# REGRESSION LINEAIRE SIMPLE

Ref : « exploration des données – Méthodes et modèles du data mining » Daniel T. Larose.  
<http://spss.espaceweb.usherbrooke.ca/pages/stat-inferentielles/regression-simple.php>  
site de Frédéric Bertrand, maître de conférences, université de Strasbourg (+ Myriam Maumy)  
« Statistiques pour l’économie et la gestion » Anderson-Sweeney-Williams-Camm-Cochran Ed. de boeck

## Introduction :

Alliance Data Systems (ADS) fournit des moyens de traitement des transactions, des services de crédit et des services de marketing à ses clients dans le domaine de la gestion des relations client, aujourd’hui en croissance. Les clients de ADS sont concentrés dans quatre secteurs : le commerce de détail, les stations-service les services publics et les transports . En 1983, Alliance a commencé à proposer des services de traitement des crédits aux entreprises appartenant aux secteurs du commerce de détail (y compris les stations-service) et de la restauration : cette société emploie aujourd’hui plus de 6500 personnes et offre ses services à des clients à travers le monde. Gérant de plus de 140000 points de vente aux Etats-Unis, ADS traite plus de 2,5 milliards de transactions par an. La société se place au deuxième rang des sociétés américaines privées de services de crédit. En 2001, ADS a fait une première offre publique d’achat et est maintenant cotée à la bourse de New-York.

L’un des services marketing d’ADS consiste à élaborer des campagnes promotionnelles par courrier. Grâce à sa base de données contenant des informations sur les habitudes d’achat de plus de 100 millions de consommateurs, ADS peut cibler les consommateurs qui seront les plus sensibles à une campagne promotionnelle. Le bureau de développement analytique utilise les méthodes de régression pour construire des modèles permettant de mesurer et de prévoir la sensibilité des consommateurs à des campagnes marketing ciblées. Certains modèles de régression prédisent la probabilité d’achat des individus recevant une réduction, d’autres prédisent le montant dépensé par les consommateurs qui effectuent un achat.

Lors d’une campagne promotionnelle particulière, une chaine de magasins souhaitait attirer de nouveaux consommateurs. Pour prévoir l’effet de la campagne, les analystes de ADS ont sélectionné un échantillon de consommateurs dans leur base de données, ont envoyé à ces individus des bons d’achat et ont ensuite collecté des données sur les transactions de ces clients : le montant ‘achat ainsi que plusieurs variables spécifiques à chaque consommateur susceptibles d’être utiles pour prévoir les ventes. La variable spécifique à chaque consommateur la plus pertinente pour prévoir le montant des achats, était le montant total des dépense s effectuées dans des magasins similaires au cours des 39 derniers mois. Les analystes d’ADS ont effectué une régression entre le montant des achats et le montant dépensé dans des magasins similaires : où correspond au montant des achats et le montant estimé des achats et x le montant dépensé dans des magasins similaires.

En utilisant cette équation, nous pouvons prédire qu’une personne qui a dépensé 10000 dollars au cours des 39 derniers mois dans des magasins similaires, dépensera 47,2 dollars en réponse à la campagne promotionnelle ciblée. Dans ce chapitre, vous apprendrez à effectuer ce type de régression.

NB : Le modèle final développe par les analystes de AS incluait également plusieurs autres variables, augmentant ainsi le pouvoir prédictif de l’équation précédente, telles que la possession ou non d’une carte de crédit bancaire, le revenu estimé et le montant moyen dépensé par visite dans un magasin particulier. Dans le chapitre suivant, nous verrons comment de telles variables additionnelles peuvent être incorporées dans un modèle de régression multiple.

## Objectifs

Les décisions prises par un responsable sont souvent basées sur la relation qui existe entre deux ou plusieurs variables. Par exemple, après avoir considéré la relation entre les dépenses publicitaires et les ventes, un responsable marketing peut essayer de prévoir les ventes pour un montant donné de dépenses publicitaires. Autre exemple, un fournisseur d’électricité peut se servir de la relation entre la température journalière maximale et la demande d’électricité pour prévoir la demande en électricité, en se basant sur les températures journalières maximales prévues pour le mois suivant. Parfois, un responsable peut se fier à son intuition pour déterminer le type de relation qui lie deux variables. Cependant, s’il est possible d’obtenir des données, une procédure statistique, appelée analyse de la régression, permet de construire une équation indiquant de quelle manière les variables sont liées.

Dans la terminologie utilisée dans le cadre d’une analyse de la régression, la variable que l’on cherche à prévoir, notée y, est la variable dépendante ou variable cible ou variable expliquée. La variable ou les variables utilisées pour prévoir la valeur de la variable dépendant sont appelées variables indépendantes ou variables explicatives.

