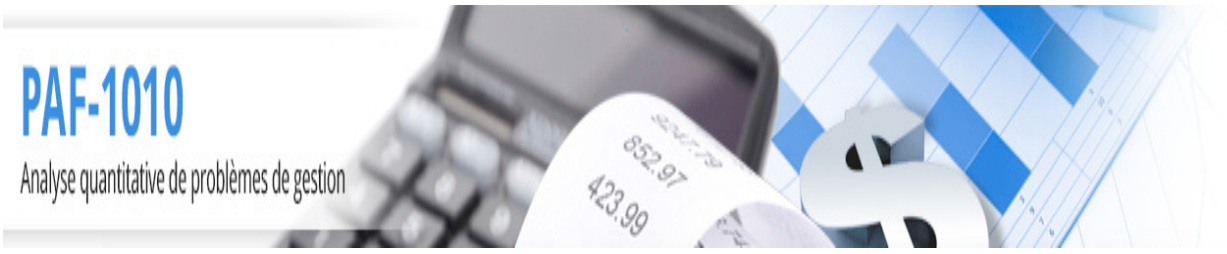
MODULE 13

RÉGRESSION



Louis Houde
Département de Mathématiques et d'informatique
Université du Québec à Trois-Rivières

13 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
- Faire exécuter une régression linéaire par le logiciel EXCEL
- Effectuer un test statistique sur les paramètres du modèle
- Vérifier les hypothèses de base de la modélisation

Modèlisation déterministe

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

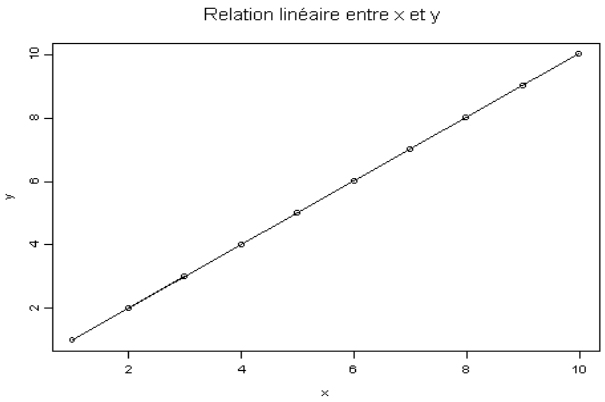
$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$$

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.

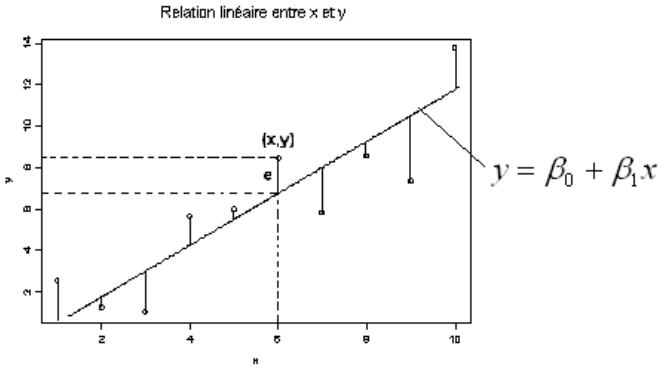
En considérant les n couples de valeurs fixées

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

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 $\sum e_i^2$ 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 13.1 ★★★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 13.1 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 13.2 ★★★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 obtient 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 deuxième partie de la formule est $\sum \widehat{e}_i^2$ c'est-à-dire la différence entre la valeur observée de y et la 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 13.2 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 13.3 ★★★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

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

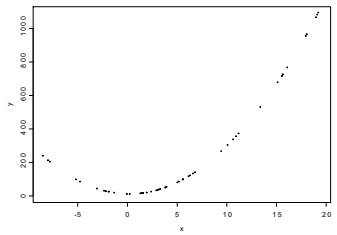
et ainsi $SST = (n-1)\, s_y^2 = 292\,332\,000$

