Régression linéaire simple

L'analyse de régression linéaire simple permet de **quantifier le lien de causalité** entre deux variables pour, entre autre, pouvoir **faire des prédictions**. Par exemple, soit le **nombre de questions de référence** reçues par un bibliothécaire par jour ainsi que le **nombre de recherches automatisées** que ce même bibliothécaire a faites (cf tableau à droite). Est-ce que le nombre de recherches automatisées dépend du nombre de questions de référence? Si oui, comment?

Pour le savoir, il faut faire une **analyse de régression linéaire simple** en utilisant le nombre de recherches automatisées comme variable dépendante (Y) et le nombre de questions de référence comme variable indépendante (X).

Données

nombre de questions de référence	nombre de recherches automatisées
10	7
8	7
6	8
4	2
11	10
3	5
6	6
5	3

Analyse de régression linéaire simple dans Excel [*Utilitaire d'analyse – Régression linéaire*]

Résultats retournés par Excel avec l'utilitaire d'analyse Régression linéaire

Statistiques de la régression	
Coefficient de détermination multiple	**0,77237926**
Coefficient de détermination R^2	**0,59656972**
Coefficient de détermination R^2	0,529331345
Erreur-type	1,796508338
Observations	8

Coefficient de corrélation (r)
Valeur = 0,8, donc une relation trèsforte entre le nombre de questions de référence et le nombre de recherches automatisées. Pour pouvoir connaître le sens de la relation, il faut jeter un coup d'oeil sur la pente de la droite de régression (b) plus bas car Excel dans cet utilitaire nous donne le coefficient de corrélation en valeur absolue... Ici, c'est donc positif!

Coefficient de détermination R^2
(carré du coefficient de corrélation)
Proportion de Y qui peut être expliquée par X (pourcentage)
Plus c'est élevé, plus X est utile pour expliquer Y

Ici : = 60% c'est-à-dire que les variations du nombre de recherches automatisées s'expliquent à 60% par les variations du nombre de questions de référence

ANALYSE DE VARIANCE					
	Degré de liberté	*mme des carvenne des ca*		*F*	*Valeur critique de F*
Régression	1	28,6353468	28,6353468	8,87245841	**0,024679143**
Résidus	6	19,3646532	3,22744221		
Total	7	48			

Valeur critique de F
Permet de vérifier si le lien observé est significatif ou simplement dû au hasard. Si plus petit ou égal au seuil critique, c'est significatif; sinon c'est dû au hasard. Habituellement on prend un seuil critique de 0,05 (5%). Ici : =0,02 < 0,05 donc c'est significatif, il y a bien un lien entre les deux

	$\bar{i}$ *Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95,0%*	*Limite supérieure pour seuil de confiance = 95,0%*
Constante	**1,25727069**	1,71424417	0,73342568	0,49095063	-2,937333664	5,451875051	-2,93733366	5,45187505
nombre de questions de référence	**0,71588367**	0,24033691	2,97866722	0,02467914	0,127800442	1,303966896	0,12780044	1,3039669

Ordonnée à l'origine de la droite de régression (a)

Pente de la droite de régression (b)

Ici, la droite de régression serait (en utilisant les coefficients *a* et *b*): Y = 1,26 + 0,72 X
Ce qui se traduit par : Nombre de recherches automatisées = 1,26 + 0,72 * Nombre de questions de référence

Utilité? Pour faire des prédictions! Par exemple, combien de recherches automatisées un bibliothécaire doit-il s'attendre à faire s'il reçoit 22 questions de référence? Nombre de recherches automatisées = 1,26 + 0,72 * 22 = 17,1 (~17 recherches automatisées)

Graphique présentant : (1) les données observées (données originales), et (2) les données "projetées" i.e. les données correspondant à la droite de régression linéaire modélisant la relation entre la variable dépendante et la variable indépendante.