Compléments sur la régression linéaire simple

Anova et inférence sur les paramètres

Frédéric Bertrand et Myriam Maumy-Bertrand[1]

[1]IRMA, Université de Strasbourg
France

Master 1
2018

Ce chapitre s'appuie essentiellement sur deux livres :

1. « Analyse de régression appliquée »,
   Deuxième édition,
   de Y. Dodge et V. Rousson,
   2004, Dunod.
2. « Régression non linéaire et applications »,
   de A. Antoniadis, J. Berruyer, R. Carmona,
   1999, Economica.

Sommaire

Sommaire

1. Test et analyse de variance de la régression
2. Distribution des paramètres
   - Modèle de régression linéaire simple
   - Distribution de la pente du modèle
   - Distribution de l'ordonnée à l'origine
3. Tests et intervalles de confiance sur les paramètres
   - Test sur la pente
   - Intervalle de confiance pour la pente
   - Test sur l'ordonnée à l'origine
   - Intervalle de confiance pour l'ordonnée à l'origine
4. Distribution et intervalle de confiance pour une valeur moyenne ou une prévision
5. Exemple

- Il existe plusieurs démarches pour tester la validité de la linéarité d'une régression linéaire simple.
- Nous montrons l'équivalence de ces différents tests.
- Conséquence : Cela revient à faire **le test du coefficient de corrélation linéaire**, appelé aussi le coefficient de Bravais-Pearson.

Problème

Nous souhaitons tester l'hypothèse nulle :

$$\mathcal{H}_0 : \rho(X, Y) = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \rho(X, Y) \neq 0$$

où

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}},$$

avec

$$Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = Cov(Y, X).$$

Solution

La méthode que nous employerons ici est :

la méthode de l'ANOVA

utilisée par les logiciels de statistique.

Remarque

- ANOVA pour ANalysis Of VAriance ou encore analyse de la variance.

Remarque

Nous avons établi dans le cours précédent :

Somme des Carrés Totale = Somme des Carrés Expliquée + Somme des Carrés Résiduelle

ce qui s'écrit mathématiquement par :

$$\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y}_n)^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

À chaque somme de carrés est associé son nombre de degrés de liberté (*ddl*). Ces *ddl* sont présents dans le tableau de l'ANOVA.

Tableau de l'ANOVA

Source de variation	sc	*ddl*	cm
expliquée sc_{reg}	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_n)^2$	1	$sc_{reg}/1$
résiduelle sc_{res}	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$sc_{res}/(n-2)$
totale sc_{tot}	$\sum_{i=1}^{n}(y_i - \overline{y}_n)^2$	$n-1$	

Remarques

1. Le **coefficient de détermination**
$$R^2 = \frac{sc_{reg}}{sc_{tot}}$$
mesure le pourcentage d'explication du modèle par la régression linéaire.
2. Le rapport
$$cm_{res} = \frac{sc_{res}}{n-2}$$
est l'**estimation de la variance résiduelle**.

À partir du tableau de l'ANOVA, nous effectuons **le test de la linéarité de la régression** en calculant **la statistique de Fisher** F qui suit une loi de Fisher $F(1, n-2)$.

Cette variable aléatoire F se réalise en :

$$F_{obs} = \frac{SC_{reg}/1}{SC_{res}/(n-2)} = (n-2)\frac{SC_{reg}}{SC_{res}}.$$

Décision

Si
$$F_{obs} \geqslant F_{1-\alpha}(1, n-2),$$
alors nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$ au risque α, c'est-à-dire qu'il existe une liaison linéaire significative entre X et Y.
Si
$$F_{obs} < F_{1-\alpha}(1, n-2),$$
alors nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent de l'accepter, c'est-à-dire nous concluons qu'il n'existe pas de liaison linéaire entre X et Y.

Remarque

En effet, si l'hypothèse nulle $\mathcal{H}_0$ est vérifiée alors cela implique que $\rho(X, Y) = 0$ c'est-à-dire $Cov(X, Y) = 0$. Donc il n'existe aucune liaison linéaire entre X et Y.

Sommaire

Sommaire

Modélisation

Le modèle de régression linéaire simple est

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

où les ε_i sont des variables aléatoires inobservables, appelées **les erreurs**.

