Université Paris Ouest Nanterre La Défense
UFR SPSE-Master 1
PMP STA 21 Méthodes statistiques pour l'analyse des données en psychologie

Chapitre 4 : Régression linéaire

I Introduction

Le but de la régression simple (resp. multiple) est d'expliquer une variable Y à l'aide d'une variable X (resp. plusieurs variables $X_1, ..., X_q$). La variable Y est appelée variable *dépendante*, ou variable *à expliquer* et les variables X_j (j=1,...,q) sont appelées variables *indépendantes*, ou variables *explicatives*.

Remarque : La régression diffère de l'analyse de la corrélation où toutes les variables jouent un rôle symétrique (pas de variable *dépendante* versus *indépendante*). Toutefois, tout comme dans le contexte de l'analyse de la corrélation, il faut être prudent lorsqu'on formule des relations de causalité ! L'existence d'une relation entre X et Y n'implique pas nécessairement une relation de causalité entre elles.

II Représentation graphique

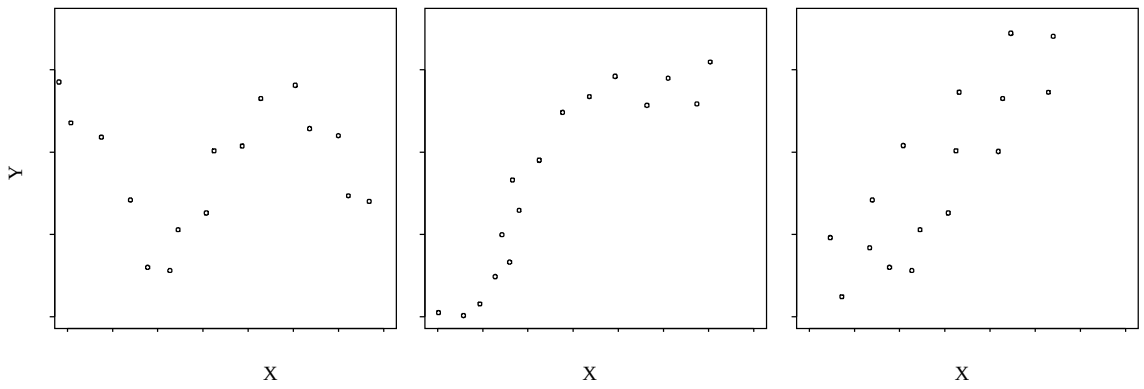
Avant toute analyse, il est intéressant de représenter les données. Le but de la régression simple est de chercher une fonction f telle que

$$y_i \approx f(x_i).$$

Pour définir $\approx$, il faut se donner un critère quantifiant la qualité de l'ajustement de la fonction f aux données.

Ainsi une étude de régression simple débute toujours par un tracé des observations $(x_i, y_i), i = 1, ..., n$. Cette première représentation permet de savoir si le modèle linéaire est pertinent. Le graphique suivant représente trois nuages de points différents.

Graphique 1 :



Au vue du graphique, il semble inadéquat de proposer une régression linéaire pour les 2 premiers graphiques, le tracé présentant une forme sinusoïdale ou sigmoïdale. Par contre, la modélisation par une droite de la relation entre X et Y pour le dernier graphique semble correspondre à une bonne approxiamation de la liaison.
Dans la suite de ce chapitre, nous étudierons le cas $f(x) = b_o + b_1\ x$.

III Modélisation du problème

Dans ce chapitre, nous allons analyser la régression linéaire simple sur un exemple. Cette présentation va nous permettre d'exposer la régression linéaire dans un cas simple afin de bien comprendre les enjeux de cette méthode, les problèmes posés et les réponses apportées.

1) Le problème

Exemple utilisé :
Etude de la relation entre la tension artérielle et l'âge d'un individu. Les données sont extraites de Bouyer et al. (1995) *Epidémiologie. Principes et méthodes quantitatives, Les éditions INSERM.*

1. *Objectif*

   On souhaite savoir si, de façon générale, l'âge a une influence sur la tension artérielle et sous quelle forme cette influence peut être exprimée.

   Le but est d'expliquer au mieux comment la tension artérielle varie en fonction de l'âge et éventuellement de prédire la tension à partir de l'âge.

2. *Population et variables étudiées*
   - Population générale d'individus.
     Sur cette population, on définit deux variables.
   - La variable Y : variable `tension` ; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante (VD).
   - La variable X : variable `âge` ; c'est la variable explicative, appelée également régresseur, variable indépendante (VI).

3. *Echantillon aléatoire d'individus*

   Pour l'étude, on doit faire des mesures sur n individus tirés au sort dans la population.

   Données numériques : on observe deux échantillons appariés de X et Y de taille n :

   $(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)$

   où x_i et y_i sont les valeurs de X et Y observées sur le i^e individu tiré au sort.

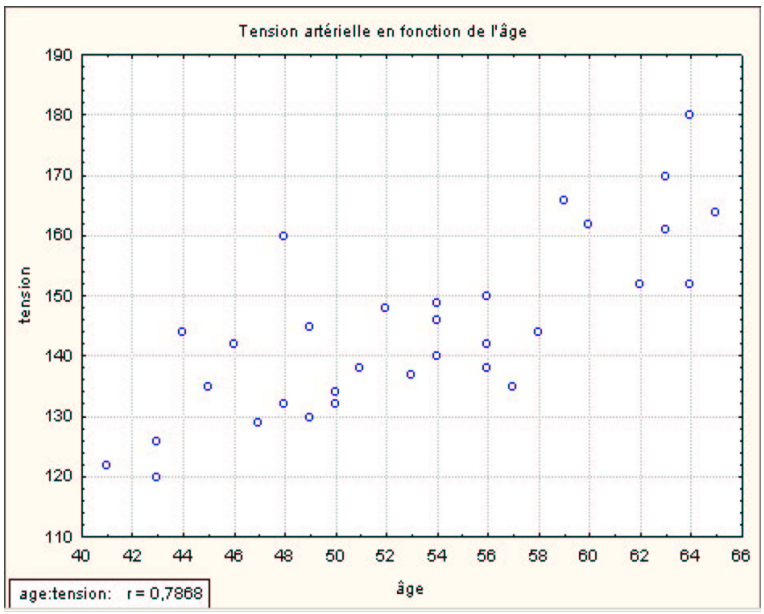
   Exemple : on a relevé la tension artérielle et l'âge sur un échantillon de 34 personnes. (extrait du fichier de données).