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

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

Remarque 13.3 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 13.1 Pour $\widehat{\beta}_0$ on obtient

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

et

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

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

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

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

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum \widehat{e}_i^2 = \frac{1}{n-2} \sum \left(\widehat{Y}_i - Y_i\right)^2 \tag{13.1}$$

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 13.4** 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 13.2 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 13.5 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 13.4 ★★★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

$$s_{\widehat{\beta}_1} = \frac{\sqrt{269.49}}{\sqrt{2240.7}} = 0.346\,8$$

En supposant que l'erreur est une v.a. normale, un intervalle de confiance de niveau 95% pour β_1 est donné par

$$\begin{aligned} \beta_1 &\in (0.8613 \pm 0.784\,52) \\ &\in (0.076\,78, 1.645\,8) \end{aligned}$$

Tests d'hypothèses

La loi des estimateurs donnée ci-haut permet aussi de construire un test d'hypothèses pour confronter

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0 \end{aligned}$$

pour $i = 0$ ou 1. Le test est de rejeter H_0 si

$$\frac{\left|\widehat{\beta}_i\right|}{S_{\widehat{\beta}_i}} \geq t_{n-2;\alpha/2}$$

Le test sur β_1 permet de vérifier si la régression est "significative" : le modèle de régression s'écrit

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

et si on accepte l'hypothèse $H_0 : \beta_1 = 0$ alors cela veut dire que x n'est pas présent dans le modèle donc qu'il n'explique en rien Y.

Une autre façon de vérifier si la régression est significative est de regarder les sommes de carrés permettant de diviser la variation observée sur la variable expliquée : sous l'hypothèse nulle ($H_0 : \beta_1 = 0$) les sommes de carrés SS_{reg} et SS_{err} sont distribués comme des Khi-deux et ainsi le rapport des carrés moyens est distribué comme une loi de Fisher à 1 et $n-2$ degrés de liberté. Le test est alors de rejeter l'hypothèse nulle si

$$F = \frac{CM_{reg}}{CM_{err}} \geq F_{1,n-2;\alpha}$$

où $F_{1,n-2;\alpha}$ est le point critique de niveau α pour une loi de Fisher à 1 et $n-2$ degrés de liberté et

$$CM_{reg} = SS_{reg} \text{ et } CM_{err} = SS_{err}/(n-2)$$

Les sommes de carrés et le test sont habituellement résumés dans un tableau ANOVA qui contient les sommes de carrés, les degrés de liberté pour chaque composante, les carrés moyens et la statistique F. Les logiciels statistiques ajoute une colonne donnant le seuil de signification expérimental (p-value) permettant d'effectuer le test sans les tables de

la loi de Fisher. Le tableau se présente comme suit

Variation	**SS**	**d.l.**	**CM**	**F**	**SIG**
Régression	SS_{reg}	1	$\frac{SS_{reg}}{1}$	$\frac{CM_{reg}}{CM_{err}}$	$\Pr\left(F_{1,n-2} \geq \frac{CM_{reg}}{CM_{err}}\right)$
Erreur	SS_{err}	$n-2$	$\frac{SS_{err}}{n-2}$		
Totale	SST	$n-1$			

Le test devient

$$\text{Rejeter } H_0 \text{ si } \widehat{\alpha} \leq \alpha$$

où $\widehat{\alpha}$ est justement le seuil de signification empirique.

Dans le cas d'une régression simple il faut nécessairement que le degré de liberté associés à la régression soit de 1, que celui lié à l'erreur soit $n-2$ et que celui lié au total soit n-1. Cela permet de s'assurer que le modèle de régression a bien une variable et que toutes les observations ont bien été prise en compte.

Les logiciels d'analyse des données deviennent de plus en plus complexes et puisque l'utilisateur n'a aucun contrôle réel sur les éléments pris en compte par l'ordinateur lors du calcul, il est nécessaire de comprendre le plus d'éléments possibles de la sortie informatique pour détecter d'éventuelles incompréhensions entre la machine et l'utilisateur.... Les degrés de liberté sont une façons de vérifier qu'il y a bien une variable dans le modèle de régression et que toutes les données disponibles sont prises en compte.

