Régression linéaire: Partie II

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1ère année Médecine

24 Avril 2024

Plan du cours

Plan du cours

Plan du cours

Plan du cours

Plan du cours

Objectif de la régression

- On dispose de 2 caractères (variables) quantitatives X et Y. On distingue deux objectifs:
  1. On cherche à savoir s'il existe un lien entre X et Y.
  2. On cherche à savoir si X a une influence sur Y et éventuellement prédire Y à partir de X.

Régression: définition

- Il s'agit ici d'étudier le **lien entre 2 variables quantitatives**.
- La variable que l'on veut modéliser est appelée variable **a expliquer** ou variable **dépendante**, **réponse**, **diagnostique (médecine)**.
- La ou les variables qui sont utilisées pour modéliser la variable a expliquer sont appelées variables **explicatives** ou variables **indépendantes** (ce terme est à éviter), **imposée** ou **symptômes** (en médecine).

Régression

- Lorsque les valeurs prices par une variable explicative sont choisies pas l'expérimentateur, on dit que la variable explicative est **contrôlée** (on parle encore de **facteur contrôlé**). Lorsque les valeurs ne sont pas choisies, mais simplement mesurées, on parle de variables **non contrôlées**.
- Les paramètres qui interviennent dans les formules de modélisation s'appellent **cœffcients** du modèles.
- La partie non expliquée désignée dans les formules par ε est appelée "**reste**" ou "**résidu**" ou "**erreur**" du modèle.

Démarche pour la régression

La régression comporte 4 étapes:

1. Choix d'un modèle $Y = f(X)$;
2. Détermination de la valeur numérique des paramètres du modèle;
3. Détermination de la signification statistique des paramètres du modèle (statistique inférentielle).
4. Validation du modèle (statistique inférentielle).

Démarche pour la régression

La régression comporte 4 étapes:

1. Choix d'un modèle $Y = f(X)$;
2. Détermination de la valeur numérique des paramètres du modèle;
3. Détermination de la signification statistique des paramètres du modèle (statistique inférentielle).
4. Validation du modèle (statistique inférentielle).

Démarche pour la régression

La régression comporte 4 étapes:

1. Choix d'un modèle $Y = f(X)$;
2. Détermination de la valeur numérique des paramètres du modèle;
3. Détermination de la signification statistique des paramètres du modèle (statistique inférentielle).
4. Validation du modèle (statistique inférentielle).

Démarche pour la régression

La régression comporte 4 étapes:

1. Choix d'un modèle $Y = f(X)$;
2. Détermination de la valeur numérique des paramètres du modèle;
3. Détermination de la signification statistique des paramètres du modèle (statistique inférentielle).
4. Validation du modèle (statistique inférentielle).

Démarche pour la régression

La régression comporte 4 étapes:

1. Choix d'un modèle $Y = f(X)$;
2. Détermination de la valeur numérique des paramètres du modèle;
3. Détermination de la signification statistique des paramètres du modèle (statistique inférentielle).
4. Validation du modèle (statistique inférentielle).

Régression: objectif

- La régression est une forme de modélisation. Elle peut avoir plusieurs objectifs:
  - **Description:** trouver le meilleur modèle fonctionnel liant la variable dépendante y à la (aux) variable(s) indépendante(s) x. Estimer la valeur la plus probable des paramètres du modèle, ainsi que leur intervalle de confiance (Statistique inférentielle).
  - **Inférence**(Statistique inférentielle): tester des hypothèses précises se rapportant aux paramètres du modèle dans la population statistique: ordonnée à l'origine, pente(s).
  - **Prédiction:** prévoir ou prédire les valeurs de la variable dépendante pour de nouvelles valeurs de la (des) variable(s) indépendante(s).

Plan de cours

Régression linéaire: modèle

- **Regression linéaire:** modèle le plus simple:

$$Y = f(X) + \varepsilon = \theta + \beta \times X + \varepsilon$$

$$y_i = \theta + \beta \times x_i + \varepsilon_i$$

- Interprétation
- Estimations des paramètres
- Prédiction.