4. *Modèle exprimant la relation entre Y et X*

   On cherche à exprimer la relation entre la variable `tension` et la variable `âge` à l'aide d'une fonction mathématique du type $y = f(x)$. Graphiquement cela revient à représenter cette relation à l'aide d'une courbe (graphe de la fonction).

5. *Choix du modèle*
   - Quelle fonction mathématique utiliser ?
     ➡ Pour choisir le modèle de relation, on doit faire des observations sur un échantillon d'individus.
     Les données recueillies sur ces individus sont représentées graphiquement à l'aide d'un nuage de points.
     Si le nuage a une forme particulière s'apparentant à une courbe mathématique, on choisira la fonction mathématique correspondant à cette courbe.

– Exemple :
Graphique 2 :



La forme étirée et croissante du nuage suggère une relation positive de type linéaire entre la tension et l'âge.
Le coefficient de corrélation linéaire observé sur l'échantillon est $r = 0,7868$.

Modèle de régression linéaire : modèle le plus simple qui exprime la relation entre Y et X à l'aide une fonction linéaire. Graphiquement, la relation est représentée par une droite d'équation $y = b_0 + b_1 x$.
Ce modèle particulier comporte deux paramètres (coefficients) :
– le coefficient b_1 : la pente de la droite ; $b_1 > 0$ si la droite est croissante, $b_1 = 0$ si la droite est horizontale et $b_1 < 0$ si la droite est décroissante ;
– le coefficient b_0 : l'ordonnée du point d'intersection de la droite avec l'axe vertical en $x = 0$.

On va modéliser la relation entre la tension et l'âge à l'aide d'une droite.

6. *Equation générale du modèle de régression linéaire simple*
– *Si la relation était parfaitement linéaire* : sur l'échantillon, cela se traduirait par des points alignés et l'on pourrait écrire la relation entre Y et X sous la forme :

$$Y = b_0 + b_1 X$$

Connaissant l'âge x d'un individu, l'équation permettrait de déterminer exactement la tension artérielle y.

– *La relation observée sur l'échantillon n'est pas exacte.* Le nuage est étiré mais les points ne sont pas alignés. De plus, on voit que des personnes du même âge ont des tensions artérielles différentes.
Ces différences peuvent être expliquées par d'autres variables ayant une influence sur la variable `tension` et qui ne sont pas prises en compte dans le modèle, ou encore par des erreurs de mesures.

– *Equation générale*
Pour rendre compte de cette situation, on écrit la relation entre la tension et l'âge sous la forme générale suivante : droite + erreur

$$Y = b_0 + b_1 X + \varepsilon$$

Pour un âge x donné, la tension d'un individu est la somme de deux termes :

- 1er terme : $b_0 + b_1 x$ entièrement déterminé par l'âge ;
- 2ème terme : le terme d'erreur ε qui varie de façon aléatoire d'un individu à l'autre.

Le terme d'erreur ε est une variable aléatoire. Elle synthétise toutes les variables influant sur la tension et qui ne sont pas prises en compte.

La variable Y est aléatoire. La variable X est supposée non aléatoire, on la mesure sans erreur sur chaque individu.

– *Modèle appliqué aux observations*
On applique cette équation générale aux n observations de Y et aux valeurs correspondantes de X. On écrit le modèle sous la forme suivante :

$Y_i = b_0 + b_1 X_i + \varepsilon_i$ pour $i = 1, \ldots, n$

Pour chaque individu i, la variable aléatoire ε_i représente l'erreur commise, c'est-à-dire l'écart entre la valeur de Y observée et la valeur $b_0 + b_1 X_i$ donnée par la relation linéaire.

Dans le modèle, les variables ε_i ne sont pas observées et les coefficients b_1 et b_0 ne sont pas connus.

– *Conditions sur les erreurs*
Pour étudier le modèle, on pose des conditions sur les erreurs. On supposera que les erreurs sont des variables indépendantes, de même loi, centrées et de même variance (que l'on notera σ^2, condition d'homoscédasticité qu'il faudra vérifier).

2) Ajustement du modèle aux données. Estimation des coefficients de la droite par la méthode des moindres carrés

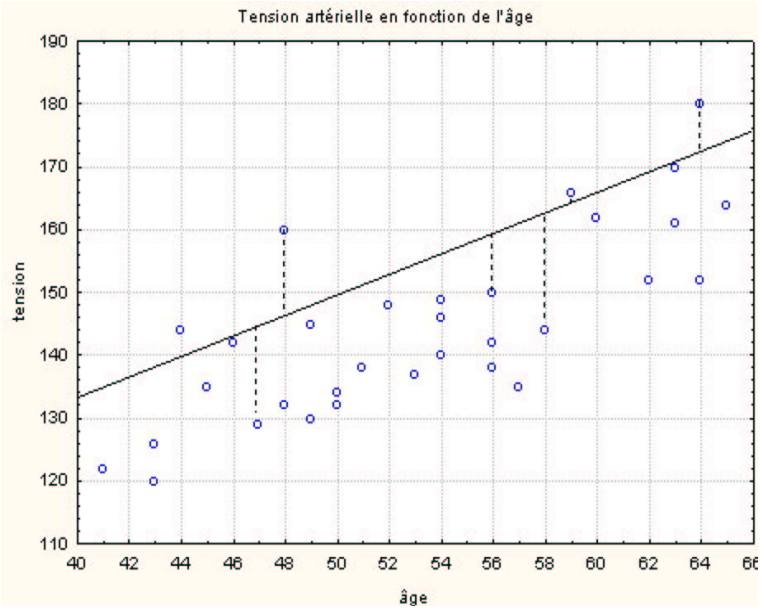
Le modèle étant posé, il faut estimer numériquement les paramètres du modèle, c'est-à-dire calculer les valeurs numériques des coefficients qui correspondent le mieux aux données.
Cela revient à déterminer la droite qui s'ajuste le mieux aux données, c'est-à-dire la droite qui est la plus proche des points.
Selon quel critère et quelles sont les formules permettant d'obtenir des valeurs estimées des coefficients ?

