Chapitre 1

Régression linéaire simple

La régression linéaire est une méthode de modélisation permettant d'établir une relation linéaire entre une variable continue dite "variable expliquée" ou dépendante et un ensemble d'autres variables continues dites "variables explicatives" ou indépendantes. Plus spécifiquement elle propose un modèle explicatif qui permet de prédire la variable dépendante en fonction des variables indépendantes.

Ce module est consacrée à l'étude de la régression linéaire simple pour modéliser la relation prédictive entre la variable dépendante et **une seule** variable indépendante. Cette modélisation permet d'élaborer les concepts de base de la régression à plusieurs variables.

La régression peut servir à remplacer une variable difficile à observer par une autre variable qui elle est relativement simple à mesurer. On peut penser au modèle qui prédit le rendement d'une entreprise en fonction du taux de change pour le $US ou celui qui donne le nombre d'hospitalisations dans une grande ville en fonction de la quantité de smog. L'objectif est de prédire la valeur du rendement ou du nombre d'hospitalisations si on connaît le taux de change ou la concentration de smog.

Elle peut aussi servir à comprendre les liens existants entre les variables pour établir les principales causes d'un phénomène. C'est le lien entre les variables et la force de ce lien qui sont d'intérêt. On peut penser à la relation entre la criminalité et le taux de chômage dans les villes nord américaines ou la relation entre l'âge des travailleurs et la productivité. Dans ces deux cas on ne veut pas prédire mais simplement vérifier l'existance d'un lien.

On donne dans ces notes les différentes formules pour effectuer le calcul des coefficients du modèle et pour faire des tests d'hypothèses. Ces calculs ne sont là que pour montrer comment on en arrive à dériver le modèle. Pour des cas concrets on utilisera Excel qui permet d'effectuer tous ces calculs sans trop de mal.

Objectifs et compétences

L'objectif de cette partie est de donner à l'étudiant les outils nécessaires pour modéliser un problème de régression linéaire simple, calculer les différents paramètres et inter-

préter les résultats.

L'étudiant sera en mesure de

- Modéliser sous forme de régression linéaire simple le lien entre deux variables
- Identifier et calculer les estimateurs des principaux paramètres statistiques
- Interpréter les paramètres et la mesure d'adéquation du modèle

Modèlisation déterministe

Considérons deux mesures continues, (x, y) sur une unité statistique. Pour un ensemble de n unités statistiques on a :

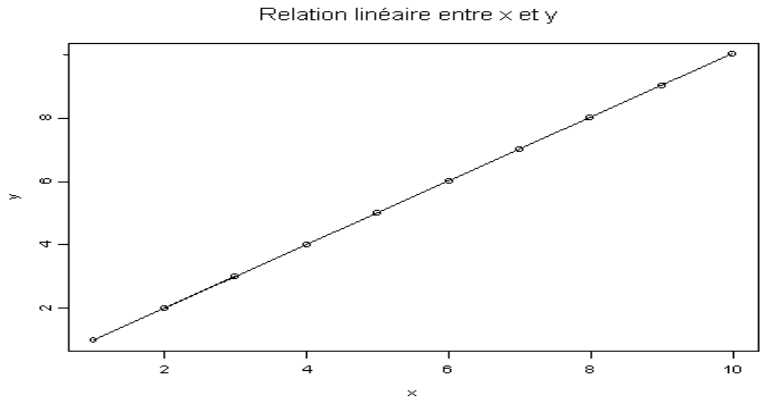
$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$$

On veut construire une relation linéaire entre les mesures x_i et y_i. Le modèle linéaire déterministe régissant ces deux variables est donné par l'équation suivante :

$$y = \beta_0 + \beta_1 x$$

où les coefficients[1] β_0 et β_1 sont respectivement l'ordonnée à l'origine et la pente de la droite et c'est pour cette raison que l'on parle de modèle "linéaire".