Régression linéaire: modèle

- **Regression linéaire:** modèle le plus simple:

$$Y = f(X) + \varepsilon = \theta + \beta \times X + \varepsilon$$

$$y_i = \theta + \beta \times x_i + \varepsilon_i$$

  - Interprétation
  - Estimations des paramètres
  - Prédiction.

Régression linéaire: modèle

- **Regression linéaire:** modèle le plus simple:

$$Y = f(X) + \varepsilon = \theta + \beta \times X + \varepsilon$$

$$y_i = \theta + \beta \times x_i + \varepsilon_i$$

  - Interprétation
  - Estimations des paramètres
  - Prédiction.

Régression linéaire: modèle

- **Regression linéaire:** modèle le plus simple:

$$Y = f(X) + \varepsilon = \theta + \beta \times X + \varepsilon$$

$$y_i = \theta + \beta \times x_i + \varepsilon_i$$

  - Interprétation
  - Estimations des paramètres
  - Prédiction.

Régression linéaire: modèle

- θ représente l'ordonnée à l'origine et β représente la pente de la droite.
- On utilise des lettres grecques pour représenter l'ordonnée à l'origine et la pente pour bien insister sur le fait que ce sont des paramètres **inconnus**.
- Leur valeur respective serait connue si on avait accès à toute la population, ce qui n'est jamais le cas en pratique. Il nous faudra donc les estimer.

Plan de cours

Principe de l'estimation

- Estimer θ et β tel que ε petits +++
- ε_i: écart entre la droite et le point i

$$y_i = \theta + \beta \times x_i + \varepsilon_i$$

$$\Rightarrow \varepsilon_i = y_i - (\theta + \beta \times X)$$

- Somme des Carrés des Écarts

$$SCE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

- Estimer θ et β tel que: SCE minimum

Estimation de la pente β

- La pente β est donnée par la formule suivante: $\beta = \frac{Cov(X,Y)}{Var(X)}$
  - La variance de X est **estimé** par (dans le cas d'un échantillon):

$$S'^2_X = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i)^2 - n(\overline{X})^2}{n-1}$$

  - La covaiance de X et Y est **estimé** par:

$$\widehat{Cov}(X,Y) = S'_{X,Y} = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n-1}$$

- D'où β est estimé par $\quad b = \frac{S'_{X,Y}}{S'^2_X}$

Estimation de θ:

- La droite passe par m_Y et m_X: $m_Y = \theta + \beta m_X$; où $m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

  D'où θ est estimé par a:

$$a = m_Y - b m_X$$

Remarque

Une fois les paramètres a et b sont estimés, on en déduit les valeurs ajustées $\hat{y}_i = a + b x_i$ puis les résidus estimés $\varepsilon_i = y_i - \hat{y}_i$.

Estimation de θ:

- La droite passe par m_Y et m_X: $m_Y = \theta + \beta m_X$; où

  $m_Y = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $m_X = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

  D'où θ est estimé par a:

$$a = m_Y - b m_X$$

Remarque

Une fois les paramètres a et b sont estimés, on en déduit les valeurs ajustées $\hat{y}_i = a + b x_i$ puis les résidus estimés $\varepsilon_i = y_i - \hat{y}_i$.

Plan de cours

Coefficient de corrélation

- Pour deux variables X et Y, le cœfficient de corrélation linéaire $r = \rho(X, Y) = cor(X, Y)$ vaut:

$$r = cor(X, Y) = \frac{cov(X, Y)}{S_X S_Y}$$

- Estimation du coefficient de corrélation entre X et Y:

$$\hat{r} = \hat{cor}(X, Y) = \frac{S'_{(X,Y)}}{S'_X S'_Y} .$$

Coefficient de détermination

- Soit $\hat{y}_i = a + b \times x_i$
- $\varepsilon_i = y_i - \hat{y}_i$.
- La variation résiduelle est

$$S_e'^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2 = MSE$$

- Soit la décomposition suivante:

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

- D'où l'équation d'analyse des variances:

$$V(Y) = V(\hat{Y}) + V(\varepsilon).$$

Coefficient de détermination

- Total sum of squares (variabilité totale de Y):

  $SCT = \sum (y_i - \overline{Y})^2 = (n-1) S'^2_Y$.