Conséquence : Les variables Y_i sont aléatoires.

Première hypothèse : $\mathbb{E}[\varepsilon_i] = 0$.
Conséquence : $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$.

D'autre part, nous avons :

$$Var[Y_i] = Var[\varepsilon_i].$$

Les trois hypothèses indispensables pour construire la théorie :

1. Les variables aléatoires ε_i sont indépendantes.
2. Les variables aléatoires ε_i sont normalement distribuées.
3. La variance des variables aléatoires ε_i est égale à σ^2 (inconnue) ne dépendant pas de x_i.
   Nous avons donc pour tout $i = 1, \ldots, n$:

$$Var[\varepsilon_i] = Var[Y_i] = \sigma^2.$$

Résumons-nous

Ces trois hypothèses sont équivalentes à :

les variables aléatoires ε_i sont indépendantes et identiquement distribuées selon une loi normale de moyenne nulle et de variance σ^2.

Nous notons :

$$\varepsilon_i \quad \text{i.i.d.} \sim \mathcal{N}(0; \sigma^2).$$

Conséquences importantes :

1. La normalité des variables aléatoires ε_i implique la normalité des variables aléatoires Y_i.
2. L'indépendance des variables aléatoires ε_i implique l'indépendance des variables aléatoires Y_i.
   En effet, nous montrons en calculant que :

$$\begin{aligned} Cov[Y_i, Y_j] &= Cov[\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j] \\ &= Cov[\varepsilon_i, \varepsilon_j] \\ &= 0. \end{aligned}$$

Sommaire

Nous avons :

$$\widehat{\beta}_1 = \frac{\sum(x_i - \overline{x}_n)Y_i}{\sum(x_i - \overline{x}_n)^2},$$

où

$$\overline{x}_n = \frac{\sum x_i}{n}.$$

Il en résulte que :

- $\widehat{\beta}_1$ **est une variable aléatoire** car $\widehat{\beta}_1$ dépend des variables Y_i qui sont des variables aléatoires.
- $\widehat{\beta}_1$ **est une fonction linéaire des variables aléatoires** Y_i.
- Comme les variables aléatoires Y_i par hypothèse sont normalement distribuées, alors $\widehat{\beta}_1$ **est normalement distribuée**.

Il reste donc à calculer ces deux valeurs pour caractériser l'estimateur $\widehat{\beta}_1$:

1. $\mathbb{E}\left[\widehat{\beta}_1\right]$
2. $Var\left[\widehat{\beta}_1\right]$.

Par calcul, nous montrons que :

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_1\right] &= \mathbb{E}\left[\frac{\sum(x_i - \overline{x}_n)Y_i}{\sum(x_i - \overline{x}_n)^2}\right] \\
&= \frac{\sum(x_i - \overline{x}_n)\mathbb{E}[Y_i]}{\sum(x_i - \overline{x}_n)^2} \\
&= \frac{\sum(x_i - \overline{x}_n)(\beta_0 + \beta_1 x_i)}{\sum(x_i - \overline{x}_n)^2} \\
&= \frac{\beta_0 \sum(x_i - \overline{x}_n) + \beta_1 \sum(x_i - \overline{x}_n)x_i}{\sum(x_i - \overline{x}_n)^2} \\
&= \frac{0 + \beta_1 \sum(x_i - \overline{x}_n)x_i}{\sum(x_i - \overline{x}_n)^2}.
\end{aligned}$$

En effet, nous montrons que :

$$\sum(x_i - \overline{x}_n) = 0.$$

De plus, comme nous avons :

$$\sum(x_i - \overline{x}_n)^2 = \sum(x_i - \overline{x}_n)x_i$$

alors nous obtenons :

$$\mathbb{E}\left[\widehat{\beta}_1\right] = \beta_1.$$

Donc la variable aléatoire $\widehat{\beta}_1$ est **un estimateur sans biais** du coefficient β_1.