Décider : de réformer le matériel car on a mis en relation le temps d’utilisation (y) et l’usure (x). Ainsi, il est possible d’anticiper la décision.

Prévoir : la consommation d’énergie nécessaire pour une production prévue (si on a mis en relation la production (x) et la consommation d’énergie (y)).

En maintenance, prévoir un modèle qui prédit la durée de vie d’un composant sur le site où il se trouve : on peut déterminer lors de petites défectuosités, des caractéristiques (des facteurs) qui influencent celles-ci et donc prédire la suite du fonctionnement.

Contrôler : exemple : dans les tunnels où on utilise de l’explosif pour creuser, on peut mettre en relation l’avancement des travaux du tunnel et la consommation d’explosifs. On peut utiliser cette relation pour voir si on n’a pas volé des explosifs.

## PRINCIPE DE LA REGRESSION DES MOINDRES CARRES:

Nous disposons de n points (x1,y1),(x2,y2),…,(xn,yn) et nous souhaitons trouver la « meilleure » droite.   
Comment trouver la « meilleure droite » ?

Voici le nuage de points :

Parmi toutes les droites qui approchent un ensemble donné de points, celle vérifiant la propriété suivante :  
 est minimum   
est la meilleure droite d’ajustement. (b : ordonnée origine) (a : pente de la droite)

Ceci revient au même de minimiser f(a,b)=

Une droite présentant cette propriété est dite s’ajuster aux données au sens des moindres carrés et est appelée droite de régression des moindres carrés. Après démonstration, on obtient :

Donc et

L’équation estimée de la régression linéaire simple :

b s’appelle l’intercepte. C’est l’endroit sur l’axe des y où la droite de régression croise l’axe des y, ce qui donne la valeur estimée de la variable cible quand la variable prédictive est égale à 0. Dans certaines situations de régression, une valeur 0 pour la variable prédictive n’a pas de sens. Par exemple, si la variable cible est le poids des élèves avec comme variable prédictive la taille, l’image de la valeur 0 de la variable prédictive n’a pas d’interprétation. Par contre, dans l’exemple « céréales » (sucre – taux nutritionnel), une valeur de 0 pour la teneur en sucre a du sens et, dans ce cas, le taux nutritionnel estimé pour une céréale ayant une teneur en sucre de 0 est de 59,3.

a est la pente de la droite de régression. Cette pente indique le changement estimé dans la variable cible par unité d’accroissement de x. Dans l’exemple « céréales », nous interprèterons a = -2.4008 comme suit : pour tout accroissement de 1gr dans la teneur en sucre, le taux nutritionnel estimé décroît de 2,4008 points.