Le graphique suivant illustre une relation linéaire parfaite :



La relation ainsi représentée est parfaite dans le sens que tous les points (x_i, y_i) sont sur la droite. De plus, ce modèle déterministe implique une relation inversible permettant

[1] Les coefficients sont souvent représentés par la lettre grecque béta noté β.

de déduire x si on connaît y:

$$x = \frac{1}{\beta_1} y - \frac{\beta_0}{\beta_1}$$

C'est un modèle idéal pour lequel la connaissance d'une des deux variables donne toute l'information nécessaire pour la deuxième. Il n'est malheureusement pas réaliste en pratique.

Un modèle plus réaliste et adapté à l'administration est de considérer

- Une variable d'intérêt dont on veut connaître la valeur mais qui est difficile à observer : y.
- Une variable dont la valeur peut être connue et qui permet des observations directes : x.
- Un écart entre la valeur idéale donnée par le modèle ci-haut et la réalité : e.

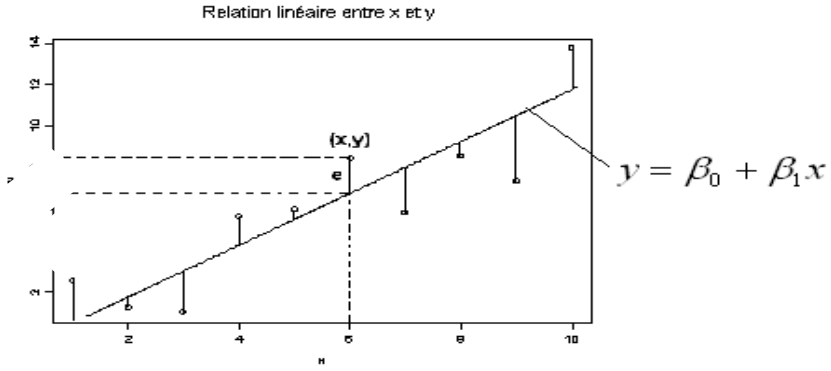
En considérant les n couples de valeurs fixées

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \cdots, (x_n, y_n)$$

obtient un modèle plus réaliste par la possibilité que la relation entre les deux variables ne soit pas exacte :

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Voici une représentation graphique de ce modèle :



Pour chaque valeur x_i observée il y a une valeur y_i qui est plus ou moins loin de la relation parfaite et la différence, e_i, est la distance entre la valeur de la droite $\beta_0 + \beta_1 x_i$ et la valeur de y_i c'est-à-dire la distance pour une valeur x_i fixée entre l'idéal pour y et la valeur observée. Le fait de considérer un écart dans le modèle en fonction de la variable y est un choix arbitraire mais qui permet de simplifier les calculs. La question n'est pas d'obtenir "la relation" entre x et y mais d'obtenir la "meilleure" droite permettant de lier les deux variables observées.

En considérant le nuage de points $((x_i, y_i))$ et la notion de "meilleure droite" il y a deux

questions auxquelles il faut répondre

- Quelles sont les valeurs de β_0 et de β_1 ?
- Quelle mesure permet de dire si la modélisation est adéquate ?

Valeur des paramètres

On considère le nuage de points et la question est de déterminer les constantes du modèle, β_0 et β_1.

Méthode des moindres carrés

Dans le but de définir la notion de "meilleure droite" on se base sur la distance moyenne entre le modèle et chacun des points. La différence entre le modèle et l'observation pour le point (x_i, y_i) est donnée par e_i : la distance étant prise comme le carré de la différence. C'est un choix purement arbitraire dicté par la simplicité : le carré se travaille très bien et une distance qui ne dépend que de x est plus simple à modéliser qu'une distance tangentielle qui dépendrait des deux éléments en même temps (x et y).