Remarque 13.6 En supposant la normalité des observations il est possible d'effectuer techniquement 3 tests d'hypothèses, un stipulant que β_0 est nul, un autre stipulant que β_1 est nul et finalement un dernier stipulant que β_1 est nul. Il est clair que le deuxième et le troisième test font la même chose mais avec des formulation différentes.

Remarque 13.7 Le test pour vérifier si la régression est significative peut être effectué sur le paramètre

$$\rho = \frac{\sigma_{xY}}{\sigma_x \sigma_Y}$$

où $\sigma^2_{xY} = E\left((x-\overline{x})\left(Y-\overline{Y}\right)\right)$, $\sigma_x = \sqrt{s_{xx}}$ et $\sigma^2_Y = E\left(Y-\overline{Y}\right)^2$. Ce paramètre est similaire à un coefficient de corrélation entre deux variables aléatoires mais dans le contexte de la régression il faut se rappeler que la variable x est fixée.

La valeur R^2 est une estimation du paramètre ρ^2.

Régression avec EXCEL

Estimation des paramètres

Les logiciels statistiques permettent de faire les calculs de la régression simple avec une grande efficacité et de plus, ils sont disponibles et d'une utilisation aisée. Les logiciels donnent généralement les éléments suivants :

- le coefficient R^2
- le tableau de l'analyse de la variance
- un tableau des coefficients β_i avec les tests d'hypothèses associés.

Tous les logiciels donnent au moins ces résultats avec un peu plus dans certains cas et beaucoup plus dans d'autres.

Une sortie classique donnée par EXCEL, outil "utilitaire d'analyse", est de la forme

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,06300437
Coefficient de détermination R^2	0,00396955
Coefficient de détermination R^2	-0,1067005
Erreur-type	12,9427281
Observations	11

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	6,00845666	6,00845666	0,03586834	0,853989935
Résidus	9	1507,627907	167,5142119		
Total	10	1513,636364			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	69,6046512	7,50238961	9,277664155	6,6545E-06
Subalternes	0,1627907	0,859555547	0,189389386	0,85398994

On y retrouve le coefficient de détermination R^2 dans la deuxième ligne de la table "Statistiques de la régression", le tableau ANOVA, "Analyse de Variance" permet de vérifier si la régression est significative avec le seuil de signification expérimental (Valeur critique de F) et finalement le troisième tableau donne les coefficients β_0(constante) et β_1 avec le seuil de signification expérimental pour chaque test.

Cette sortie n'est pas un exemple de limpidité et bien qu'elle soit destinée à un utilisateur débutant, il y a certains cellules qui ne sont compréhensibles que par un utilisateur averti : Il y a deux coefficients de détermination R^2 et un coefficient dont le nom est semblable mais il a généralement la désignation de "multiple" dans la majorité des logiciels statistiques.

Dans le cas présent le coefficeint R^2 qui nous intéresse se trouve à être le premier "coefficient de détermination multiple R ^2". La première ligne est simplement le R, soit l'estimation de ρ, tandis que la troisième ligne est un coefficient R^2 ajusté. Ce dernier n'est utile que lorsqu'il y a plusieurs variables explicatives (x) dans le modèle.

La deuxième partie donne le tableau d'analyse de la variance tel que défini ci-haut et

finalement la dernière partie donne les coefficients, l'écart type des coefficients (donné ci-haut), la valeur de la statistique utilisée pour faire le test, le $\widehat{\alpha}$ correspondant à ce test (rejeter β_i si $\widehat{\alpha} \leq \alpha$).

Le dernier tableau donne aussi les bornes des intervalles de confiance de niveau 95% pour chacun des coefficients mais il a été omis dans cette sortie d'Excel pour ne pas trop surcharger la page.

