Statistique Descriptive
Chapitre V : Statistique descriptive pour deux variables quantitatives

Anas KNEFATI

Université Rennes 2

Plan

Plan

Données et objectif

Données

- X : Variable quantitative d'observations : $x_1, ..., x_n$
- Y : Variable quantitative d'observations : $y_1, ..., y_n$

Objectif

- Expliquer Y en fonction de X.
- Établir un indicateur de liaison entre X et Y

- X : C'est la variable explicative
- Y : C'est la variable à expliquer

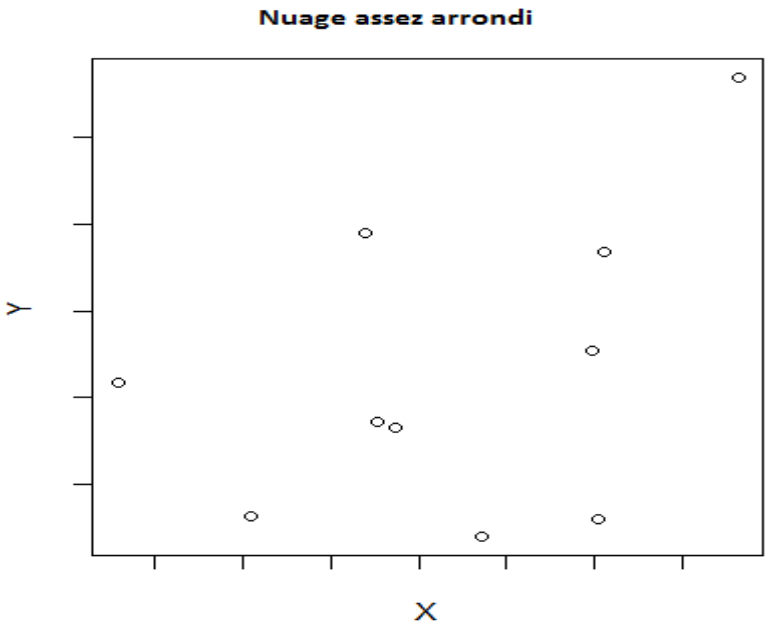
Exemple

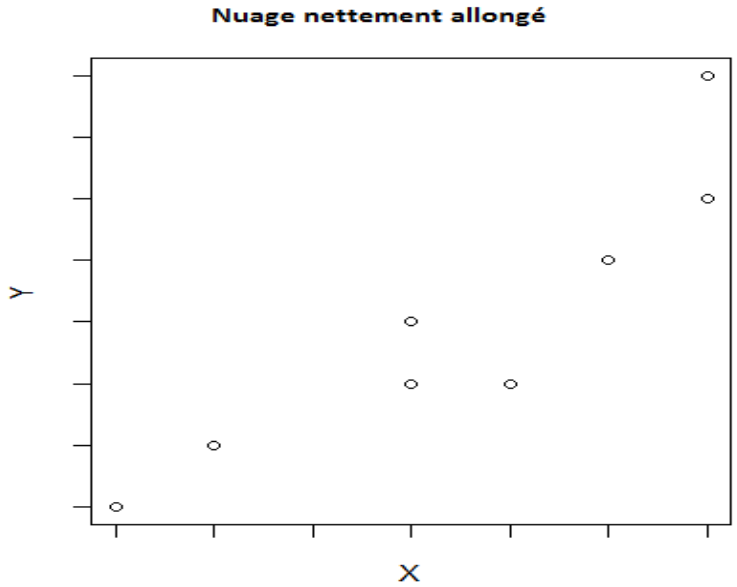
Année	Prix moyen annuel du gazole (1 Litre) X	Prix moyen annuel de la "baguette de pains" (1 kg) Y
1992	0.54	2.15
1993	0.56	2.23
1994	0.60	2.30
1995	0.59	2.34
1996	0.66	2.39
1997	0.68	2.42
1998	0.64	2.45
1999	0.69	2.50
2000	0.85	2.56
2001	0.80	2.63
2002	0.77	2.73
2003	0.80	2.84
2004	0.89	2.95
2005	1.03	3.00
2006	1.08	3.07
2007	1.10	3.18
2008	1.28	3.32
2009	1.01	3.35
2010	1.16	3.35
2011	1.34	3.42
2012	1.41	3.46
2013	1.36	3.47
2014	1.30	3.48
2015	1.17	3.46