La méthode des moindres carrés est parfaitement adaptée à la résolution du premier problème : en considérant la différence e_i on peut la transcrire en fonction de la droite théorique $\beta_0 + \beta_1 x_i$ et de l'observation réelle y_i

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

C'est le segment de droite qui lie le point et la droite théorique sur le graphique ci-haut. L'idée de la méthode est de trouver les valeurs des paramètres β_0 et β_1 qui minimisent le critère $\sum$, c'est-à-dire la somme des distances entre le modèle et les observations. L'équation permettant de résoudre en fonction de β_0 et β_1 est donnée par

$$\min_{\beta_0,\beta_1} \sum e_i^2 = \min_{\beta_0,\beta_1} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Par la technique de la dérivée[2] il suffit de dériver la fonction par rapport à β_0 puis à β_1 et d'égaler les deux résultats à 0.

La solution des équations notées $\widehat{\beta}_0$ et $\widehat{\beta}_1$ est donnée par

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} \\ \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned}$$

où $S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y})$.

[2] La technique de la dérivée consiste à dériver la fonction par rapport à chacun des paramètres d'intérêt puis d'égaler chacune de ces dérivées à 0. Cela forme un système avec autant d'équations que d'inconnues qu'il suffit de solutionner pour obtenir le maximun ou le minimun de la fonction.

En appliquant ce principe, cela veut dire que si un ensemble d'observations du type

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \cdots, (x_n, y_n)$$

est donné alors la droite

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

est celle qui minimise les écarts en terme de distance entre les observations et le modèle idéal toujours en considérant que la variable x est explicative et la variable y expliquée.

Le modèle ainsi obtenu peut servir à "deviner" ou prédire y si on connaît le point x : l'équation de régression est donnée par

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Si $\widehat{\beta}_0 = 3$ et $\widehat{\beta}_1 = 100$ alors pour $x = 255$ la valeur de y donné par le modèle est de

$$\begin{aligned} \widehat{y} &= 3 + 100 \times 255 \\ &= 25503 \end{aligned}$$

On utilise ici $\widehat{y}$ pour indiquer que c'est la valeur obtenue en fonction de la valeur de x et des estimations des paramètres.

Ce modèle donne une prévision de y pour une valeur de x donnée mais on obtient aussi "l'effet" d'un changement dans la valeur de x : si x augmente de 1 unité alors y augmente de 100 unités.

Exemple 1 : Considérons la relation entre le nombre d'employés d'une usine et le taux d'absentéisme. Une théorie veut que ce taux augmente si le nombre d'employés est plus grand puisque les responsabilités sont divisées. On veut donc prévoir le taux d'absentéisme étant donné la taille de l'entreprise en terme d'employés.

La relation est donnée par le modèle linéaire

$$y_i = \beta_0 + \beta_1 x_i$$

où y_i est le taux d'absentéisme à l'usine i et x_i est la taille de l'entreprise. L'idée est de modéliser ce taux en fonction de la taille de l'entreprise pour déterminer dans un premier temps si cela est relié et dans un deuxième temps quel est l'influence du premier sur le deuxième.

On a observé des valeurs suivantes dans 7 entreprises :

Nombre d'employés	356	67	25	157	589	557	78
Taux d'absentéisme %	5	3	2	4	7	3	8

La variable x est le nombre d'employés dans l'entreprise et y est le taux d'absentéisme en %.

On obtient $\overline{y} = (5, 3, 2, 4, 7, 3, 8), 4.\,571\,4$, $\overline{x} = 261.\,29$,

$$S_{xx} = \sum (x_i - \overline{x})^2 = 341861.4$$

et

$$S_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = 715.8571$$

ainsi

$$\begin{aligned} \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = 2.0940 \times 10^{-3} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} = 4.0243 \end{aligned}$$

L'équation de régression est

$$\widehat{y} = 4.024 + 0.002x$$

Selon ce modèle une entreprise ayant 200 employés devrait avoir un taux d'absentéisme en % de

$$4.024 + 0.002(200) = 4.424$$

De plus, une augmentation de 100 du nombre d'employés augmente de $0.002 * 100 = 0.2$ le taux (en %).

