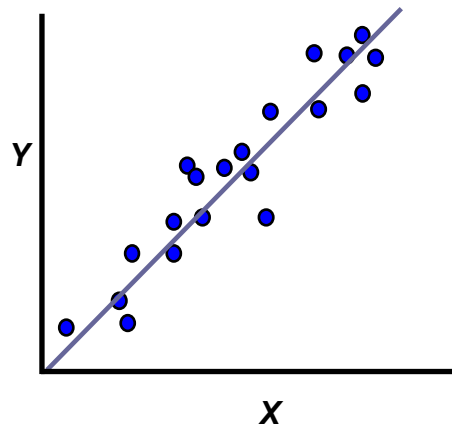
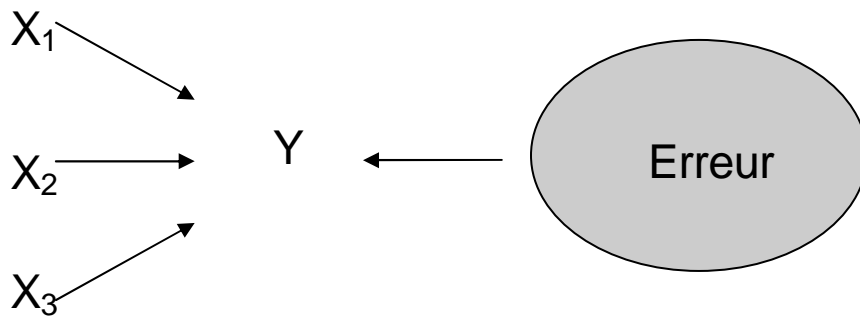


Méthodes de régression

Régression linéaire simple et multiple



Copyright Creascience Inc.

Tous droits réservés. Aucune partie de cet ouvrage ne peut être reproduite, enregistrée ou retransmise sous aucune forme, ni par aucun procédé électronique, mécanique, photocopique ou autres, sans la permission expresse de Creascience Inc.

Creascience Inc
3947 Saint-Hubert
Montréal, Qc
Canada
H2L 4A6

Téléphone : (514) 840-9220

© 2010

Table des matières

1	RÉGRESSION LINÉAIRE SIMPLE (RLS)	1
1.1	Objectifs de la régression linéaire simple	1
1.2	Terminologie.....	1
1.3	Qu'est-ce qu'un modèle ?	2
1.4	Spécification du modèle en RLS	2
1.5	Principe d'estimation des moindres carrés	3
1.6	Interprétation des coefficients des modèles de régression	3
1.7	Différence entre la corrélation et la régression.....	4
1.8	Exemple de RLS avec les données sur la pression sanguine.....	5
1.9	Test statistique sur β_0	8
1.10	Test statistique sur β_1	9
1.11	Qualité de l'ajustement du modèle	9
1.12	Conditions d'utilisation du modèle et diagnostics.....	13
1.13	Le problème des observations influentes.....	24
1.14	Prédiction en régression.....	28
1.15	Extrapolation	32

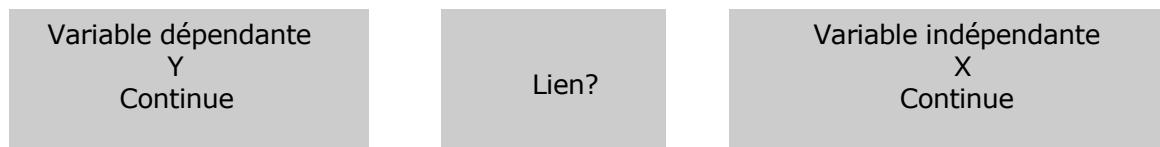
2	RÉGRESSION LINÉAIRE MULTIPLE	33
2.1	Objectifs de la RLM	33
2.2	Aspects communs à la RLS et à la RLM	33
2.3	Interprétation des coefficients de régression partiels β_i	34
2.4	Construction d'un modèle de RLM.....	34
2.5	Une statistique importante pour mesurer la qualité du modèle : le coefficient de détermination ajusté	37
2.6	Vérification de l'adéquation du modèle.....	38
2.7	Sélection de variables.....	45
2.8	Multicollinéarité.....	56
2.9	Termes spéciaux dans les modèles de régression multiple	63
3	ALTERNATIVES À LA RÉGRESSION LINÉAIRE STANDARD.....	69
3.1	Régression non-linéaire (RNL).....	69
3.2	Applications de la régression non-linéaire	69
3.3	Autres manières de lutter contre la multicollinéarité	70
4	RÉSUMÉ.....	71
4.1	Étapes dans la construction d'un modèle de régression.....	71
4.2	Robustesse de la régression aux déviations des hypothèses sous-jacentes	72
4.3	Quelques références.....	72

1 Régression linéaire simple (RLS)

1.1 Objectifs de la régression linéaire simple

Le but de la régression linéaire simple (RLS) consiste à relier une variable dépendante, communément notée Y , à une seule variable explicative ou indépendante, dénotée X .