- Regression sum of squares (variabilité expliquée par le modèle, c-à-d $\hat{y}_i$): $SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{Y}})^2 = (n-1) \frac{S'_{(X,Y)}}{S'^2_X}$.

- error sum of squares (variabilité des résidus):

  $SCR = \sum_{i=1}^{n} \varepsilon_i^2 = (n-2) S'^2_e$.

- $SCT = SCE + SCR$.

Coefficient de détermination

- Le Coefficient de détermination R^2 est donnée par le rapport suivant:

$$\begin{aligned} R^2 &= \frac{SCE}{SCT} = \frac{\text{Part de variance expliquée par la régression}}{\text{Variance totale}} \\ &= 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i - \overline{Y})^2} = 1 - \frac{n-2}{n-1} \frac{S_e'^2}{S_Y'^2} \\ &= \frac{b^2 (\sum_{i=1}^{n} x_i^2 - n(\overline{X})^2)}{\sum_{i=1}^{n} y_i^2 - n(\overline{Y})^2} = (corr(Y, \hat{Y}))^2 \end{aligned}$$

  - Si $R^2 = 0$, le modèle de régression linéaire est inadapté (le modèle ne sert à rien où les variables X et Y ne sont pas corrélées linéairement).
  - Si R^2 est proche de 0, cela vaut dire que la variable X n'explique pas bien la variable réponse Y (au moins de façons linéaire).
  - Si $R^2 = 1$ le modèle explique tout: les points de l'échantillon sont parfaitement alignés (X explique bien la variable réponse Y).

Plan de cours

Conditions sur les erreurs

- $C1$: $\mathbb{E}(\varepsilon_i) = 0$, $\forall i$
- $C2$: $Cov(\varepsilon_i, \varepsilon_j) = 0$, si $i \neq j$ et $Var(\varepsilon_i) = \sigma^2$, $\forall i$

Donc, les erreurs sont supposées centrées, de même variance et non corrélées.

Hypothèses sur les erreurs

Sous les hypothèses précedentes, on a:

- b est un estimateur sans biais et efficace de β.
- a est un estimateur sans biais et efficace de θ.
- $S_e'^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2 = MSE$ est un estimateur sans biais de σ^2

Plan de cours

Hypothèses sur les erreurs

Dans cette section, on suppose de plus que les conditions suivantes sont vérifiées:

- $\varepsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$.
- ε_i sont mutuellement indépendants.

Intervalle de confiance des paramètres

Sous les hypothèses précédentes, on a

- Intervalle de confiance pour θ est:

$$IC(\theta) = \left[a \pm t_{(n-2,1-\alpha/2)} s'_e \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s'^2_x}} \right]$$

- Intervalle de confiance pour β est:

$$IC(\beta) = \left[b \pm t_{(n-2,1-\alpha/2)} s'_e \sqrt{\frac{1}{(n-1)s'^2_x}} \right]$$

Intervalles de confiance pour la droite de régression

- Il s'agit d'un intervalle de confiance pour $\mathbb{E}(Y_{n+1}|x_{n+1})$, la réponse moyenne à la valeur x_{n+1}.
- Pour x_{n+1} donnée, soit $\hat{y}_{n+1} = a + b\,x_{n+1}$ (l'estimateur de $\mathbb{E}(Y_{n+1}|x_{n+1})$).
- Intervalle de confiance pour $\mathbb{E}(Y_{n+1}|x_{n+1})$ au niveau de confiance $1-\alpha$ est:

$$IC(y_{n+1}) = \left[\hat{y_{n+1}} \pm t_{(n-2,1-\alpha/2)} s'_e \sqrt{\frac{1}{n} + \frac{(x_{n+1}-\overline{X})^2}{(n-1)s'^2_x}}\right]$$

Plan de cours

Intervalles de prédiction

- On désire prévoir à l'aide du modèle, la valeur de la variable y pour une valeur non observé de x.
- Soit x_{n+1} une nouvelle valeur, pour laquelle nous voulons prédire y_{n+1}.
- D'après le modèle on a $y_{n+1} = \theta + \beta x_{n+1} + \varepsilon_{n+1}$, où y_{n+1} et ε_{n+1} sont des variables aléatoires.
- La prédiction naturelle est alors : $\hat{y}_{n+1} = a + b x_{n+1}$.
- L'erreur de prédiction est définie par $\hat{\varepsilon}_{n+1} = \hat{y}_{n+1} - y_{n+1}$.