Remarque 1 Lorsque l'équation de régression est présentée il est possible de remplacer le "y" et le "x" par des noms qui font directement référence aux variables du problème. Dans l'exemple précédant on peut, et c'est habituellement mieux, présenter l'équation de régression sous la forme

$$Abs = 4.024 + 0.002Empl$$

Cette présentation permet de voir immédiatement la variable expliquée et la variable explicative. Il est recommandé de prendre des noms cours pour les variables quitte à donner une abréviation.

Exemple 1: Dans le but d'expliquer la consommation sur carte de crédit, des données sur le revenu et sur la dépense sont obtenues :

Dépenses	Revenu
8900	21000
9400	25000
14500	30000
25400	45000
26600	50000

Le modèle à estimer doit permettre d'obtenir les dépenses sur carte de crédit en fonction des revenus. La variable dépendante est $y =$"Dépenses" et la variable indépendante est $x =$"Revenu". Pour obtenir l'équation de régression il faut obtenir $\overline{x}$, $\overline{y}$, S_{xy} et S_{xx}. Or

$$\begin{aligned} \overline{x} &= 34\,200 \qquad \overline{y} = 16960 \\ S_{xx} &= 642\,800\,000 \qquad S_{xy} = 429740000 \end{aligned}$$

et ainsi

$$\begin{aligned}\widehat{\beta}_1 &= \frac{429740000}{642\,800\,000} = 0.668\,54\\ \widehat{\beta}_0 &= 16960 - 0.668\,54 * 34\,200 = -5904.\,1\end{aligned}$$

L'équation de régression devient

$$\widehat{y} = -5904.\,1 + 0.668\,54 * x$$

ce qui veut dire que pour un revenu de 20000 les dépenses estimées par ce modèle seront de

$$-5904.\,1 + 0.668\,54 * 20000 = 7466.\,7$$

Mesure d'adéquation

Les paramètres étant estimés, l'étape suivante consiste à définir une mesure "raisonnable" de l'adéquation du modèle en fonction des données. Pour établir cette mesure on considère la mesure y seule. Si on ne connaît pas x alors la variance de y, c'est-à-dire l'incertitude liée à cette variable est donnée par $s_y = \frac{1}{n-1}\sum (y_i - \overline{y})^2$ et notons $SST = (n-1)\, s_y^2$, soit la somme des carrés brute. On obient alors

$$SST = \sum (y_i - \overline{y})^2$$

Si on ajoute et enlève la valeur de la droite théorique, cette somme peut se décomposer en deux sommes de carrés[3]

$$\begin{aligned}\sum (y_i - \overline{y})^2 &= \sum (y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2\\ &= \sum (\widehat{y}_i - \overline{y})^2 + \sum (y_i - \widehat{y}_i)^2\end{aligned}$$

La deuxièr rtie de la formule est $\sum \widehat{e}_i^2$ c'est-à-dire la différence entre la valeur observée de y valeur prédite par le modèle estimé. C'est en fait l'erreur par rapport à ce qui est estimé donc ce qui reste à expliquer entre x et y. Notons

$$SS_{err} = \sum e_i^2 = \sum (y_i - \widehat{y}_i)^2$$

Si SST représente la variations des données y et que SS_{err} représente la variation non expliquée par x alors la différence

$$SS_{reg} = SST - SS_{err}$$

est la réduction de l'incertitude à propos de y si on connaît x.

Une mesure de la qualité de la modélisation ou de l'adéquation du modèle est donnée par

$$R^2 = \frac{SS_{reg}}{SST} = \frac{SST - SS_{err}}{SST}$$

[3] Cette relation peut se démontrer avec quelques manipulations algébriques.

c'est-à-dire la proportion de la variance de y qui a été expliquée en considérant x comme une variable explicative. On dit que R^2 est de coefficient de détermination du modèle par rapport aux données. Il est interprété, si multiplié par 100, comme le % d'explication de la variable x, sur y. Cette interprétation est basée uniquement sur la réduction de la variance des données y si on connaît x et elle est justifiée sur ce point.