Tests d'hypothèses

Il y a 3 tests possibles pour une régression linéaire simple (une seule variable explicative) ;

- un test pour le coefficient β_1 par la table de l'analyse de la variance,
- un pour le coefficient β_1 basé sur la loi de Student
- un pour le coefficient β_0.

Les deux premiers tests sont équivalents s'il n'y a qu'une variable dans l'équation (les seuils de signification empirique doivent être égaux) tandis que le dernier sert à vérifier si une relation passe par le point (0,0).

Remarque 13.8 Il peut sembler étonnant de proposer deux tests identiques pour vérifier si la régression est "significative" mais cela se comprend mieux si on considère que la régression linéaire simple est un cas particulier de la régression linéaire où il y a plusieurs variables x pour expliquer la variable y. Dans un tel contexte les tests ne sont pas équivalents.

Test ANOVA

Les hypothèses statistiques sont

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

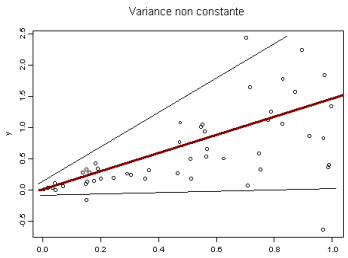
c'est-à-dire H_0 indique qu'il n'y a pas de lien entre les deux variables tandis que H_1 indique qu'il y a un lien. Le coefficient R^2 ne devrait être interprété que si on rejette H_0 puisque sinon cela revient à quantifier un lien qui n'existe pas.

Le seuil de signification empirique pour faire le test est donné dans la table de l'analyse de la variance et il est valide aux conditions suivantes :

- Les résidus sont normaux (à tout le moins une distribution symétrique sans valeurs aberrantes)

- La variance des résidus est constante
- Les valeurs de x sont fixées a priori

Pour vérifier ces conditions il faut faire l'analyse des résidus avec un diagramme en rectangle (Box-plot) et un graphique des points. Une variance non constante est illustrée par l'exemple suivant :



On remarque que pour des petites valeurs de x la variance autour de la droite de régression (ligne rouge) est assez faible mais plus x devient grand, plus la variance des écarts devient grande.

Le diagramme en rectangle permet de bien cibler les valeurs aberrantes et extrêmes qui pourraient invalider la première hypothèse.

Exemple 13.5 ★Si on reprend l'exemple des dépenses sur les cartes de crédit avec un niveau de 5% pour tester les hypothèses, on observe le tableau d'analyse de la variance suivant sur EXCEL

ANALYSE DE VARIANCE

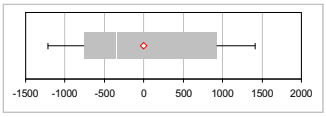
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	287300042,9	287300042,9	171,285271	0,000963468
Résidus	3	5031957,063	1677319,021		
Total	4	292332000			

Le seuil de signification empirique pour confronter les hypothèses

$$H_0 \quad : \quad \beta_1 = 0$$
$$H_1 \quad : \quad \beta_1 \neq 0$$

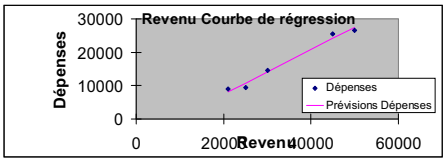
est $\widehat{\alpha} = 0.00096$. On doit donc rejeter H_0 au niveau 5% et conclure que la régression est significative.

Les conditions d'application sont respectées pour le test puisque les résidus sont relativement symétriques sans valeurs aberrantes selon le diagramme en rectangle.



Il est assez difficile de dire si la variance est constante selon le graphique des points

étant donné le très petit nombre de ceux-ci :



Il n'y a certainement aucune indication d'une variance non constante dans les résidus.

Test sur β_1 par Student

Le test sur β_1 permet de vérifier si la variable associé (x) est présente dans l'équation de régression. Les hypothèses statistiques sont

$$H_0 \quad : \quad \beta_1 = 0$$
$$H_1 \quad : \quad \beta_1 \neq 0$$

c'est-à-dire H_0 : la pente de la droite est 0 donc aucune influence de la variable x contre H_1 : la pente est différente de 0 donc qu'il y a une influence de la variable x sur la prédiction de la variable y.