D'autre part, nous calculons la variance de $\widehat{\beta}_1$ ainsi :

$$\begin{aligned}
Var\left[\widehat{\beta}_1\right] &= Var\left[\frac{\sum(x_i-\overline{x}_n)Y_i}{\sum(x_i-\overline{x}_n)^2}\right]\\
&= \frac{\sum(x_i-\overline{x}_n)^2 Var[Y_i]}{\left(\sum(x_i-\overline{x}_n)^2\right)^2}\\
&= \frac{\sum(x_i-\overline{x}_n)^2\sigma^2}{\left(\sum(x_i-\overline{x}_n)^2\right)^2}\\
&= \frac{\sigma^2}{\sum(x_i-\overline{x}_n)^2},
\end{aligned}$$

ce qui achève la caractérisation de $\widehat{\beta}_1$.

Sommaire

Nous avons :

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1\overline{x}_n$$

où

$$\overline{x}_n = \frac{\sum x_i}{n} \quad \text{et} \quad \overline{Y}_n = \frac{\sum Y_i}{n}.$$

- $\widehat{\beta}_0$ **est une variable aléatoire** car $\widehat{\beta}_0$ dépend de $\widehat{\beta}_1$ qui est une variable aléatoire.
- $\widehat{\beta}_0$ **est une fonction linéaire de** $\widehat{\beta}_1$.
- Comme $\widehat{\beta}_1$ est normalement distribuée, alors $\widehat{\beta}_0$ **est normalement distribuée.**

Il reste donc à calculer ces deux valeurs pour caractériser l'estimateur $\widehat{\beta}_0$:

1. $\mathbb{E}\left[\widehat{\beta}_0\right]$
2. $Var\left[\widehat{\beta}_0\right]$.

Par calcul, nous montrons que :

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_0\right] &= \mathbb{E}\left[\overline{Y}_n - \widehat{\beta}_1 \overline{x}_n\right]\\ &= \mathbb{E}\left[\overline{Y}_n\right] - \overline{x}_n \mathbb{E}\left[\widehat{\beta}_1\right]\\ &= \mathbb{E}\left[\overline{Y}_n\right] - \overline{x}_n \beta_1,\end{aligned}$$

car nous venons de démontrer que $\widehat{\beta}_1$ est un estimateur sans biais du coefficient β_1.

Il reste à calculer la valeur :

$$\mathbb{E}\left[\overline{Y}_n\right].$$

Or nous avons :

$$\begin{aligned}\mathbb{E}\left[\overline{Y}_n\right] &= \mathbb{E}\left[\frac{\sum Y_i}{n}\right]\\ &= \frac{\sum \mathbb{E}[Y_i]}{n}\\ &= \frac{\sum(\beta_0 + \beta_1 x_i)}{n}\\ &= \frac{n\beta_0 + \beta_1 \sum x_i}{n}\\ &= \beta_0 + \overline{x}_n \beta_1.\end{aligned}$$

Nous obtenons donc :

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_0\right] &= \mathbb{E}\left[\overline{Y}_n\right] - \overline{x}_n \beta_1\\ &= (\beta_0 + \overline{x}_n \beta_1) - \overline{x}_n \beta_1\\ &= \beta_0.\end{aligned}$$

Donc la variable aléatoire $\widehat{\beta}_0$ est **un estimateur sans biais** du coefficient β_0.

D'autre part, nous calculons la variance de $\widehat{\beta}_0$ ainsi :

$$\begin{aligned}Var\left[\widehat{\beta}_0\right] &= Var\left[\overline{Y}_n - \widehat{\beta}_1 \overline{x}_n\right]\\ &= Var\left[\overline{Y}_n\right] + \overline{x}_n^2\, Var\left[\widehat{\beta}_1\right] - 2\,\overline{x}_n\, Cov\left[\overline{Y}_n, \widehat{\beta}_1\right].\end{aligned}$$

Il reste donc à calculer la valeur :

$$Cov\left[\overline{Y}_n, \widehat{\beta}_1\right].$$

Par les calculs, nous montrons que :

$$\begin{aligned} Cov\left[\overline{Y}_n, \widehat{\beta}_1\right] &= Cov\left[\frac{\sum Y_i}{n}, \frac{\sum(x_j-\overline{x}_n)Y_j}{\sum(x_i-\overline{x}_n)^2}\right] \\ &= \frac{\sum_i\sum_j(x_j-\overline{x}_n)Cov[Y_i,Y_j]}{n\sum(x_i-\overline{x}_n)^2} \\ &= \frac{\sum_i(x_i-\overline{x}_n)Var[Y_i]}{n\sum(x_i-\overline{x}_n)^2} \\ &= \frac{\sigma^2\sum_i(x_i-\overline{x}_n)}{n\sum(x_i-\overline{x}_n)^2} \\ &= 0. \end{aligned}$$