Remarque 2 Si un modèle colle parfaitement aux données alors tous les points observés sont sur la droite estimée. Cela veut dire que $SS_{err} = 0$ puisqu'il n'y a aucun écart entre une observation et la droite. On a alors que $SS_{reg} = SST$ et ainsi $R^2 = 1$. Cela veut dire que lorsque R^2 est proche de 1 le modèle est bon.

Si par contre la valeur de R^2 est proche de 0 cela veut dire que le fait d'observer x ne réduit en rien l'incertitude sur la variable y et ainsi la modèlisation n'apporte aucune information supplémentaire.

Exemple 3 : En reprenant l'exemple des dépenses de carte de crédit, l'équation de régression est

$$\widehat{y} = -5904.1 + 0.668\,54 * x$$

et on obtient le tableau suivant :

Dépenses	Revenu	*Prévisions Dépenses*
8900	21000	8135,220909
9400	25000	10809,39639
14500	30000	14152,11574
25400	45000	24180,2738
26600	50000	27522,99315

où "Prévisions Dépenses" représentent les $\widehat{y}_i$. On obtient

$$s_y^2 = 73083000 \text{ et } \overline{y} = 16960$$

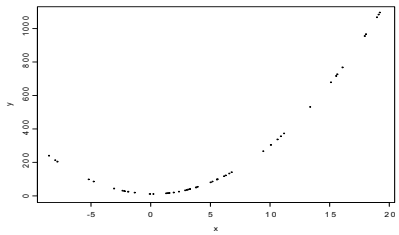
et ainsi $SST = (n-1)\, s_y^2 = 292\,332\,000$

$$SS_{reg} = \sum \left(\widehat{y}_i - \overline{y}\right)^2 = 287300042,9$$

Donc $R^2 = 287300042.9/292\,332\,000 = 0.982\,79$. On a une relation presque parfaite.

Remarque 3 Il se peut que la relation soit parfaite mais qu'elle ne soit pas linéaire. Le coefficient R^2 n'est plus un bon indicateur de l'adéquation comme dans l'exemple

suivant :



La relation est parfaite mais $R^2 = 0.68$. Il faut toujours vérifier qu'on a une relation linéaire ou presque linéaire avant d'interpréter le coefficient. Pour faire cette vérification il suffit de produire le graphique y en fonction de x.

Modèle aléatoire

La modélisation déterministe supposait un ensemble de données fixe et la droite résultante est le meilleur modèle en fonction des choix de la modélisation et des observations c'est-à-dire par rapport à des données fixes. Le modèle aléatoire suppose une erreur qui est certe réelle mais pas reproductible exactement, seulement en probabilité.

Dans le modèle aléatoire on considère l'erreur entre la valeur estimée par le modèle et la valeur observée comme étant aléatoire donc pas fixée par les observations, celles-ci sont simplement le résultat d'une réalisation particulière d'un processus aléatoire. Pour une observation associée à une valeur x_i l'équation de régression est donnée par

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

où e_i est une variable aléatoire de moyenne 0 et de variance σ^2 constante pour toutes les valeurs de x.

On remarque que la variable dépendante est en majuscule puisque c'est une v.a. aléatoire et que la variable indépendante est en minuscule parce qu'on suppose qu'elle est fixée au départ (on observe Y selon une certaine valeur de x).

Dans ce modèle on suppose que les erreurs ont la même loi de probabilité et qu'elles ne sont pas liées entre elles. Cela veut dire qu'une valeur forte pour l'erreur ne peut en aucun cas influer sur l'erreur à l'observation suivante.