a) Le critère des moindres carrés

Parmi toutes les droites possibles, on cherche la droite pour laquelle la somme des carrés des écarts verticaux des points à la droite est minimale.

Sur le graphique, on a tracé une droite quelconque à travers les données et on représente les erreurs pour quelques points.

Graphique 3 :



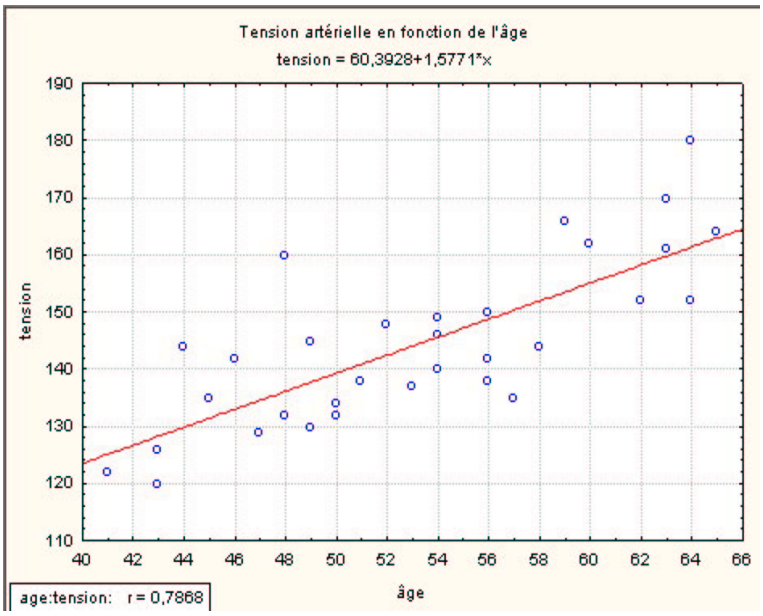
b) Formules de calcul des coefficients estimés

$$\hat{b}_1 = \frac{cov(x,y)}{var(x)} = r(x,y)\sqrt{\frac{var(y)}{var(x)}}$$

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

c) Résultats : droite de régression estimée ou droite des moindres carrés

Graphique 4 :



Remarque : La droite passe par le centre de gravité du nuage, le point moyen $(\overline{x}, \overline{y})$.

Sur l'exemple, on obtient : $\bar{x} = 53,85,\ \bar{y} = 145,32,\ r = 0,7868,\ s_X^* = 7,18,\ s_Y^* = 14,39.$

La méthode des moindres carrés fournit les coefficients estimés suivants sur l'exemple :

$$\hat{b}_1 = 1,5771 \text{ et } \hat{b}_0 = 60,3928.$$

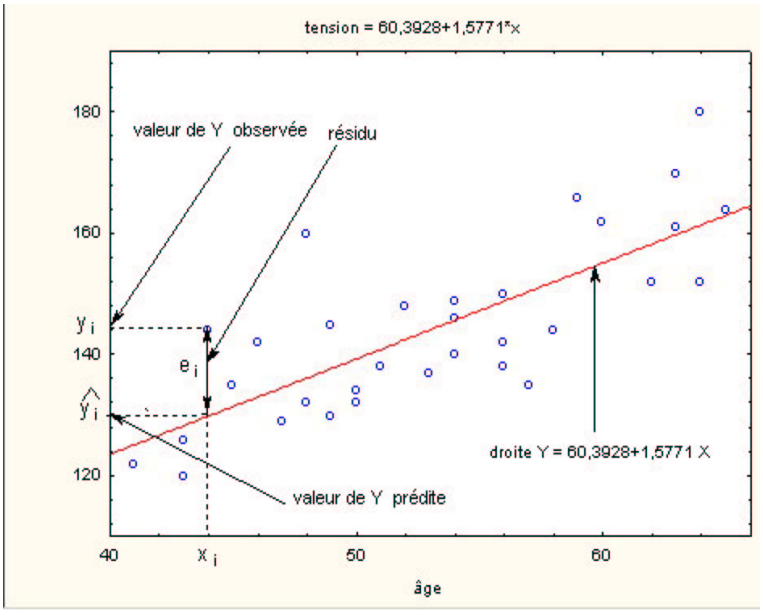
- La pente estimée de la droite : $\hat{b}_1 = 1,5771$.
  Interprétation : une augmentation de l'âge d'un an se traduit par une augmentation ($\hat{b}_1 > 0$) de la tension estimée à 1,5771.

- L'ordonnée à l'origine estimée : $\hat{b}_0 = 60,3928$.
  Interprétation : Ne pas extrapoler la droite au delà des limites du domaine observé de X. Ici, la droite a été ajustée pour des âges compris entre 40 et 66 ans. Le coefficient fixe la « hauteur » de la droite.
- la droite d'équation $y = 60,3928 + 1,5771\,x$ s'appelle la droite de régression estimée de Y sur X.

d) Valeurs ajustées, résidus et somme des carrés des résidus

Une fois les coefficients de la droite estimés, on calcule
- *pour chaque individu :*
- $\hat{y}_i = \hat{b}_0 + \hat{b}_1\,x_i$ s'appelle la valeur ajustée ou prédite de Y par le modèle.
- $e_i = y_i - \hat{y}_i$ s'appelle le résidu de l'observation i. C'est l'écart entre la valeur de Y observée sur l'individu $n^\circ i$ et la valeur prédite. Le résidu e_i est une approximation du terme d'erreur ε_i.
- la somme des carrés des résidus est $SCR = \sum e_i^2$. Elle mesure la distance de la droite de régression aux points du nuage de points qui est minimale au sens des moindres carrés.
- La statistique $\hat{\sigma}^2 = SCR/(n-2)$ est un estimateur sans biais de σ^2.