Comme nous avons que

$$Var\left[\overline{Y}_n\right] = \frac{\sigma^2}{n},$$

nous obtenons, alors :

$$\begin{aligned} Var\left[\widehat{\beta}_0\right] &= Var\left[\overline{Y}_n\right] + \overline{x}_n^2 Var\left[\widehat{\beta}_1\right] \\ &= \frac{\sigma^2}{n} + \frac{\overline{x}_n^2\sigma^2}{\sum(x_i-\overline{x}_n)^2} \\ &= \frac{\sigma^2\left(\sum(x_i-\overline{x}_n)^2 + n\overline{x}_n^2\right)}{n\sum(x_i-\overline{x}_n)^2}. \end{aligned}$$

En rappelant que :

$$\sum(x_i-\overline{x}_n)^2 = \sum x_i^2 - n\overline{x}_n^2,$$

nous avons finalement :

$$Var\left[\widehat{\beta}_0\right] = \frac{\sigma^2\sum x_i^2}{n\sum(x_i-\overline{x}_n)^2}.$$

Sommaire

Sommaire

Nous rappelons que :

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1; \sigma^2(\widehat{\beta}_1))$$

où

$$\sigma^2(\widehat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \overline{x}_n)^2}.$$

Nous obtenons alors :

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma(\widehat{\beta}_1)} \sim \mathcal{N}(0; 1).$$

Problème

Nous ne connaissons pas le paramètre σ^2, c'est-à-dire la variance des variables aléatoires ε_i.

Que pouvons-nous faire alors pour résoudre ce problème ?

Solution

Estimer ce paramètre !

- Nous estimons d'abord σ^2 par CM_{res} l'estimateur sans biais de σ^2 :

$$CM_{res} = \frac{||\varepsilon||^2}{n-2} = \frac{\sum(Y_i - \widehat{Y}_i)^2}{n-2}.$$

- Nous estimons ensuite $\sigma^2(\widehat{\beta}_1)$ par :

$$s^2(\widehat{\beta}_1) = \frac{CM_{res}}{\sum(x_i - \overline{x}_n)^2}.$$

- Nous montrons alors que :

$$\left(\widehat{\beta}_1 - \beta_1\right) / s(\widehat{\beta}_1) \sim T_{n-2},$$

où T_{n-2} désigne une v.a. de Student avec $(n-2)$ *ddl*.

Nous souhaitons tester l'hypothèse nulle :

$$\mathcal{H}_0 : \beta_1 = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \beta_1 \neq 0.$$

Nous utilisons alors la statistique de Student suivante :

$$t_{obs} = \frac{\widehat{\beta}_1}{s(\widehat{\beta}_1)}$$

pour décider de l'acceptation ou du rejet de $\mathcal{H}_0$.

Décision

Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc d'accepter l'hypothèse alternative $\mathcal{H}_1$ au seuil de signification α si

$$|t_{obs}| \geqslant t_{n-2;1-\alpha/2}$$

où la valeur critique $t_{n-2;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-2)$ *ddl*.

Dans ce cas, nous disons que la relation linéaire entre X et Y est significative au seuil α.

Décision - Suite et fin

Nous décidons d'accepter l'hypothèse nulle $\mathcal{H}_0$ au seuil de signification α si

$$|t_{obs}| < t_{n-2;1-\alpha/2}$$

où la valeur $t_{n-2;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-2)$ *ddl*.

Dans ce cas, Y ne dépend pas linéairement de X. Le modèle devient alors :

$$Y_i = \beta_0 + \varepsilon_i$$

Le modèle proposé $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ est inadéquat. Nous testons alors un nouveau modèle.