La sortie informatique donne un tableau avec la valeur du coefficient β_0 sous le vocable "constante" et la valeur du coefficient β_1 sous le nom de la variable indépendante. En plus de la valeur des paramètres, la sortie donne des informations complémentaires dont, à tout le moins, l'erreur type (l'écart type du coefficient), la statistique t et le seuil de signification empirique $\widehat{\alpha}$.

On remarquera que dans le cas d'une régression simple, le seuil de signification empirique pour les hypothèses sur le coefficient β_1 est toujours le même que le seuil de signification empirique par la table ANOVA. De plus, on remarquera que la statistique pour tester les hypothèses sur β_1 par le test de Student est exactement la racine de la statistique F de la table de l'analyse de la variance.

Test sur β_0

Le coefficient β_0 représente l'ordonné à l'origine de la droite de régression, c'est-à-dire la valeur de y lorsque la valeur de x est 0. Les hypothèses statistiques sont

$$H_0 \quad : \quad \beta_0 = 0$$
$$H_1 \quad : \quad \beta_0 \neq 0$$

En regardant l'équation de régression

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

si $\beta_0 = 0$ alors il y a une relation qui dépend uniquement de β_1 et qui passe par l'origine.

Considérons l'exemple des ventes en fonction du nombre d'employés (directeur compris) dans les PME. Il est évident que si une PME n'a pas d'employés alors il ne peut y avoir de vente. On veut donc que la valeur de β_0 soit 0 et il peut être intéressant de le tester.

Dans d'autres cas ce paramètre n'est pas explicable: il est possible de lier le revenu par habitant d'un pays au % de la main d'oeuvre dans le secteur de l'agriculture. Or de vérifier si le fait de ne pas avoir d'agriculture dans un pays (0 pour le %AGR) implique 0 de PCINC[5] n'est pas informatif.

Dans le cas de l'analyse des ventes par employés on obtient la sortie EXCEL suivante :

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,793475763
Coefficient de détermination R^2	0,629603787
Coefficient de détermination R^2	0,607815774
Erreur-type	11,46798749
Observations	19

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	3800,354731	3800,354731	28,8967975	5,03317E-05
Résidus	17	2235,750532	131,5147372		
Total	18	6036,105263			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	8,247135626	3,582200703	2,302253925	0,03423264
Employé	0,465861957	0,086662735	5,375574157	5,0332E-05

Selon la sortie et avec un niveau de 5% on peut dire que le modèle de régression est significatif ($\widehat{\alpha} = 0.00005$) et que le coefficient β_0 est significatif à 5% puisque $\widehat{\alpha} = 0.034$. Cela veut dire que le modèle ne peut rendre compte des valeurs près ce 0.

Remarque 13.9 Il existe des modèles qui sont linéaire sur une portion des valeurs mais pas sur la totalité. Cela explique les modèles, comme le précédant, qui n'est pas conséquent pour la valeur $x = 0$.

Exemple 13.6 ★Dans une grande entreprise on cherche à savoir si un test d'aptitude à la gestion permet de bien prédire la performance d'un cadre dans un poste de direction. Chaque cadre a été soumis à un test d'aptitude avant sa promotion et après 2 ans un questionnaire a été rempli par les employés sous sa direction. On a observé les

[5] Per Capitae INCom

résultats suivants :

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,409915661
Coefficient de détermination R^2	0,168030849
Coefficient de détermination R^2	0,126432391
Erreur-type	7,487926834
Observations	22

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	226,4826711	226,4826711	4,03935287	0,058135128
Résidus	20	1121,380965	56,06904826		
Total	21	1347,863636			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	5,196544069	8,134417587	0,63883419	0,53017958
Score	0,147563193	0,073421313	2,009814138	0,05813513

Au niveau 5% on rejette l'hypothèse d'un lien entre ces deux variables donc il n'y a pas de lien entre les deux variables et l'analyse de l'équation de régression est non opportune.