Graphique 5 :



Exemple numérique :
l'individu n° 5 a pour âge $x_5 = 44$.
La tension observée est $y_5 = 144$.
La tension prédite (ou estimée) par le modèle est
$\hat{y}_5 = 60,3928 + 1,5771 \times 44 = 129,7852$.
Le résidu pour l'observation n° 5 est $e_5 = 14,2148$. Le résidu est positif (point au dessus de la droite).
$SCR = 2605,569$
$\hat{\sigma}^2 = 2605,569/32 = 81,424$

e) Comment mesurer la qualité de l'ajustement

Pour le modèle choisi, Y peut varier en fonction :
- de X, selon la relation linéaire postulée
- d'autres variables non prises en compte et synthétisées dans le terme d'erreur.

On va mesurer la part de chacune de ces deux sources de variation pour évaluer la qualité de l'ajustement du modèle aux données.

1. *Décomposition de la variation totale des observations*
   On peut tout d'abord écrire la décomposition suivante :

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

   On peut montrer la propriété suivante :

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

   Ainsi, la variation totale des observations y_i autour de leur moyenne $\bar{y}$,

$$SCT = \sum(y_i - \bar{y})^2$$

   peut être décomposée en deux parties :

$$SCT = SCR + SCE$$

   où $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ représente la variance expliquée par la régression (mesure la variation des valeurs ajustées autour de la moyenne $\bar{y}$) et $SCR = \sum_{i=1}^{n} e_i^2$ représente la variance résiduelle ou non expliquée (partie de la variation totale qui n'est pas expliquée par le modèle de régression).

   Dans l'exemple, on obtient $SCT = 6839,442$, $SCE = 4233,873$ et $SCR = 2605,569$.

2. *Le coefficient de détermination R^2*
   Afin d'avoir une idée globale de la qualité de l'ajustement linéaire, on définit R^2 le coefficient de détermination qui est le carré du coefficient de corrélation R :

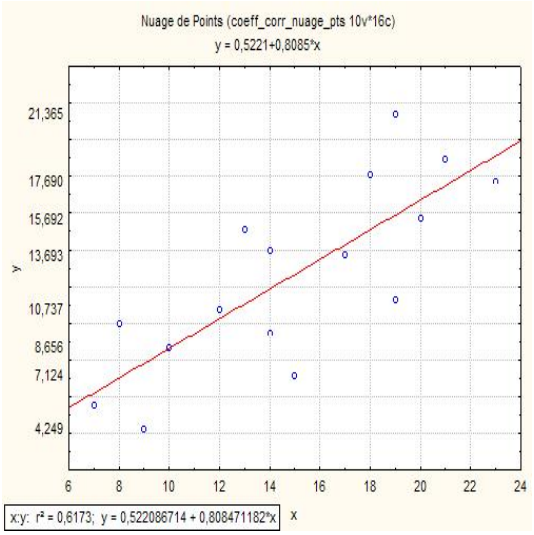
$$R^2 = \frac{SCE}{SCT}$$

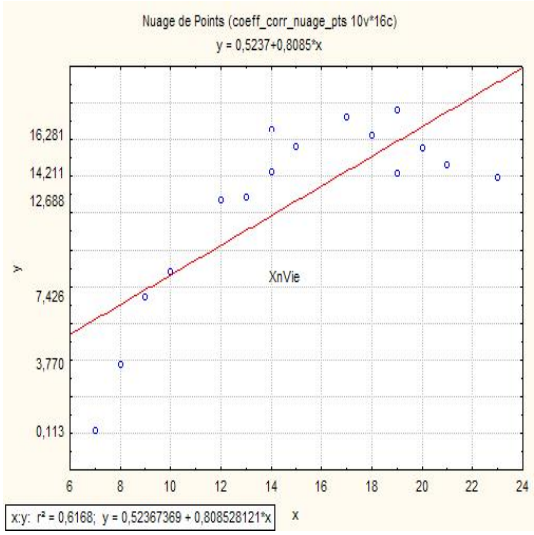
   Il mesure la part de la variation totale de Y expliquée par le modèle de régression sur X.

   Pour l'exemple, on a $r^2 = 0,619 = \dfrac{4233,873}{6839,442}$. Le modèle de régression explique 61,91% de la variation totale.

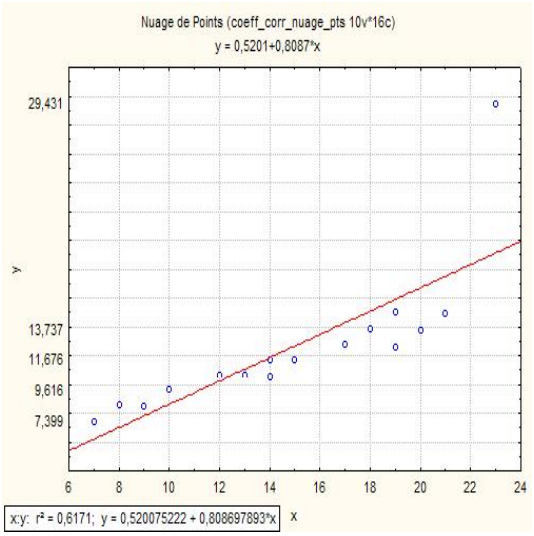
   Cas particuliers :
   - si $R^2 = 0$, le modèle n'explique rien, les variables X et Y ne sont pas corrélées linéairement.
   - si $R^2 = 1$, les points sont alignés sur la droite, la relation linéaire explique toute la variation.
   - une valeur de R^2 proche de 1 (voir chapitre corrélation de Pearson) est *nécessaire* pour avoir un ajustement raisonnable mais en aucun cas suffisant, par exemple :
   Les données sont extraites du livre "La régression : nouveaux regards sur une ancienne méthode statistique" de Tomassone, Lesquoy et Millier (1983) .