Intervalle de prévision: remarque

- De types d'erreurs vont entacher notre prévision:
  - La première est due à la non connaissance de ε_{n+1}.
  - La second à l'incertitude sur les estimateurs a et b.

Intervalle de prévision

- L'intervalle de prédiction est donc un intervalle dans lequel une future observation y_{n+1} va tomber avec une certaine probabilité (différent d'un intervalle de confiance).
- sous les hypothèses du modèle (incluant l'hypothèse de normalité), on a, l'intervalle de prédiction pour y_{n+1} en x_{n+1}, au niveau de confiance $1-\alpha$ (par la droite de régression $Y = a + bX$) est

$$IP(y_{n+1}) = \left[\hat{y_{n+1}} \pm t_{(n-2,1-\alpha/2)} s'_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{X})^2}{(n-1)s'^2_x}} \right]$$

Remarques: IC vs IP

- Les longueurs des deux types d'intervalles croissent lorsque x_{n+1} s'éloigne de $\overline{X}$.
- L'IC de la droite de régression ne convient pas pour effectuer des prévisions puisqu'il concerne la vraie réponse moyenne au point $X = x_{n+1}$, soit un paramètre de la population, et non une nouvelle observation.
- L'IP en x_{n+1} est toujours plus grand que l'IC en x_{n+1} car il dépend de l'erreur associée aux futures observations.
- L'IP n'est valide que pour une nouvelle observation à la fois. Pour une série de nouvelles observations, il faut mettre à jour le modèle au fur et à mesure.

Plan de cours

Test de la pente

- Pratiquement, les hypothèses portant sur β ont plus d'intérêt que celles portant sur θ.
- On va donc se limiter à tester la nullité de la pente β (absence de liaison linéaire entre x et y).
  - Si $\beta = 0 \Rightarrow$ pas de lien entre Y et X.
  - Lien entre Y et X est-il significatif? $\Rightarrow \beta \neq 0$?
  - b estimation de β, donc le hasard $\Rightarrow$ fluctuation de b observé $\Rightarrow$ Test statistique

Hypothèses:

- Il s'agit de tester les hypothèses

$$\begin{cases} H0: \quad \beta = 0 \equiv r' = 0, & \text{il n' y a pas de lien entre } X \text{ et } Y \\ H1: \quad \beta \neq 0 \equiv r' \neq 0, & \text{il y a un lien entre } X \text{ et} Y. \end{cases}$$

- Accepter $H0$ implique que l'on conclut qu'il n'y a pas de relation linéaire entre X et Y . Ceci peut signifier que
  - La relation entre X et Y n'est pas linéaire.
  - La variation de X influe peu ou pas sur la variation de Y.
- Au contraire, rejeter $H0$ implique que l'on conclut que la variation de X influe sur la variation de Y.

Conditions d'applications

- Condition $C1$ et $C2$.
- à X donné, Y_i indépendants.
- La régression est linéaire.

Variable de décision: test de Student

- Sous l'hypothèse H_0, la statistique

$$T_n = \frac{b}{s'_e / \sum (x_i - \overline{X})^2} \rightsquigarrow T_{(n-2)}$$

- Pour une hypothèses alternative: $H1 : \beta \neq 0$ bilatéral, on rejette H_0 avec un risque α si

$$|t| \geq t_{(n-2,1-\alpha/2)} \, ,$$

où t est la réalisation de T_n et $t_{(n-2,1-\alpha/2)}$ est le fractile d'ordre $1 - \alpha/2$ de la loi Student ($T(n-2)$) à $(n-2)$ degrés de liberé.

- **Commentaire:** On rejette H_0 si $0 \notin IC(\beta)$.