En régression linéaire simple, les deux variables à lier sont de nature continue. Cela signifie que le nombre des valeurs possibles qu'elles peuvent prendre est très grand par opposition à une variable catégorique ou discrète.



L'objectif de l'analyse de régression linéaire simple est de résumer la relation entre les deux variables à l'aide d'une équation mathématique.

De plus, il arrive que l'on connaisse une valeur donnée de la variable explicative X et que l'on veuille prédire la valeur de Y . L'équation mathématique de la régression linéaire permet de déterminer quelle serait cette valeur. Afin d'effectuer de la prédiction à l'aide du modèle de régression, il faut tout d'abord s'assurer que celui-ci s'ajuste assez bien aux données.

Exemple : En utilisant la relation entre les dépenses en publicité et les ventes, il est possible de prédire les ventes à partir de l'équation de régression linéaire. Une telle prédiction sera valide et utile s'il existe un lien linéaire entre les deux variables et si le modèle fournit une prédiction relativement précise.

1.2 Terminologie

Dans les logiciels statistiques, plusieurs termes sont utilisés pour faire référence aux variables Y et X . Le tableau suivant présente les plus fréquents.


	Y	X
	Variable dépendante	Variable indépendante
	Variable réponse	Prédicteur
	Variable endogène	Variable explicative
		Variable exogène

Tableau 1 : Différents termes utilisés pour faire référence à Y et X

1.3 Qu'est-ce qu'un modèle ?

Les modèles sont utilisés dans toutes les analyses de régression. Il est donc important de comprendre pourquoi ils sont utilisés et comment ils fonctionnent.

Définition tirée du petit Robert

Sc. Représentation simplifiée d'un processus, d'un système.

Modèle mathématique, modèle formé par des expressions mathématiques et destiné à simuler un tel processus.

Il faut réaliser qu'un modèle explique rarement **entièrement** un phénomène complexe à l'étude. En ce sens, un modèle est sujet à l'erreur, ce qui signifie qu'aucun modèle n'est parfait !

En RLS, une approximation importante est effectuée : le modèle linéaire résume le lien entre deux variables continues à l'aide d'une droite. Cette approximation ne représente pas le meilleur modèle pour toutes les situations puisque tous les phénomènes ne sont pas strictement linéaires.

Le choix d'un modèle mathématique qui décrit adéquatement la relation entre les variables implique un compromis entre la complexité et le pouvoir de prédiction du modèle.



1.4 Spécification du modèle en RLS

Dans la section précédente, il a été mentionné qu'en RLS l'équation de la droite est utilisée afin de relier la variable indépendante (X) et la variable dépendante (Y).

L'idée est d'établir une relation de la forme suivante :

$$Y = b_0 + b_1X$$

Étant donnée les erreurs sur les Y, les point expérimentaux (x_k, y_k) ne tombent pas tous directement sur la droite de régression. Comment déterminer la meilleure droite ?

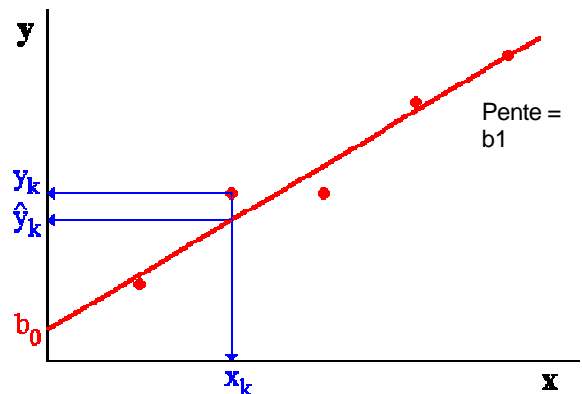


Figure 1 : Droite de régression représentant le lien linéaire entre X et Y

1.5 Principe d'estimation des moindres carrés

Idée : Trouver la droite qui passe le plus près possible des points.

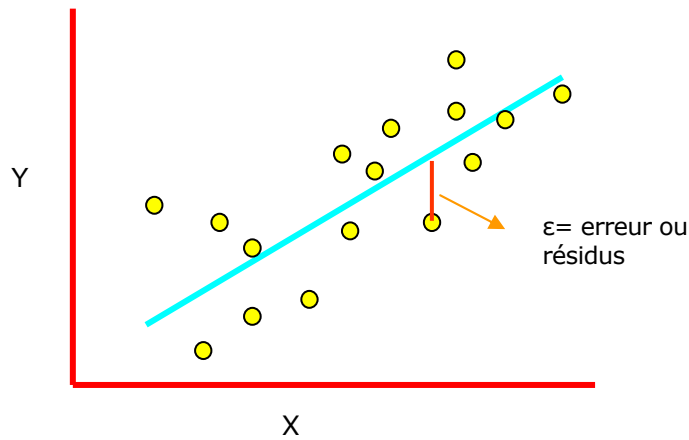


Figure 2 : Principe des moindres carrés

La régression produit la droite qui fournit le « meilleur » ajustement aux données

► **Critère des moindres carrés** : Minimiser les différences entre les valeurs observées et les valeurs prédites par le modèle (distances verticales).