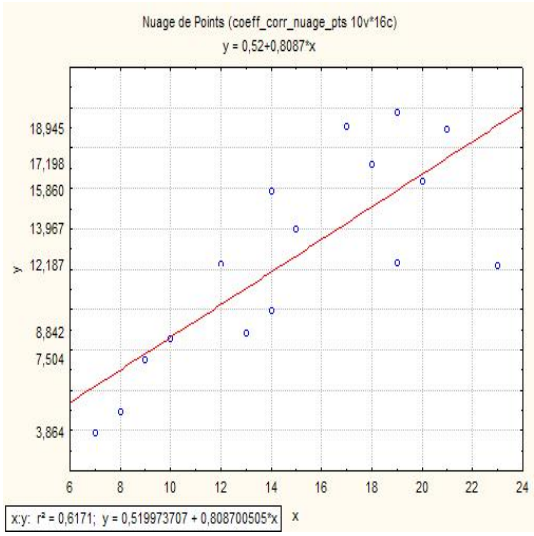
Graphique 6 :



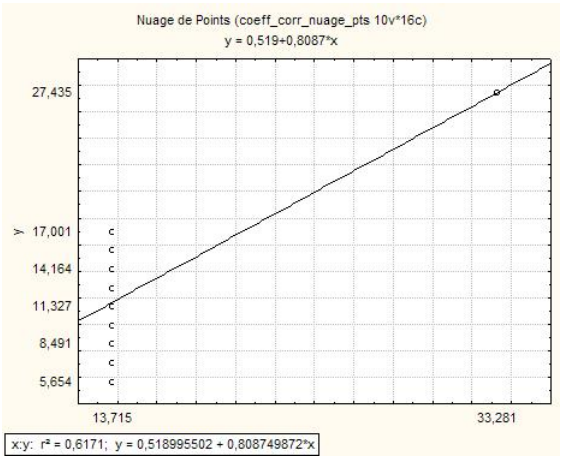


Graphique 7 :





Graphique 8 :



3) Tests

a) Test global de significativité de la régression

Il paraît raisonnable de tester la significativité globale du modèle, c'est à dire tester si tous les coefficients sont supposés nuls, excepté la constante.

Cela correspond dans le cas de la régression linéaire simple à $H_0 : b_1 = 0$ contre $H_1 : b_1 \neq 0$

La statistique du test : statistique F de Fisher
On utilise la statistique, notée F définie par la formule :

$$F = (n-2)\frac{R^2}{1-R^2} = \frac{SCE/1}{SCR/(n-2)}$$

.

Loi de F sous H_0
La statistique F suit la loi de Fisher à $(1, n-2)$ ddl.

Région de rejet de H_0
Sous H_0, on s'attend à observer une valeur de F proche de 0. Plus la valeur de F est grande et plus elle est en faveur de H_1.
La région de rejet est située à l'extrémité droite du domaine .

Décision
Règle basée sur la p-valeur : si $\alpha_{obs} \leq \alpha$, on rejette H_0 au risque d'erreur α.

$$\alpha_{obs} = P_{H_0}(F(1, n-2) > (n-2)\frac{r^2}{1-r^2})$$

Dans Statistica, les valeurs observées de F sont données ainsi que la p-valeur.

b) Tests sur les paramètres

Reprenons l'exemple de la tension en fonction de l'âge. Nous avons modélisé la tension Y par l'âge X. Il paraît raisonnable de se poser les questions suivantes :
(a) est-ce-que le coefficient b_1 est non nul, autrement dit la variable X a-t-elle réellement une influence sur Y ?
(b) est-ce-que le coefficient b_0 est non nul, autrement dit faut-il une constante dans le modèle ?

Rappelons que le modèle utilisé est le suivant

$$y_i = b_0 + b_1 x_i + \epsilon_i.$$

Nous pouvons expliciter les questions précédentes en terme de test d'hypothèse :
(a) correspond à $H_0 : b_1 = 0$, contre $H_1 : b_1 \neq 0$
(b) correspond à $H_0 : b_0 = 0$, contre $H_1 : b_0 \neq 0$

– *La statistique du test : statistique T de Student*
On utilise la statistique, notée T définie par la formule, pour $j = 0, 1$:

$$T = \frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}$$

- *loi de T sous H_0*
  La statistique T suit la loi de Student à $n-2$ ddl.
- *Région de rejet de H_0*
  Sous H_0, on s'attend à observer une valeur de T proche de 0. Plus la valeur de $|T|$ est grande et plus elle est en faveur de H_1.
  La région de rejet est située à l'extrémité droite et à l'extrémité gauche du domaine (test bilatéral).
- *Décision*
  Règle basée sur la p-valeur : si $\alpha_{obs} \leq \alpha$, on rejette H_0 au risque d'erreur α.

$$\alpha_{obs} = 2\ P_{H_0}(Student(n-2) > |\frac{\hat{b}_j}{\hat{\sigma}_{\hat{b}_j}}|)$$

Dans Statistica, les valeurs observées de T sont données ainsi que la p-valeur.

Remarque : on peut remarquer que dans le cas de la régression linéaire simple, il est équivalent de tester la significativité globale du modèle ou bien de tester $H_0 : b_1 = 0$, contre $H_1 : b_1 \neq 0$. Effectivement, on a $T^2 = F$. Au niveau des lois l'égalité est aussi valable bien évidemment et nous avons que le carré d'un Student à $(n-2)$ ddl est une loi de Fisher à $(1, n-2)$ ddl. Bien entendu le quantile $(1-\alpha)$ d'une loi de Fisher correspond au quantile $1-\alpha/2$ d'une loi de Student.