## PARAMETRES DE LA REGRESSION et INFERENCE

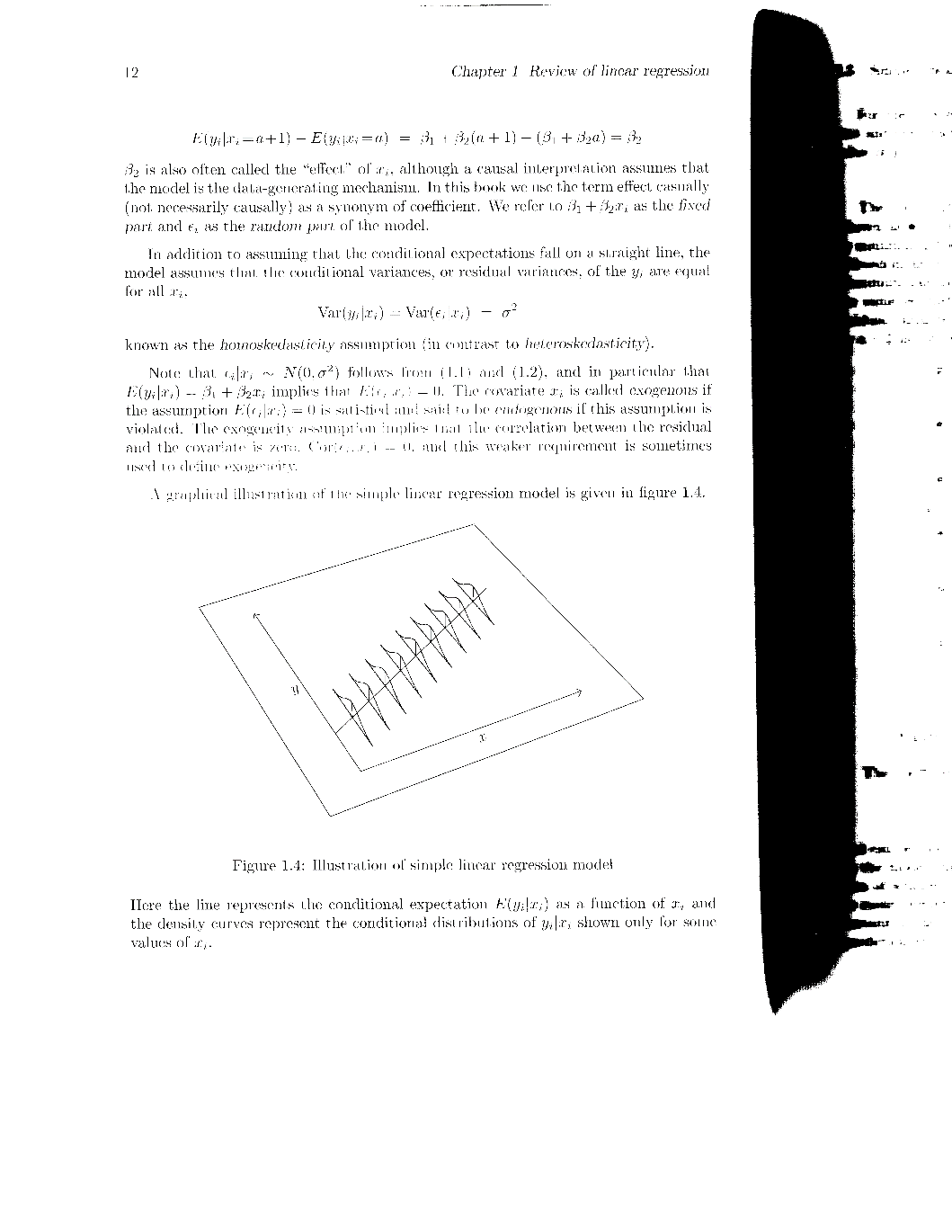
Bien sûr, une droite de régression des moindres carrés pourra toujours être trouvée pour estimer la relation entre deux variables continues mais cela ne garantit pas que cette régression soit utile. La question est « comment peut-on déterminer si une équation de régression estimée particulière est utile pour faire des prévisions » ?

Dans l’exemple « sucre – taux nutritionnel», si l’idée est de réaliser une régression linéaire sur les quelques valeurs données, nous nous arrêtons là. Si l’idée est que ces valeurs représentent un échantillon pour créer un modèle qui lie la taux nutritionnel d’une céréale au sucre qu’elle contient de manière générale alors la lecture sera :

nous possédons un échantillon de taille N et nous cherchons un modèle linéaire de type dans lequel

* le meilleur estimateur de est
* le meilleur estimateur de est
* est une variable qui représente le comportement individuel

Si nous avions accès à toute l ‘information, nous pourrions constater qu’une même abscisse correspond à différentes valeurs de .