Sommaire

IC pour β_1

Un intervalle de confiance au niveau $(1-\alpha)$ pour le coefficient inconnu β_1 est défini par

$$\left]\widehat{\beta}_1 - t_{n-2;1-\alpha/2} \times s(\widehat{\beta}_1)\,;\,\widehat{\beta}_1 + t_{n-2;1-\alpha/2} \times s(\widehat{\beta}_1)\right[.$$

Cet intervalle de confiance est construit pour que, $(1-\alpha)$% de ses réalisations contiennent la vraie valeur inconnue du coefficient β_1.

Sommaire

Nous rappelons que :

$$\widehat{\beta}_0 \sim \mathcal{N}(\beta_0; \sigma^2(\widehat{\beta}_0))$$

où

$$\sigma^2(\widehat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \overline{x}_n)^2}.$$

Nous obtenons alors :

$$\frac{\widehat{\beta}_0 - \beta_0}{\sigma(\widehat{\beta}_0)} \sim \mathcal{N}(0;1).$$

Problème

Nous ne connaissons pas le paramètre σ^2, c'est-à-dire la variance des variables aléatoires ε_i.

Que pouvons-nous faire alors pour résoudre ce problème ?

Solution

Estimer ce paramètre !

- Nous estimons d'abord σ^2 par CM_{res} l'estimateur sans biais de σ^2 :

$$CM_{res} = \frac{||\varepsilon||^2}{n-2} = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2}.$$

- Nous estimons ensuite $\sigma^2(\widehat{\beta}_0)$ par :

$$s^2(\widehat{\beta}_0) = \frac{CM_{res} \sum x_i^2}{n \sum (x_i - \overline{x}_n)^2}.$$

- Nous montrons alors que :

$$\left(\widehat{\beta}_0 - \beta_0\right) / s(\widehat{\beta}_0) \sim T_{n-2}$$

où T_{n-2} désigne une v.a. de Student avec $(n-2)$ *ddl*.

Nous souhaitons tester l'hypothèse nulle

$$\mathcal{H}_0 : \beta_0 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_0 \neq 0.$$

Nous utilisons la statistique de Student suivante :

$$t_{obs} = \frac{\widehat{\beta}_0}{s(\widehat{\beta}_0)}$$

pour décider de l'acceptation ou du rejet de l'hypothèse nulle $\mathcal{H}_0$.

Décision

Nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$ au seuil de signification α si :

$$|t_{obs}| \geqslant t_{n-2;1-\alpha/2}$$

où la valeur critique $t_{n-2;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-2)$ ddl.

Dans ce cas, le coefficient β_0 du modèle est dit significatif au seuil α.

Décision - Suite et fin

Nous décidons de ne pas refuser et donc d'accepter l'hypothèse nulle $\mathcal{H}_0$ au seuil de signification α si

$$|t_{obs}| < t_{n-2;1-\alpha/2}$$

où la valeur critique $t_{n-2;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-2)$ *ddl*.

Dans ce cas, l'ordonnée de la droite de régression passe par l'origine :

$$Y_i = \beta_1 x_i + \varepsilon_i.$$

Sommaire

IC pour β_0

Un intervalle de confiance au niveau $(1-\alpha)$ pour le coefficient inconnu β_0 est défini par :

$$\left]\widehat{\beta}_0 - t_{n-2;1-\alpha/2} \times s(\widehat{\beta}_0)\,;\, \widehat{\beta}_0 + t_{n-2;1-\alpha/2} \times s(\widehat{\beta}_0)\right[.$$

Cet intervalle de confiance est construit pour que, $(1-\alpha)$% de ses réalisations contiennent la vraie valeur inconnue du coefficient β_0.

Sommaire

Nous allons voir comment trouver un intervalle de confiance pour la valeur moyenne

$$\mu_Y(x) = \beta_0 + \beta_1 x,$$

c'est-à-dire pour l'ordonnée du point d'abscisse x se trouvant sur la droite de régression.