La régression est alors une moyenne conditionnelle

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

c'est-à-dire la moyenne des valeurs observables pour la variable aléatoire Y étant donné

une certaine valeur x fixée. Selon la distribution des erreurs les valeurs observables réellement seront plus ou moins éloignées de cette moyenne pour un x donné.

Les estimateurs des moindres carrés pour β_0 et β_1 tels que décrits dans la section précédante sont les estimateurs de forme linéaire non biaisés les plus intéressants, c'est-à-dire de variance minimale et sans biais[4].

Propriété des estimateurs

La méthode des moindres carrés donne le même résultat que pour le modèle déterministe : une réécriture des estimateurs en fonction des données aléatoires donne

$$\begin{aligned} \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} \\ \widehat{\beta}_1 &= \frac{S_{xY}}{S_{xx}} \end{aligned}$$

où Y est une variable aléatoire.

Cela implique que les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont aussi des variables aléatoires donc dépendants des échantillons qui seront choisis. Comme variables aléatoires elles ont une moyenne, une variance et une loi de probabilité.

Proposition 1 Pour $\widehat{\beta}_0$ on obtient

$$E\left(\widehat{\beta}_0\right) = \beta_0$$

et

$$Var\left(\widehat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)$$

De plus, si on suppose que les erreurs sont de distribution normale alors

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}} \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)^{-1/2} \sim t_{n-2}$$

où $\widehat{\sigma}^2$, un estinateur de σ^2, est donné par

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum \widehat{e}_i^2 = \frac{1}{n-2} \sum \left(\widehat{Y}_i - Y_i\right)^2 \quad (1)$$

c'est-à-dire la variance des erreurs observées en considérant l'équation de régression estimée.

[4] Il peut sembler naturel que les deux denières conditions soient respectées dans tous les cas mais ce n'est pas toujours possibles. Il existe des modélisations pour lesquelles ces propriétés naturelles des estimateurs ne peuvent être respectées.

- **Remarque 4** Ce résultat permet de construire un intervalle de confiance de niveau $1-\alpha$ par la formule

$$\beta_0 \in \left(\beta_0 \pm t_{n-2;\alpha/2} S_{\widehat{\beta}_0}\right)$$

où $t_{n-2;\alpha/2}$ est le point critique d'une loi de Student à $n-2$ degrés de liberté et

$$S_{\widehat{\beta}_0} = \widehat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)}$$

Proposition 2 Pour $\widehat{\beta}_1$ on obtient

$$E\left(\widehat{\beta}_1\right) = \beta_1$$

et

$$Var\left(\widehat{\beta}_1\right) = \sigma^2\left(\frac{1}{\sum (x_i - \overline{x})^2}\right) = \frac{\sigma^2}{S_{xx}}$$

De plus, si on suppose que les erreurs sont de distribution normale alors

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}}\sqrt{S_{xx}} \sim t_{n-2}$$

où $\widehat{\sigma}^2$ est donné par la formule.

Remarque 5 Cela permet de construire un intervalle de confiance de niveau $1-\alpha$ pour le paramètre β_1:

$$\beta_1 \in \left(\beta_1 \pm t_{n-2;\alpha/2} S_{\widehat{\beta}_1}\right)$$

où $t_{n-2;\alpha/2}$ est le point critique d'une loi de Student à $n-2$ degrés de liberté et

$$S_{\widehat{\beta}_1} = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}$$

Exemple 4: On considère un modèle de régression pour lequel on a observé le poids des individus par rapport à la taille (grandeur) en m. On observe les valeurs suivantes :

Taille(cm)	175	165	187	152	145	189	170	165	160	157	145
Poids (kg)	60	81	97	57	61	97	109	104	59	74	61

L'équation de régression estimée est donnée par

$$P = -65.539 + 0.8613T$$

avec $R^2 = 0.4066$ et

$$s_{\widehat{\beta}_1} = \frac{\sqrt{269.49}}{\sqrt{2240.7}} = 0.3468$$