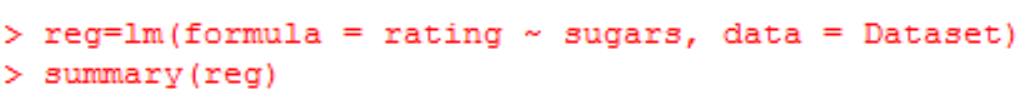


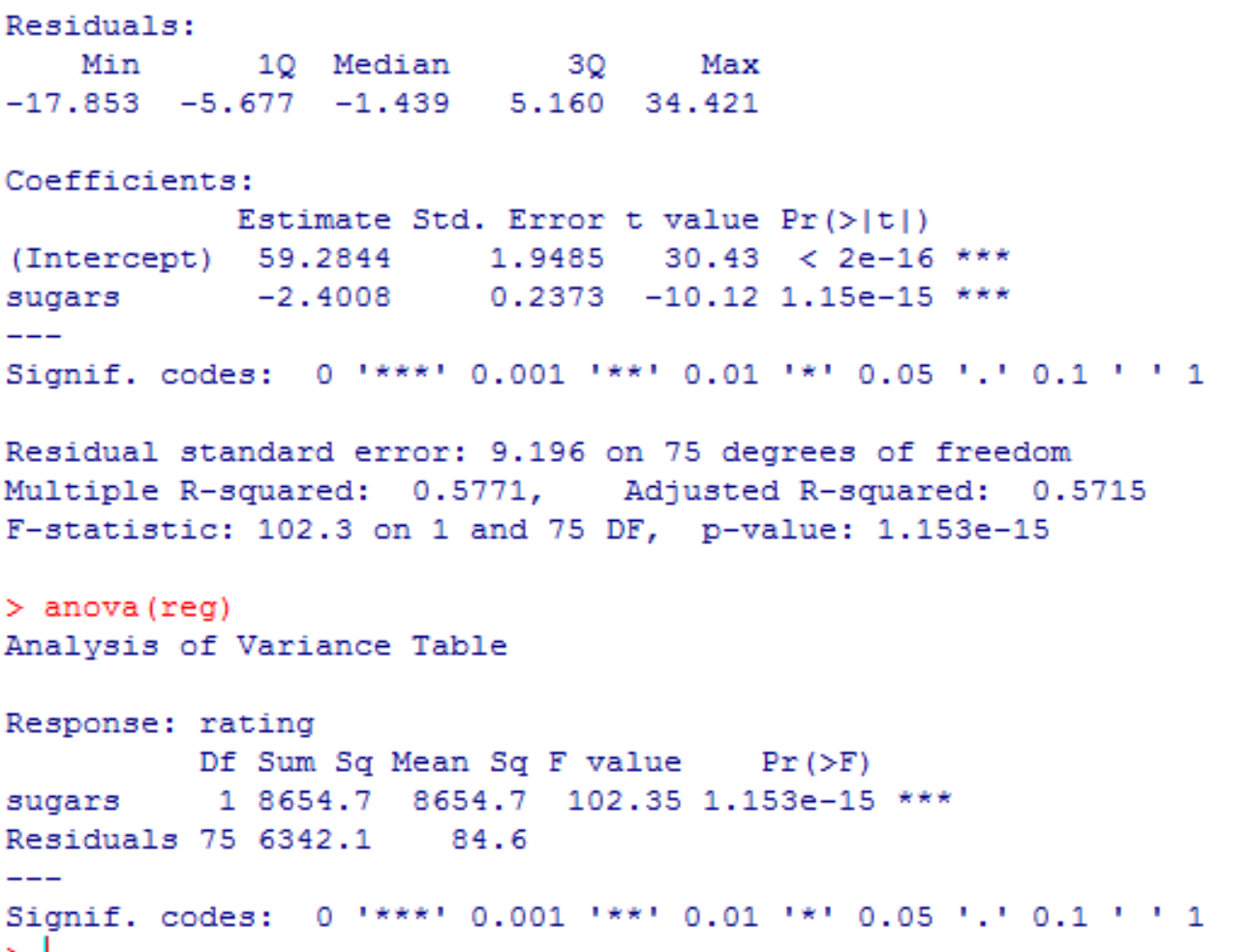
Ainsi, lorsque est fixé, est aléatoire et la composante aléatoire de est le .

Plusieurs droites de régression existent. La droite des moindres carrés consiste à minimiser, en principe, mais vu que nous ne les connaissons pas, cette somme sera estimée par = (cfr ci-dessus). Les estimateurs de sont respectivement et . Nous noterons donc la valeur estimée correspondant à .

Les logiciels statistiques donnent quelques indicateurs permettant de se faire une idée sur la pertinence de la régression et l’utilité du modèle.

### Analyse de la régression dans R





### Coefficient de détermination.

L’indicateur statistique sert à mesurer la qualité de l’ajustement de la régression càd mesure comment l’approximation linéaire s’ajuste réellement aux données observées.

Calcul :

* Si représente la valeur estimée de la variable cible, alors représente l’erreur de prévision ou le résidu,   
  SSE (som of squares error) =SCE (somme des carrés des erreurs) = représente une mesure globale de l’erreur dans la prévision résultant de l’usage de l’équation de régression estimée.   
  Dans notre exemple « céréales », SSE=6342,1. Est-ce beaucoup ?

* Supposons que nous n’ayons pas accès à la variable sucre, le meilleur estimateur pour les y serait , la moyenne de l’échantillon. Clairement, notre estimateur du taux nutritionnel serait dégradé dans l’ensemble puisque la droite d’estimation aurait pour équation y=42,558. Sur le schéma ci-dessous, nous voyons que les observations du départ semblent plus groupées autour de la droite de régression que de la droite de y=, ce qui suggère que l’erreur globale prédite est plus faible quand nous utilisons l’information sucre. Cela indique qu’utiliser l’information de la variable explicative améliore nos estimateurs du taux nutritionnel.  
  Nous définissons SST (som of squares total) = SCT (somme des carrés totale) = . La SST mesure la variabilité totale dans les valeurs de la seule variable cible. Remarquons SST=(n-1)\*Var(y).
* Nous aimerions obtenir une mesure du degré dont l’équation de la régression estimée améliore les estimateurs . L’indicateur statistique SSR (som of square regression) est une mesure de l’amélioration globale de la fiabilité de prévision en utilisant la régression par opposition à ignorer l’information de la variable explicative SSR =. Nous avons SST=SSR+SSE. SSR mesure la portion de la variabilité dans la variable cible y qui est prise en compte par la régression linéaire. Ici, la variable sucre explique 8654,7 dans le total de la variabilité à expliquer.
* En résumé : variabilité totale sur Y = variabilité de Y expliquée par la régression + variabilité de Y non expliquée par la régression.
* Nous avons mesure la qualité de l’ajustement de la régression. Ici vaut 0,5771.