L'estimateur de $\beta_0 + \beta_1 x$ est donné par la droite des moindre carrés :

$$\widehat{Y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

où

- $\widehat{Y}(x) \sim \mathcal{N}\left(\beta_0 + \beta_1 x; \sigma^2(\widehat{Y}(x))\right)$

où

$$\sigma^2\left(\widehat{Y}(x)\right) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right).$$

Ce qui peut s'écrire aussi :

- $\dfrac{\widehat{Y}(x) - \mu_Y(x)}{\sigma\left(\widehat{Y}(x)\right)} \sim \mathcal{N}(0;1).$

Problème

La variance σ^2 est inconnue.

Solution

- Nous estimons d'abord σ^2 par l'estimateur CM_{res}.
- Nous estimons ensuite $\sigma^2(\widehat{Y}(x))$ par :

$$s^2(\widehat{Y}(x)) = CM_{res}\left(\frac{1}{n} + \frac{(x-\overline{x}_n)^2}{\sum(x_i-\overline{x}_n)^2}\right).$$

- Ainsi nous obtenons :

$$\frac{\widehat{Y}(x) - \mu_Y(x)}{s\left(\widehat{Y}(x)\right)} \sim T_{n-2}.$$

Intervalle de confiance de la valeur moyenne

Il est possible de construire une intervalle de confiance de la valeur moyenne de Y sachant que $X = x_0$. L'estimation ponctuelle pour cette valeur de x_0 est alors égale à $\widehat{y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.
L'intervalle de confiance de la valeur moyenne prise par la variable Y lorsque $X = x_0$ est égal à

$$\left]\widehat{y}(x_0) - t_{n-2;1-\alpha/2} \times s(\widehat{y}(x_0))\,;\,\widehat{y}(x_0) + t_{n-2;1-\alpha/2} \times s(\widehat{y}(x_0))\right[.$$

Cet intervalle de confiance est construit pour que, $(1-\alpha)$% de ses réalisations contiennent la vraie valeur moyenne inconnue $\mu_Y(x_0)$.

Intervalle de prédiction d'une valeur individuelle

L'ajustement affine peut servir à prévoir une valeur attendue pour la variable Y quand nous fixons $X = x_0$. L'estimation ponctuelle de cette valeur est alors égale à $\widehat{y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$.

Un intervalle de prévision au niveau $(1-\alpha)$ pour la variable Y sachant que $X = x_0$ est défini par :

$$\left]\widehat{y}(x_0) - t_{n-2;1-\alpha/2}\sqrt{\left(cm_{res} + s^2(\widehat{y}(x_0))\right)};\right.$$
$$\left.\widehat{y}(x_0) + t_{n-2;1-\alpha/2}\sqrt{\left(cm_{res} + s^2(\widehat{y}(x_0))\right)}\right[.$$

Intervalle de prédiction d'une valeur individuelle (suite)

Cet intervalle de prévision est construit pour que $(1-\alpha)\%$ de ses réalisations contiennent la vraie valeur individuelle inconnue $Y(x_0)$.

Précaution d'emploi

L'utilisation d'une valeur estimée $\widehat{y}(x_0)$ n'est justifiée que si R^2 est proche de 1.

Sommaire

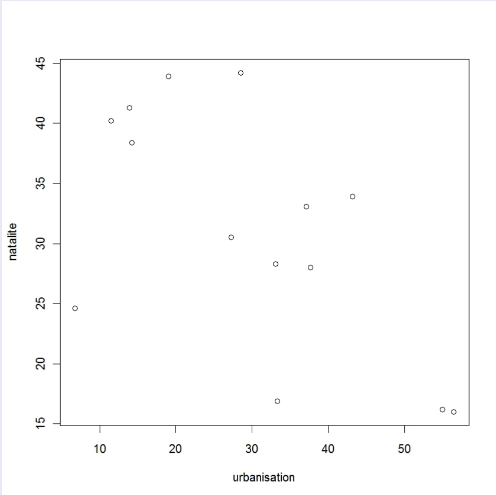
Exemple : le tableau de données. D'après Birkes et Dodge (1993)

Pays	Taux d'urbanisation x_i	Taux de natalité y_i
Canada	55,0	16,2
Costa Rica	27,3	30,5
Cuba	33,3	16,9
E.U.	56,5	16,0
El Salvador	11,5	40,2
Guatemala	14,2	38,4
Haïti	13,9	41,3