Exemple 13.7 ★Un cadre veut évaluer le lien entre le nombre de subalternes et le salaire. Il observe

Salaire K$	56	76	85	56	55	75	88	82	69	60	77
Nb subalternes	2	5	5	10	6	7	2	19	11	8	7

Que peut-on dire sur la relation au niveau 5% ?

Une analyse par EXCEL donne la sortie suivante

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,06300437
Coefficient de détermination R^2	0,00396955
Coefficient de détermination R^2	-0,1067005
Erreur-type	12,9427281
Observations	11

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	6,00845666	6,00845666	0,03586834	0,853989935
Résidus	9	1507,627907	167,5142119		
Total	10	1513,636364			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	69,6046512	7,50238961	9,277664155	6,6545E-06
Subalternes	0,1627907	0,859555547	0,189389386	0,85398994

Au niveau 5% on peut dire qu'il n'y a pas de relation linéaire entre le nombre de subalternes et le salaire.

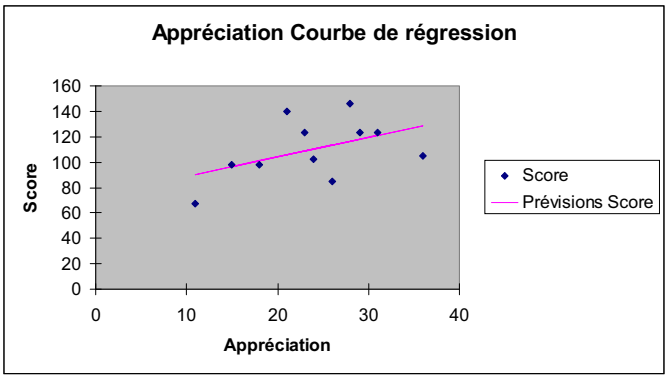
Or un test d'hypothèse sur β_1 au niveau 5% rejette l'hypothèse H_0 et on doit conclure que la relation linéaire n'existe pas.

Une analyse des résidus par EXCEL donne les valeurs suivantes :

ANALYSE DES RÉSIDUS

Observation	*Prévisions Score*	*Résidus*
1	108,7335605	14,26643952
2	128,8558773	-23,85587734
3	105,6378194	34,36218058
4	118,0207836	4,979216354
5	100,9942078	-2,994207836
6	90,15911414	-23,15911414
7	113,3771721	-28,37717206
8	96,35059625	1,649403748
9	110,281431	-8,281431005
10	116,4729131	29,52708688
11	121,1165247	1,883475298

avec les graphiques



Les tests sont donc valides.

Exemples

★Valeur des maisons dans la ville d'Albuquerque

On veut lier le prix de vente des maisons en centaine de $ en fonction des taxes annuelles en $. Le but est de faire des prévisions sur le prix des maisons si on connaît la valeur des taxes. L'équation du modèle est

$$Y = \beta_0 + \beta_1 x$$

où Y est le prix de vente de la maison et x est la valeur des taxes.

L'analyse des données donne les sorties EXCEL[6] suivantes

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multipl	0,875664739
Coefficient de détermination R^2	0,766788735
Coefficient de détermination R^2	0,764567675
Erreur-type	186,3175401
Observations	107

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	11984584,5	11984584,5	345,2355407	5,67019E-35
Résidus	105	3644993,705	34714,22576		
Total	106	15629578,21			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	211,5948434	49,95483129	4,23572331	4,90159E-05
TAX	1,091072774	0,058721342	18,58051508	5,67019E-35

On observe un modèle qui est significatif (on rejette $H_0 : \beta_1 = 0$ au niveau 5%) puisque le seuil de signification empirique est de 5.67E-35. Le coefficient de détermination R^2 est de 77% c'est-à-dire que la relation est forte.