Pour l'exemple, on obtient

N=34	Synthèse de la Régression; Variable Dép. : tension (donnees-bou R= ,78678955 R²= ,61903780 R² Ajusté = ,60713273 F(1,32)=51,998 p<,00000 Err-Type de l'Estim.: 9,0235					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(32)	niveau p
OrdOrig.			60,39282	11,87925	5,083893	0,000016
age	0,786790	0,109110	1,57709	0,21871	7,210952	0,000000

$\hat{\sigma} = \sqrt{2605,569/32} = 9,0235$

(a) $H_0 : b_1 = 0$, contre $H_1 : b_1 \neq 0$

$t^2_{obs} = 7,211^2 = 51,998 = f_{obs}$

$\alpha_{obs} = 0$, on rejette donc H_0 au risque 5%.

(b) $H_0 : b_0 = 0$, contre $H_1 : b_0 \neq 0$

$t_{obs} = 5,084$

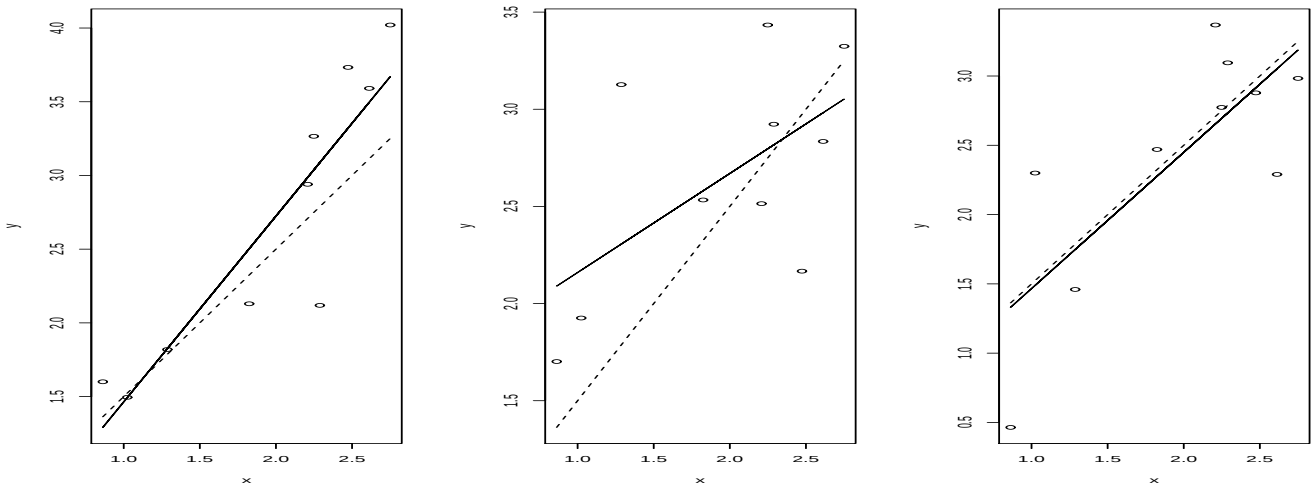
$\alpha_{obs} \approx 0$, on rejette donc H_0 au risque 5%.

4) Intervalles de confiance et intervalles de prévision

Jusqu'ici tous les calculs (estimation des paramètres de la droite, coefficient de détermination) ont été effectués sur les données de l'échantillon.

Exemple : Nous avons réalisé une expérience en mesurant 10 valeurs (x_i, y_i) issues du modèle $Y = X + 0.5 + \varepsilon$, où ε est choisi comme suivant une loi $\mathcal{N}(0, 0.025)$. A partir de ces 10 valeurs nous avons obtenu des estimateurs de b_0 et de b_1 : $\hat{b}_0 = 0.204$ et $\hat{b}_1 = 1.26$. Si nous refaisions deux fois cette expérience, nous mesurons 10 nouveaux couples de données (x_i, y_i). A partir de ces données, nous obtenons deux nouveux estimateurs de b_10 et de b_1 : $\hat{b}_0 = 1.654$ et $\hat{b}_1 = 0.51$ puis : $\hat{b}_0 = 0.482$ et $\hat{b}_1 = 0.983$. Les estimateurs sont fonctions des données mesurées et changent donc avec les observations collectées . Les vraies valeurs de b_0 et de b_1 sont inconnues et ne changent pas. Le trait en pointillé représente la vraie droite de régression et le trait plein son estimation.

Graphique 9 :



Que peut-on en conclure sur la relation entre les deux variables pour la population toute entière ?

a) Intervalles de confiance

On supposera dans la suite que les ε_i en plus d'être indépendants, de même loi, centrées et de même variance, sont distribuées suivant une loi $\mathcal{N}(0, \sigma^2)$.

La valeur ponctuelle d'un estimateur est en général insuffisante et il est nécessaire de lui adjoindre un intervalle de confiance.

(i) Un IC de b_0 au niveau $1 - \alpha$ est donné par :

$$\left[\hat{b}_0 - t\ \hat{\sigma}_{\hat{b}_0}, \hat{b}_0 + t\ \hat{\sigma}_{\hat{b}_0}\right]$$

où t représente le quantile de niveau $(1 - \alpha/2)$ d'une loi de Student $n - 2$.