Suite des données

Pays	Taux d'urbanisation x_i	Taux de natalité y_i
Honduras	19,0	43,9
Jamaïque	33,1	28,3
Mexique	43,2	33,9
Nicaragua	28,5	44,2
Trinité-et-Tobago	6,8	24,6
Panama	37,7	28,0
Rép. Dom.	37,1	33,1

Nuage de points



Analyse : calcul du coefficient de corrélation linéaire

Nous souhaitons modéliser la relation entre le taux de natalité et le taux d'urbanisation.
La première question à se poser est : « existe-t-il une relation linéaire entre les deux variables ? »
Pour y répondre, calculons le coefficient de corrélation linéaire de Bravais-Pearson à l'aide de R.

```
> cor(natalite,urbanisation)
[1] -0.6211854
```

Comment interprétons-nous cette valeur ? Il semblerait qu'il puisse exister une relation linéaire entre les deux variables. Il reste donc à réaliser le test du coefficient de corrélation linéaire.

Suite de l'analyse : test de corrélation linéaire

Mais pour cela, il faut savoir si le couple (X, Y) suit une loi normale bivariée. Utilisons R.

```
> exemple<-data.frame(urbanisation,natalite)
> transpose<-t(exemple)
> mshapiro.test(transpose)
Shapiro-Wilk normality test
data: Z
W = 0.927, p-value = 0.2771
```

La *p*-valeur (*p*-value = 0,2771) étant supérieure à $\alpha = 5\%$, nous décidons de ne pas rejeter et donc d'accepter l'hypothèse nulle $\mathcal{H}_0$ au seuil $\alpha = 5\%$. La fonction `mshapiro.test()` nécessite l'installation du package `mvnormtest`.

Suite de l'analyse : test de corrélation linéaire

Maintenant que l'hypothèse fondamentale est vérifiée, nous pouvons réaliser le test de corrélation linéaire.

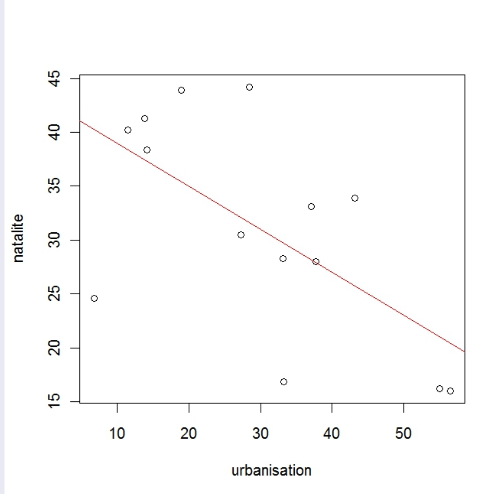
```
> cor.test(urbanisation,natalite)
Pearson's product-moment correlation
data: urbanisation and natalite
t = -2.7459, df = 12, p-value = 0.01774
alternative hypothesis: true correlation is
not equal to 0
95 percent confidence interval:
-0.8662568 -0.1351496
sample estimates:
cor
-0.6211854
```

Suite et fin de l'analyse : test de corrélation linéaire

La *p*-valeur (*p*-value = 0,01774) étant inférieure à $\alpha = 5\%$, nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc d'accepter l'hypothèse alternative $\mathcal{H}_1$ au seuil de signification $\alpha = 5\%$. Il existe donc une relation linéaire entre les deux variables. Maintenant, déterminons les coefficients de la droite les moindres carrés avec R et traçons-la.

```
> modele<-lm(natalite~urbanisation)
> coef(modele)
(Intercept) urbanisation
42.9905457 -0.3988675
> abline(coef(modele),col="red")
```


Nuage des points et droite des MCO



Calcul des résidus

Pour réaliser les tests sur la pente et sur l'ordonnée, il faut vérifier la normalité des résidus. Nous allons les calculer avec R.

```
> residus<-residuals(modele)
```

et les placer dans le tableau des données.