Une seule solution des moindres carrés est trouvée pour chaque jeu de données analysé.

1.6 Interprétation des coefficients des modèles de régression

Le modèle de régression linéaire simple est le suivant : $Y = \beta_0 + \beta_1 X$

β_0

Le premier coefficient de l'équation, dénoté par β_0 , représente l'ordonnée à l'origine du modèle. C'est la valeur que prend Y lorsque $X = 0$.

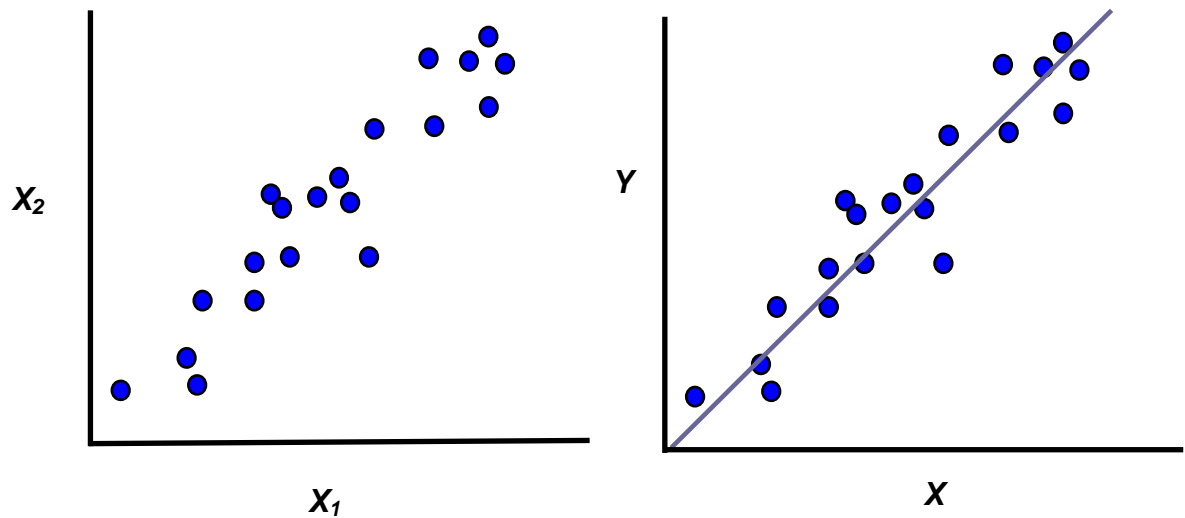
β_1

Le deuxième coefficient de l'équation, dénoté par β_1 , correspond à la pente de la droite de régression. La pente constitue un paramètre important puisqu'elle mesure l'association linéaire entre Y et X. Ce paramètre permet de quantifier le changement moyen en Y lorsque X augmente d'une unité.

Il est important de noter que les coefficients de la régression sont estimés à partir des données. Utiliser un autre jeu de données résulte en des estimations différentes.

1.7 Différence entre la corrélation et la régression

Corrélation	Régression
Mesure d'association entre 2 variables	Lien entre une variable dépendante et une variable indépendante
X_1 et X_2	Y et X
Pas de relation causale	Relation causale
	Utilisée dans un but de prédiction



1.8 Exemple de RLS avec les données sur la pression sanguine

1.8.1 Description des données

Le jeu de données suivant contient deux variables : la pression sanguine systolique du sujet ainsi que son âge. Le nombre de sujets dans le fichier de données est 69.

Variable réponse (Y) : pression sanguine systolique (SBP)

Variable explicative (X) : âge (age)

Nombre d'observations (subject) : n=69



Le tableau ci-dessous contient un extrait du fichier de données.

Obs	SUBJECT	AGE	SBP
1	1	39	144
2	2	45	138
3	3	47	145
4	4	65	162
5	5	46	142
6	6	67	170
. . .			
30	1	41	158
31	2	36	124
32	3	65	158
33	4	60	185

X Y

Objectifs de l'analyse de régression :

1. Déterminer le pouvoir explicatif de l'âge sur SBP
2. Prédire SBP à l'aide d'une équation mathématique et estimer l'erreur de prédiction