(ii) Un IC de $b_1$au niveau $1 - \alpha$ est donné par :

$$\left[\hat{b}_1 - t\ \hat{\sigma}_{\hat{b}_1}, \hat{b}_1 + t\ \hat{\sigma}_{\hat{b}_1}\right].$$

Nous pouvons également donner un intervalle de confiance de la droite de régression.
Un IC de y_i au niveau $1 - \alpha$ est donné par :

$$\left[\hat{y}_j - t\ \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \hat{y}_j + t\ \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right].$$

En calculant les IC pour tous les points de la droite, nous obtenons une hyperbole de confiance. En effet, lorsque x_j est proche de $\bar{x}$, le terme dominant de la variance est $1/n$, mais dès que x_j s'éloigne de $\bar{x}$, le terme dominant est le terme au carré.

b) Intervalles de prédiction

Un des buts de la régression est de proposer des prédictions pour la variable à expliquer Y. Soit x_{n+1} une nouvelle valeur de la variable X, nous voulons prédire y_{n+1}. Le modèle indique que

$$y_{n+1} = b_0 + b_1 x_{n+1} + \varepsilon_{n+1}$$

Nous pouvons prédire la valeur correspondante grâce au modèle estimé

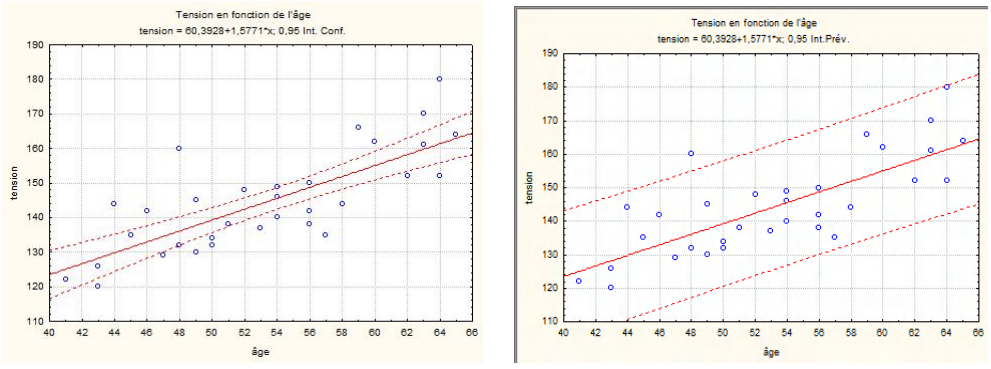
$$\hat{y}^p_{n+1} = \hat{b}_0 + \hat{b}_1 x_{n+1}.$$

En utilisant la notation $\hat{y}^p_{n+1}$, nous souhaitons insister sur la notion de prévision : la valeur pour laquelle nous effectuons la prévision ici la $(n+1)^{ème}$ n'a pas servi dans le calcul des estimateurs. Remarquons que cette quantité serait différente de la valeur ajustée, notée $\hat{y}_i$, qui elle fait intervenir la $i^{ème}$ observation.
Deux types d'erreurs vont entacher notre prévision, la première due à la non connaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.
La variance augmente lorsque x_{n+1} s'éloigne du centre de gravité du nuage. Faire de la prévision lorsque x_{n+1} est "loin" de $\bar{x}$ est donc périlleux, la variance de l'erreur de prévision peut alors être très grande.
Un IC de y_{n+1} au niveau $1-\alpha$ est donné par :

$$\left[\hat{y}^p_{n+1} - t\ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\ ,\ \hat{y}^p_{n+1} + t\ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right].$$

Cette formule exprime que plus le point à prévoir est éloigné de $\bar{x}$, plus la variance de la prévision et donc l'IC seront grands. Une approche intuitive consiste à remarquer que plus une observation est éloignée du centre de gravité, moins nous avons d'information sur elle. Lorsque la valeur à prévoir est à l'intérieur de l'étendue des x_i, le terme dominant de la variance est la valeur 1 et donc la variance est relativement constante. Lorsque x_{n+1} est en dehors de l'étendue des x_i, le terme dominant peut être le terme au carré, et la forme de l'intervalle sera à nouveau une hyperbole.

Graphique 10 :



L'intervalle de prévision est plus grand que l'intervalle de confiance. L'intervalle de confiance admet une forme hyperbolique.

5) Analyse des résidus

Les différentes phases d'un régression peuvent se résumer par trois étapes successives.

1. La première est la modélisation : nous avons supposé que la variable Y est expliquée de manière linéaire par la variable X via le modèle de régression $Y = b_0 + b_1 X + \epsilon$.
2. La seconde est l'étape d'estimation : nous avons ensuite estimé les paramètres grâce aux données récoltées. Les hypothèses sur le résidu ϵ ont permis d'établir des propriétés statistiques des estimateurs obtenus.
3. Enfin la troisième étape est celle de validation à laquelle ce paragraphe est consacré. Nous aborderons le problème de la validation des hypothèses sur les résidus et la qualité de l'ajustement observation par observation.

L'examen des résidus constitue une étape primordiale de la régression linéaire. Cette étape est essentiellement fondée sur des méthodes graphiques fournies par Statistica, et il est donc difficile d'avoir des règles strictes de décision. L'objectif de cette partie est de présenter ces méthodes graphiques.

a) Vérification des conditions

- Analyse de la normalité
  L'hypothèse de normalité sera examinée à l'aide d'un graphique comparant les quantiles des résidus estimés aux quantiles sous l'hypothèse de normalité. Ce type de graphique est appelé droite de Henry. Si les résidus ne sont pas normalement distribués, ils vont s'écarter de la droite.
- Analyse de l'homoscédasticité
  Il n'existe pas de procédure précise pour vérifier l'hypothèse d'homoscédasticité. Nous proposons plusieurs graphiques possibles pour détecter une hétéroscédasticité. Il est recommandé de tracer les résidus en fonction des valeurs ajustées $\hat{y}_i$, c'est-à-dire tracer les couples de points $(\hat{y}_i, e_i)$. Si une structure apparaît (tendance, cône, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée.
- Analyse de la structure des résidus
  Les résidus sont supposés être indépendants. L'indépendance est très difficile à tester de manière formelle. Le test de Durbin-Watson est le plus souvent utilisé, consiste à tester H_0 : l'indépendance, contre H_1 : les résidus sont non-indépendants d'un certaine façon. Cependant il existe de nombreux modèles de non-indépendance qui ne seront pas forcément détectés par ce test.

b) Ajustement individuel au modèle et valeur aberrante

Pour analyser la qualité de l'ajustement d'une observation, il faut regarder le résidu correspondant à cette observation. Si ce résidu est anormalement élevé (sens que nous allons préciser) alors l'individu i est appelé individu aberrant ou atypique. Il convient alors d'essayer d'en comprendre la raison (erreur de mesure, individu provenant d'un sous-population) et éventuellement d'éliminer cette observation car elle peut modifier les estimations.