Tableau de données avec résidus

Pays	Taux d'urbanisation x_i	Taux de natalité y_i	Valeurs estimées $\hat{y}_i$	Résidus e_i
Canada	55,0	16,2	21,05	−4,85
Costa Rica	27,3	30,5	32,10	−1,60
Cuba	33,3	16,9	29,71	−12,81
E.U.	56,5	16,0	20,45	−4,45
El Salvador	11,5	40,2	38,40	1,80
Guatemala	14,2	38,4	37,33	1,07
Haïti	13,9	41,3	37,45	3,85

Suite des données avec résidus

Pays	Taux d'urbanisation x_i	Taux de natalité y_i	Valeurs estimées $\hat{y}_i$	Résidus e_i
Honduras	$19,0$	$43,9$	$35,41$	$8,49$
Jamaïque	$33,1$	$28,3$	$29,79$	$-1,49$
Mexique	$43,2$	$33,9$	$25,76$	$8,14$
Nicaragua	$28,5$	$44,2$	$31,62$	$12,58$
Trinité-et-Tobago	$6,8$	$24,6$	$40,28$	$-15,68$
Panama	$37,7$	$28,0$	$27,95$	$0,05$
Rép. Dom.	$37,1$	$33,1$	$28,19$	$4,91$

Normalité des résidus

Réalisons donc le test de normalité, le test de Shapiro-Wilk avec `R`.

```
> shapiro.test(residus)
Shapiro-Wilk normality test
data: residus W = 0.9635, p-value = 0.7797
```

La *p*-valeur (*p*-value = 0,7797) étant supérieure à $\alpha = 5\%$, nous décidons de ne pas rejeter et donc d'accepter l'hypothèse alternative $\mathcal{H}_0$ au seuil de signification $\alpha = 5\%$.

Test sur la pente β_1.

Nous testons

$$\mathcal{H}_0 : \beta_1 = 0$$

contre

$$\mathcal{H}_1 : \beta_1 \neq 0.$$

Nous calculons

$$t_{obs} = \frac{\widehat{\beta}_1}{s(\widehat{\beta}_1)} = \frac{-0,3989}{0,1453} = -2,746.$$

Or la valeur critique est égale à pour un seuil $\alpha = 0,05$:

$$t_{(12;0,975)} = 2,178813.$$

Décision

Comme

$$|t_{obs}| > t_{n-2;1-\alpha/2},$$

nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil de signification $\alpha = 5\%$.

En conclusion : La relation linéaire entre le taux de natalité et le taux d'urbanisation est significative.

IC pour β_1

Un intervalle de confiance pour le coefficient inconnu β_1 au niveau $(1-\alpha) = 0,95$ s'obtient en calculant :

$$\widehat{\beta}_1 \pm \mathrm{t}_{n-2;1-\alpha/2} \times s(\widehat{\beta}_1) = -0,3989 \pm 2,178813 \times 0,1453.$$

Nous avons donc après simplification et approximation :

$$]-0,716; -0,082[$$

qui contient la vraie valeur du coefficient inconnu β_1 avec une probabilité de $0,95$. Nous remarquons que 0 n'est pas compris dans cet intervalle.

Test sur l'ordonnée β_0

$$\mathcal{H}_0 : \beta_0 = 0$$

contre

$$\mathcal{H}_1 : \beta_0 \neq 0.$$

Nous calculons

$$t_{obs} = \frac{\widehat{\beta}_0}{s(\widehat{\beta}_0)} = \frac{42,9905}{4,8454} = 8,872.$$

Or la valeur critique est égale à pour un seuil $\alpha = 0,05$:

$$t_{0,975;12} = 2,178813.$$

Décision

Comme

$$|t_{obs}| > t_{n-2;1-\alpha/2},$$

nous décidons de refuser l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$.

En conclusion : La droite de régression ne passe pas par l'origine.

IC pour β_0

Un intervalle de confiance pour le coefficient inconnu β_0 au niveau $(1-\alpha) = 0,95$ s'obtient en calculant :

$$\widehat{\beta}_0 \pm t_{n-2;1-\alpha/2,} \times s(\widehat{\beta}_0) = 42,9905 \pm 2,178813 \times 4,8454.$$

Nous avons donc après simplification et approximation :

$$]32,433; 53,548[$$

qui contient la vraie valeur du coefficient inconnu β_0 avec une probabilité de $0,95$. Nous remarquons que 0 n'est pas compris dans l'intervalle.