Le modèle est donné par l'équation

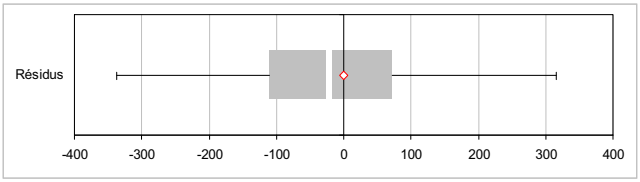
$$PRICE = 211.595 + 1.09\, TAX$$

Ce modèle permet de prédire le prix d'une maison lorsque les taxes sont connues. Prenons l'exemple d'une maison dont les taxes annuelles sont de 1000$. Le modèle donne un prix de vente de

$$PRICE = 211.595 + 1.09\ (1000) = 1301.6$$

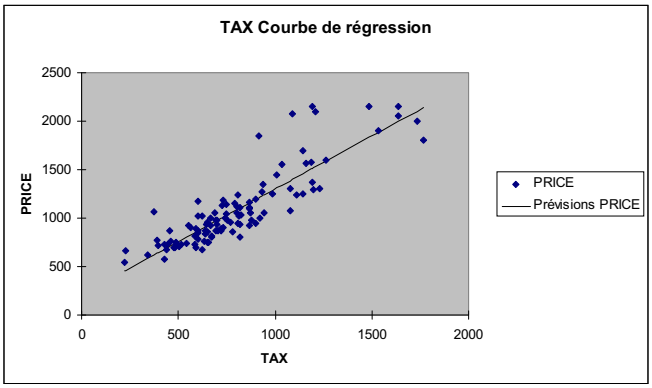
c'est-à-dire de 130160$ (l'unité de mesure étant en 100$).

Les conditions d'application sont respectées puisque le diagramme en boîte est relativement symétrique et qu'il n'y a pas de valeurs aberrantes :



[6] Les données sont disponibles sur le portail de cours sous "Albuquerque"

Il semble cependant que la variance augmente légèrement lorsque les taxes augmentent



★Parcomètres de la ville de New-York

Les grandes villes utilisent souvent des sous-contractants pour la collecte de l'argent des parcomètres. Dans le but de valider les montants qui sont recoltés $(Contr)$, la ville garde quelques parcomètres qui sont vidés par ses propres employés $(Ville)$. Cela permet d'estimer le montant que le sous-contractant devra verser à la ville[7].

L'analyse par EXCEL donne

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,651703623
Coefficient de détermination R^2	0,424717613
Coefficient de détermination R^2	0,411643013
Erreur-type	182890,213
Observations	46

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	1,08656E+12	1,08656E+12	32,4841771	9,31291E-07
Résidus	44	1,47175E+12	33448830008		
Total	45	2,55831E+12			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	325136,3211	226200,5298	1,4373809	0,157683735
Ville	186,9012517	32,79263198	5,699489196	9,31291E-07

On constate que la régression est significative au niveau 5% puisque le seuil de signification empirique pour confronter les hypothèses $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$ est de 9.31E-7. Le modèle est donné par

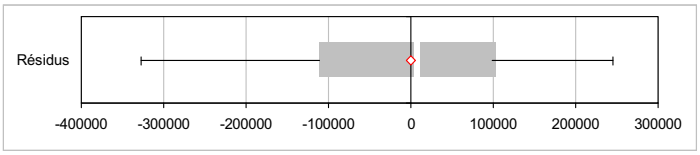
$$Contr = 325136.3211 + 186.90125\, VILLE$$

