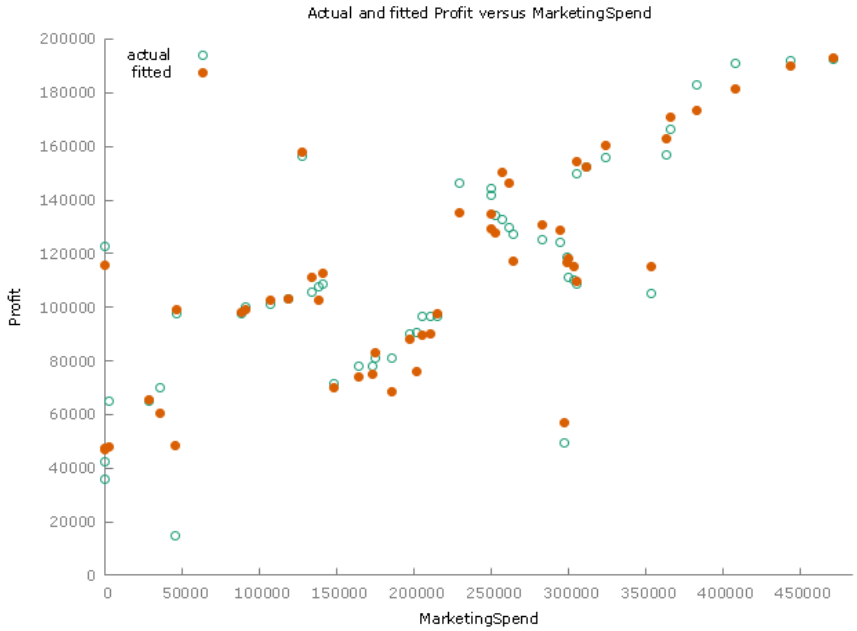
Mean dependent var	112012.6	S.D. dependent var	40306.18
Sum squared resid	4.26e+09	S.E. of regression	9416.349
R-squared	0.946535	Adjusted R-squared	0.945421
F(1, 48)	849.7889	P-value(F)	3.50e-32
Log-likelihood	−527.4365	Akaike criterion	1058.873
Schwarz criterion	1062.697	Hannan-Quinn	1060.329

Conclusion : Il nous reste que RDSpend et de plus sa p-value étant très petite (<0.05), On s'arrête donc le model 4 est le mieux qui exprime la dépendance entre le Profit et les variables indépendantes au départ.

Remarque :

Au niveau du modèle 3, on a remarqué que la p-value de MarketingSpend est égale a 0.06 ce qui est très proche du SL=0.05, ce constat nous interpelle car si l'on avait choisi un SL plus grand on aurait pas enlever la variable MarketingSpend.

Représentons alors les prédictions du Profit avec que MarketingSpend :



On observe que même si les données d'observations sont dispersées, elles semblent s'aligner le long d'une droite(Cela qui indique qu'il y'a un bon niveau de significativité pour MarketingSpend).

Ce constat visuel nous met un doute sur la non-utilisation de la variable MarketingSpend dans le modèle final. Alors comment améliorer notre modèle.

Nous allons donc nous fier sur le Adjusted R-squared (Critère du (R^2) : plus le (R^2) est grand, plus le modèle prédit mieux). Dans notre cas, l'on peut remarquer du modèle 1 au modèle 3, le (R^2) augmente, i.e. le modèle prédictif s'est amélioré au fur a mesure. Au modèle 4, le (R^2) a diminué . cela veut donc dire que le modèle 3 est plus robuste que le modèle 4. Pour conclure, on pourra dire la variable MarketingSpend n'aurait pas dû être retire au modèle 3, vu que le modèle fonctionnait mieux que les autres.

Remarque : Critère d'Akaike => Plus il diminue, mieux le modèle prédit.

Interprétations des coefficients de la régression linéaire multiple.

Le modèle 3 étant le meilleur, on a :

Model 3: OLS, using observations 1-50
Dependent variable: Profit

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	46975.9	2689.93	17.46	<0.0001	***
RDSpend	**0.796584**	0.0413476	19.27	<0.0001	***

MarketingSpend	**0.0299079**	0.0155200	1.927	0.0600	*

→Comme les coefficients RDSpend et MarketingSpend sont (>0), alors la variable indépendante (Profit) évolue dans le même sens que la variable dépendante(RDSpend et MarketingSpend).

→ RDSpend a un plus gros impact sur le Profit **par unité de RDSpend** que le MarketingSpend

Par unité de MarketingSpend.