### Coefficient de corrélation

r = = - 0,76 est un indicateur de la force de la relation linéaire entre les deux variables qualitatives. Ses valeurs peuvent être comprises entre -1 et 1.

Interprétation :

* les valeurs de r proches de 1 indiquent que les variables sont positivement corrélées càd quand la valeur x s’accroit, les valeurs y ont tendance à s’accroitre également (r > 0,7 variables corrélées ; si r est compris entre 0,33 et 0,7 légère corrélation)
* les valeurs de r proches de -1 indiquent que les variables sont négativement corrélées càd quand la valeur x s’accroit, les valeurs y ont tendance à diminuer (idem -0,7 et -0,33)
* les autres valeurs de r indiquent que les variables ne sont pas corrélées càd quand x s’accroit, les valeurs de y ne sont pas affectées.

### Ecart-type de l’estimateur s.

L’indicateur s est important à prendre en compte, c’est l’écart-type de la régression. , moyenne de la somme des erreurs au carré mesure la variabilité de la variable cible laissée non expliquée par la régression.   
NB : m est utilisé lors de la régression multiple, il représente le nombre de variables explicatives.

s = = 9,196 mesure la fiabilité des estimateurs produits par la régression. C’est l’estimateur du résidu typique, la différence typique entre la valeur prédite (estimée) et la valeur réelle. s représente la précision des prévisions générées par l’équation de la régression estimée. Plus la valeur de s est faible mieux c’est. Que veut dire une valeur faible de s ?

### Table ANOVA

En résumé, la table ANOVA présente :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source des variations | Somme des carrés | ddl | Moyenne des carrés | F |
| Régression | SSR | m | MSR=SSR/m | MSR/MSE |
| Erreur (résidus) | SSE | n-m-1 | MSE=SSE/(n-m-1) |  |
| Total | SST | n-1 |  |  |

### Exercice :

Calculez tous les paramètres de la régression tentant d’expliquer le taux nutritionnel par la quantité de fibres présentes dans la céréale – à partir d’excel et à partir de R. Interprétez tous ces résultats et comparez cette régression à la précédente.

### Test F

Il existe plusieurs démarches pour tester la validité de la linéarité d’une régression simple. Ces différents tests sont équivalents et reviennent à faire le test du coefficient de corrélation linéaire contre

Le test F dont les résultats sont donnés par l’ANOVA consiste à voir si la variance expliquée par la régression est significativement plus élevée que la variance non expliquée.

De manière plus précise, le test F réalisé ici est le suivant :

Test : (càd la relation linéaire est non significative) (càd relation linéaire significative)

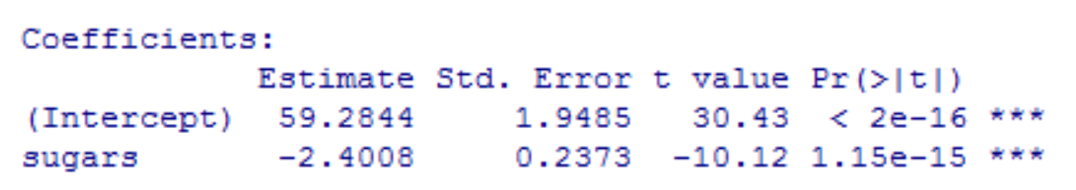
Fobs = la p-valeur est plus petite que

En supposant H0 vraie, nous avions maximum 10^-15 chances pour 10000 d’obtenir un tel échantillon. Ainsi, nous considérons avoir suffisamment de preuve pour rejeter H0 et affirmer que le modèle de régression linéaire est significatif.

### Test t

Pour rappel, le test t a pour but de déterminer si la valeur d’espérance d’une population de distribution normale et d’écart-type inconnu est égale à une valeur déterminée . Pour ce faire, nous tirons de cette population un échantillon de taille n dont on calcule la moyenne et d’écart-type s.

Ici, l’hypothèse nulle teste contre en utilisant la loi de student à (n-2) degrés de liberté



Dans l’exemple ci-dessus :

P-valeur = .

Le procédé est le même pour tester