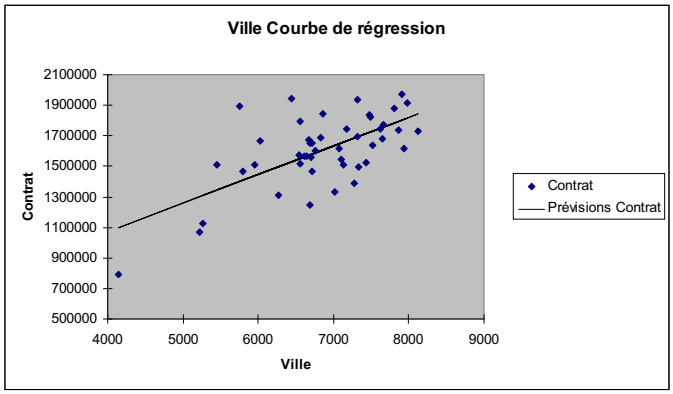
et le R^2 est de seulement 42% ce qui est relativement faible pour effectuer une prévision.

On constate que le test sur la constante ne permet pas de rejeter $H_0 : \beta_0 = 0$ au niveau 5%. Cela s'interprète comme étant un signe que le modèle n'est pas seulement local.

[7] Les données sont disponible sur le portail de cours sous "parko"

Les conditions d'application des tests sont respectées selon le diagramme en boîte et le graphique des points :





Publicité et vente de billets

Un publicitaire veut vérifier si les campagnes de publicité pour les films sont efficaces. Il note les dépenses de publicité en millier $ et les recettes pour le film en millier d'entrées :

VENTES	PUB
164	34
198	36
85	32
179	29
168	45
201	67
98	76
197	75
197	75
209	78
100	72
216	75
223	78
245	81
119	84
260	83
298	89
309	82
124	81
267	83

La sortie EXCEL pour le modèle de régression $VENTES = \beta_0 + \beta_1 PUB$ est :

RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0,400542526
Coefficient de détermination R^2	0,160434315
Coefficient de détermination R^2	0,113791777
Erreur-type	61,1586794
Observations	20

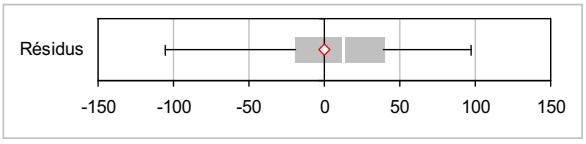
ANALYSE DE VARIANCE

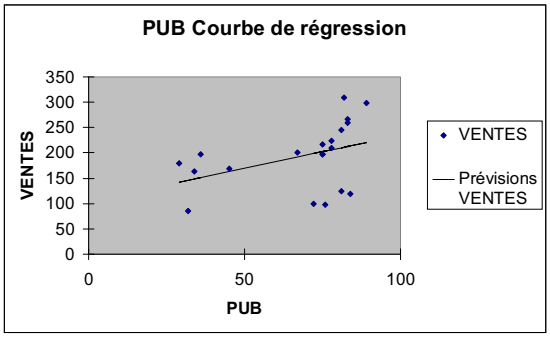
	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*
Régression	1	12865,63682	12865,63682	3,439656622	0,080105841
Résidus	18	67326,91318	3740,384066		
Total	19	80192,55			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*
Constante	104,8959227	49,35643393	2,125273533	0,047665411
PUB	1,298215163	0,699985643	1,854631128	0,080105841

Au niveau 5% on doit accepter $H_0 : \beta_1 = 0$ donc que la régression n'est pas significative. Cela veut dire qu'on ne peut pas lier les dépenses de publicité aux entrées pour les films.

Les graphiques suivants montrent que les tests sont bons.





Résumé

Le modèle de régression linéaire simple permet de mettre en relation une variable expliquée, y et une variable explicative, x via l'équation $y = \beta_0 + \beta_1 x$.

Pour déterminer les coefficients on utilise Excel : le tableau des coefficients donnent β_0 sous "constante" et β_1 sous le nom de la variable explicative.

Pour vérifier si la régression est significative il faut faire un test avec le niveau choisit sur le coefficient ρ ou sur β_1 ce qui donne exactement le même résultat.

Pour qualifier si la régression est un bon prédicteur on se base sur le coefficient de détermination R^2 qui donne une valeur entre 0 et 1. On interprète ce dernier comme étant le % d'explication.