Test de significativité globale du modèle: Test de Fisher où test de R^2

- Le test de Fisher s'intéresse à la significativité globale d'un modèle.
  - Il s'agit de tester les hypothèses

$$\begin{cases} H0: & \text{le modèle n'amène rien dans l'explication de Y} \\ H1: & \text{le modèle est pertinent (globalement significatif).} \end{cases}$$

- Dans le cas de régression simple, seul le paramètre β est concerné (cela correspond à $H0 : b = 0$ contre $H1 : b \neq 0$)

Test de significativité globale du modèle: Test de Fisher où test de R^2

- Sous $H0$, la distribution

$$F = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SSR}{SSE} \rightsquigarrow F(1,n-2) \text{ Fisher}(1,n-2)d.d.l.$$

- On rejette $H0$ au seuil α si $F > f_{(1,n-2)}(1-\alpha)$.
- Ce test est équivalent au test de Student précédent, il conduit à la même conclusion.

Exercice 1

- Soit un modèle lineaire simple: $Y = \theta + \beta X + \varepsilon$. On donne les informations suivantes: $n = 7$, $\overline{X} = 400$, $\overline{Y} = 60$, $\sum x_i^2 = 1400000$, $\sum y_i^2 = 26350$, $\sum x_i y_i = 184500$.
  - Estimer les coefficients du modèle.
  - Evaluer la qualite de cet ajustement.
  - Tester la significativite globale du modèle.

Exercice 1: solution

- Formules des calculs:
  - $\hat{\beta} = b = \frac{S'_{X,Y}}{S'^2_X} = \frac{\sum x_i y_i - n\overline{X}\,\overline{Y}}{\sum x_i^2 - n(\overline{X})^2}$, $\hat{\theta} = a = \overline{Y} - b\overline{X}$.
  - $R^2 = \frac{SCE}{SCT} = \frac{b^2[\sum x_i^2 - n\overline{X}^2]}{[\sum y_i^2 - n\overline{Y}^2]}$.
  - $F = (n-2)\frac{R^2}{1-R^2}$.
- On trouve: $b = 0.0589$; $a = 36.44$; $R^2 = 0.8455$; $F = 27.3618$.
- Le R^2 étant relativement élevé, environ 85%, l'ajustement effectué est de bonne qualité.
- Puisque $F_{cal} > F_{th} = F(1;5)(0.05) = 6.61$, on en conclut que le modèle est globalement bon.

Exercice 2

- Le tableau suivant donne l'âge et la tension artérielle Y de 12 femmes:

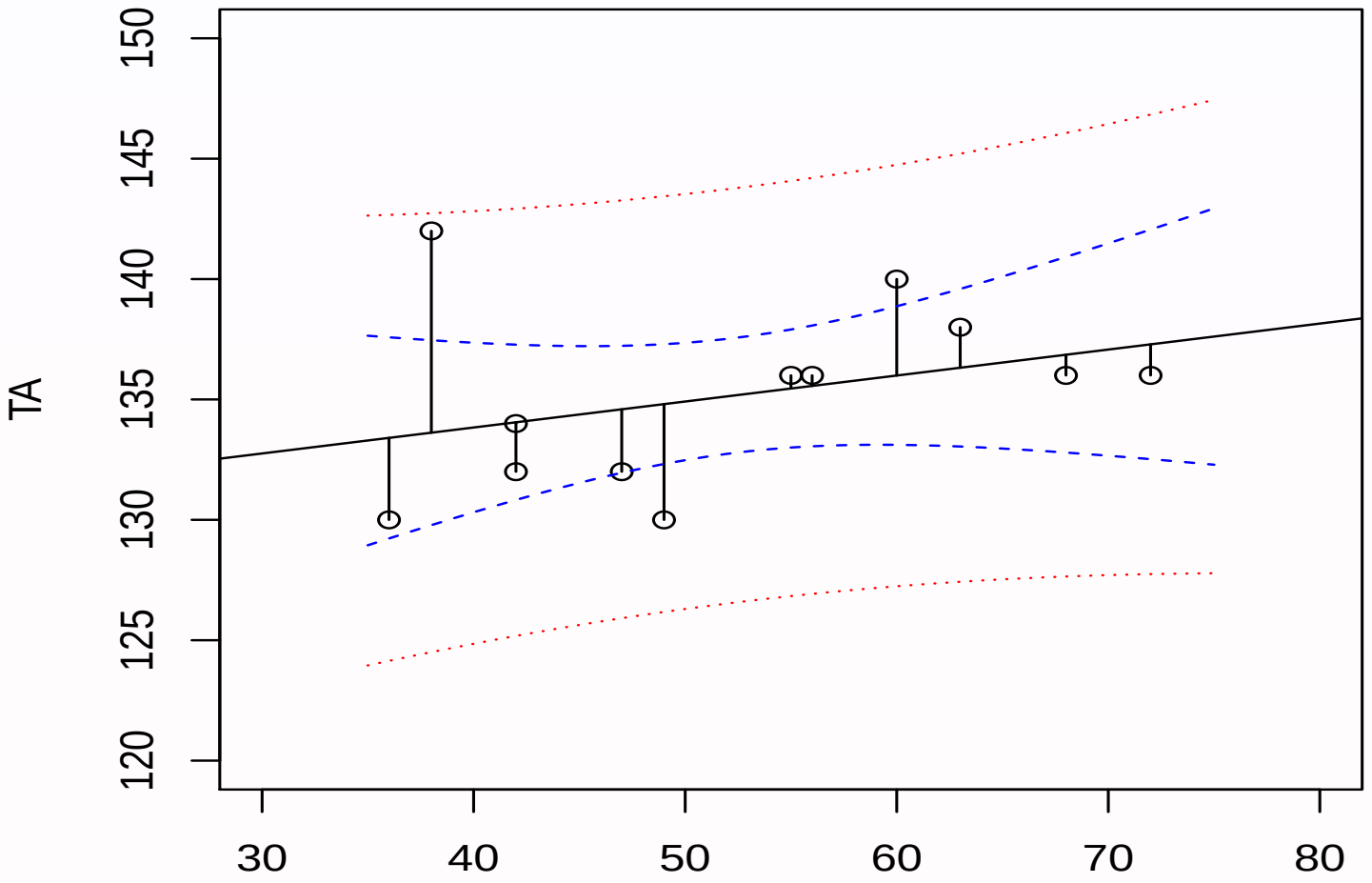
Indi	1	2	3	4	5	6	7	8	9	10	11	12
Age(X)	56	42	72	36	63	47	55	49	38	42	68	60
T-A (Y)	136	132	136	130	138	132	136	130	142	134	136	140

Exercice 2

- Donner la représentation graphique du nuages des points.
- Déterminer l'équation de la droite de régression de Y sur X.
- En déduire les valeurs estimées de $\hat{Y}$ de Y.
- Calculer les résidus et vérifier la propriété selon laquelle la moyenne des résidus est nulle.
- Calculer l'estimateur de la variance de l'erreur.
- Construire l'intervalle de confiance au niveau de confiance de 95% pour le parametre β.
- Tester la significativité de la pente. Quelle conclusion peut-on tirer ?
- Estimer la tension artérielle d'une femme âgée de 50 ans.

Exercice 2

- Nuage des points:



Exercice 2: solution

- $a = 129.53$ et $b = 0.11$.
- $\hat{y}_i$: 135.56, 134.05, 137.29, 133.40, 136.32, 134.59, 135.45, 134.81, 133.62, 134.05, 136.86,135.99.
- ε_i: 0.44, -2.05, -1.29, -3.40, 1.68, -2.59, 0.55, -4.81, 8.38, -0.05, -0.86, 4.01.
- $\sum \varepsilon_i = 0.01$, $\sum \varepsilon_i^2 = 137.64$ et $S_e'^2 = 137.64/10 = 13.76$.
- Test de Student $t_{cal} = 1.146$ et p-values $= 0.279 > 0.05 = \alpha$, donc on accepte $H0$: $\beta = 0$.

Exercice 2: solution

- Analysis of Variance Table:

  $SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{Y}})^2 = (n-1)\frac{S'_{(X,Y)}}{S'^2_X} = 18.057.$

  $SCR = \sum_{i=1}^{n} \varepsilon_i^2 = 137.61.$

  $SCT = \sum (y_i - \overline{Y})^2 = (n-1)S'^2_Y = 18.057 + 137.61.$

- $R^2 = 0.116$

- $F_{cal} = 1.3122$ et $F_{th}(1, 10) = 4.96$; p-values= $0.2787 > 0.05 = \alpha$, donc on accepte $H0$: le modèle n'amène rien dans l'explication de TA.