Une valeur aberrante ou atypique est une observation qui est mal expliquée par le modèle et admet un résidu élevé.

Généralement les données aberrantes sont détectées en traçant des graphiques. La détection des données aberrantes ne dépend que de la grandeur des résidus.

Valeurs prévues vs. résidus. Ce tracé est particulièrement utile pour tester l'hypothèse de linéarité concernant la relation entre les variables indépendantes et la variable dépendante. Plus précisément, si la relation est linéaire, les résultats des résidus doivent former un "nuage" homogène autour de la droite centrale.
Valeurs prévues vs. observées. Ce tracé est particulièrement utile pour identifier des groupes potentiels d'observations qui ne sont pas bien prévus.
Valeurs observées vs. résidus. Ce tracé est très utile pour détecter des points atypiques ou groupes d'observations qui ont systématiquement des prévisions trop fortes ou trop faibles.

Résidus vs. résidus Supprimés. Les résidus supprimés sont les résidus qui seraient obtenus si l'observation respective était exclue de l'estimation de la régression multiple (c'est-à-dire, des calculs des coefficients de régression). Ainsi, s'il existe de fortes divergences entre les résidus supprimés et les résidus, nous pouvons en conclure que les coefficients de régression ne sont pas très stables, c'est-à-dire qu'ils sont fortement affectés par l'exclusion de simples observations
Tracé des points atypiques. Cette option de statistica permet d'identifier les points atypiques même dans des très grands fichiers de données, puisque seules les observations extrêmes sont tracées.

Autres mesures diagnostiques
La distances de Cook mesure l'influence de l'observation i sur l'estimation du paramètre b_j. Pour bâtir une telle mesure, nous considérons la distance entre le coefficient estimé $\hat{b_j}$ et le coefficient $\hat{b_{j(i)}}$ que l'on estime en enlevant l'observation i, mais en gardant le même modèle et toutes les autres observations bien évidemment. Si la distance est grande alors l'observation i influence beaucoup l'estimation de β, puisque la laisser ou l'enlever conduit à des estimations éloignées. La distance de Mahalanobis mesure si l'observation est atypique par rapport aux variables explicatives uniquement (point levier).

Une fois repérées et notées, il est bon de comprendre pourquoi ces valeurs sont atypiques : est-ce une erreur de mesure ou d'enregistrement ? Proviennent-elles d'une autre population ?.. Nous recommandons d'enlever ces points de l'analyse. Si vous souhaitez les conserver malgré tout, il est indispensable de s'assurer que ce ne sont pas des valeurs influentes : les coefficients et les interprétations tirées du modèle ne doivent pas trop varier avec ou sans ces observations.

IV Regression multiple

1) Définition

Le modèle de régression multiple est une généralisation du modèle de régression simple lorsque les variables explicatives sont en nombre fini. Nous supposons donc que les données collectées suivent le modèle suivant :

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i, \qquad i = 1, \cdots, n$$

Tous les résultats précédents se généralisent dans le cas général.

2) Tests sur les paramètres

Il paraît raisonnable de se poser les questions suivantes :
(a) est-ce-que la variable X_j a-t-elle réellement une influence sur Y ?
(b) tester la significativité globale du modèle, c'est à dire est-ce-que que tous les coefficients sont supposés nuls, excepté la constante ?

Nous pouvons expliciter les questions précédentes en terme de test d'hypothèse :
(a) correspond à $H_0 : b_j = 0$, contre $H_1 : b_j \neq 0$

Cela revient au test de Student à $n-p-1$ dll ($n-2$ ddl dans le cas simple, p=1) présenté dans le paragrahe précédent.

(b) correspond à H_0 : tous les $b_j = 0, j = 1, ..., p$ contre H_1 : il y a au moins un b_j non nul

– *La statistique du test : statistique F de Fisher*

On utilise la statistique, notée F définie par la formule, pour :

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} = \frac{SCE/p}{SCR/(n-p-1)}$$

.

- *loi de F sous H_0*
  La statistique F suit la loi de Fisher à $(p, n-p-1)$ ddl.
- *Région de rejet de H_0*
  Sous H_0, on s'attend à observer une valeur de F proche de 0. Plus la valeur de F est grande et plus elle est en faveur de H_1.
  La région de rejet est située à l'extrémité droite du domaine .
- *Décision*
  Règle basée sur la p-valeur : si $\alpha_{obs} \leq \alpha$, on rejette H_0 au risque d'erreur α.

$$\alpha_{obs} = P_{H_0}(F(p, n-p-1) > \frac{r^2}{1-r^2}\frac{n-p-1}{p})$$

Dans Statistica, les valeurs observées de F sont données ainsi que la p-valeur.

3) Choix des variables

Dans une régression multiple, il se peut que le nombre p des variables disponibles soit grand. Cette quantité d'information est parfois superflue ou redondante. Ainsi la diminution du nombre de vatiables réellement intéressantes dans la régression est envisageable. Soit on part du modèle complet et on retire des variables en utilisant un critère décrit sous Statistica (pas à pas descendant). Soit on part d'une régression simple et on ajoute des variables qui enrichissent le modèle (pas à pas ascendant). Sous Statistica, dans ces deux cas, on arrête d'enlever ou d'ajouter une variable au modèle en analysant la statistique F.