Plan

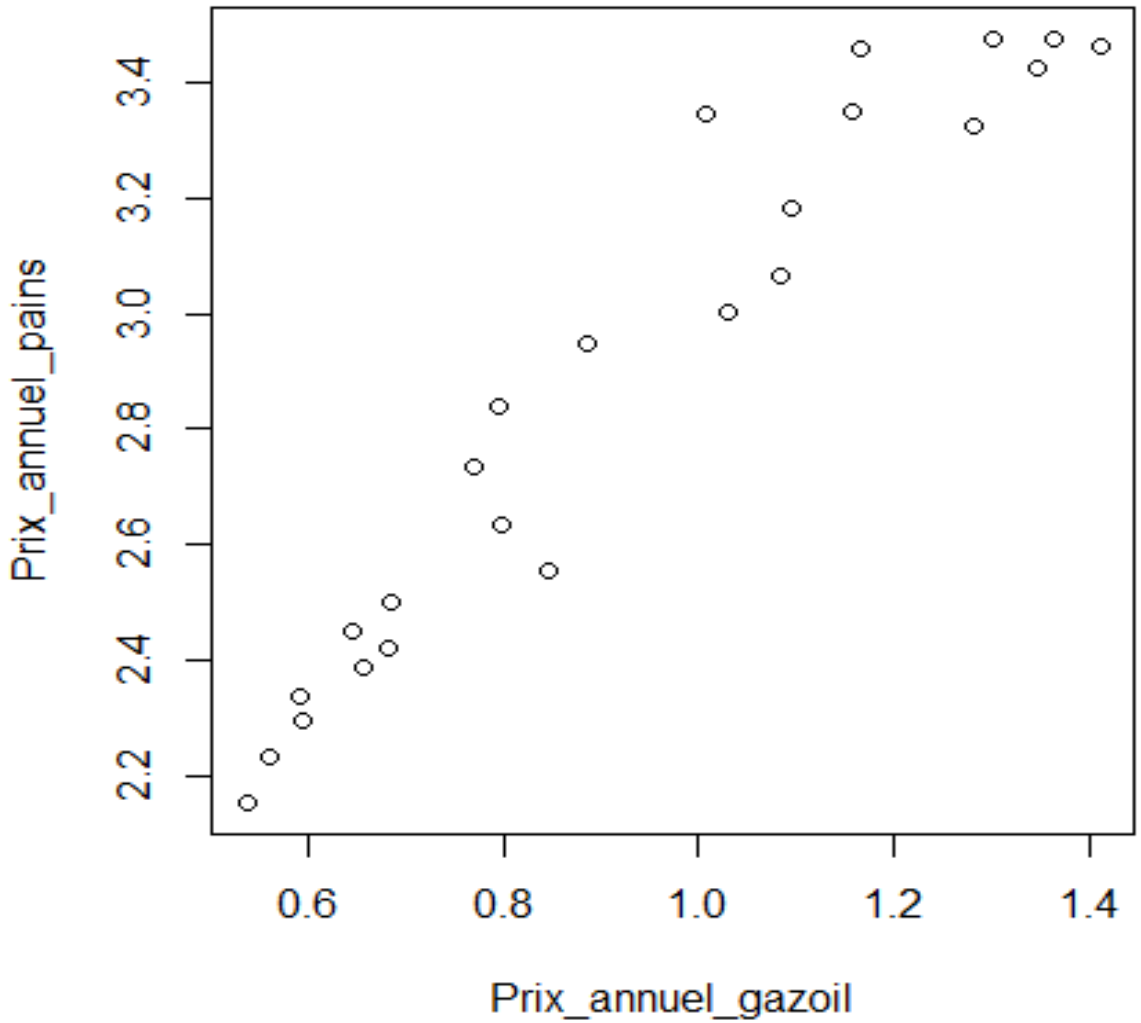
Représentation graphique : Nuage de points

- On représente les données dans un repère du plan en positionnant les points de coordonnées (x_i, y_i).
- L'ensemble de ces points donne, en général, une idée assez bonne de la variation conjointe des deux variables
  - Si le nuage est nettement allongé $\rightarrow$ Les évolutions de X et Y sont très liées $\rightarrow$ Les variables sont très liées
  - Lorsque le nuage est assez arrondi $\rightarrow$ Pas de relation nette entre les évolutions de X et Y $\rightarrow$ les variables sont peu liées.





Exemple


Prix_annuel_pains
2.2
2.4
2.6
2.8
3.0
3.2
3.4
0.6
0.8
1.0
1.2
1.4
Prix_annuel_gazoil

Plan

Ajustement linéaire (Régression linéaire)

Fonction d'ajustement

- On cherche une fonction réelle f tq $f(x_i) \approx y_i$
- Graphiquement, cela revient à cherche la courbe d'une fonction qui passe au plus près des points du nuage.
- Le choix le plus simple : $f(x) = ax + b$
- a et b sont à estimer (Paramètres de la régression linéaire)

Critères des moindres carrés

- Afin d'estimer les paramètres de la régression linéaire on cherche à minimiser le résidu global :

  $$E(a, b) = \sum_{i=1}^{n} \rho\left(y_i - f(x_i)\right)$$

  ρ est une fonction positive, elle s'appelle la fonction perte et en générale, on utilise soit la fonction carrée soit la fonction de la valeur absolue.
- ICI, on étudie le cas de la fonction carrée. Donc en minimise en a et b :

  $$E(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Estimateur

La minimisation de $E(a, b)$ en a et b fournit la solution unique suivante :

$$\begin{aligned} \hat{a} &= \frac{\mathtt{Cov}(X, Y)}{\mathtt{Var}(X)} \\ \hat{b} &= \bar{y} - \hat{a}\bar{x} \end{aligned}$$

où

$$\begin{aligned} \mathtt{Cov}(X, Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \end{aligned}$$

s'appelle la covariance entre X et Y.

Remarques

- $\hat{y} = \hat{a}x + \hat{b}$ est l'équation de la droite de régression de Y sur X.
- Le point $(\bar{x}, \bar{y})$ est appelé le centre de gravité ou le point moyen.
- La droite de régression passe par le point moyen
- Le signe de $\hat{a}$ est celui de $\mathtt{Cov}(X, Y)$
- Valeurs prédites : $\hat{y}_i = \hat{a}x_i + \hat{b}$
- Résidus : $\hat{e}_i = y_i - \hat{y}_i$. Ils sont de moyenne nulle et de variance

$$\frac{1}{n}E(\hat{a}, \hat{b}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Exemple

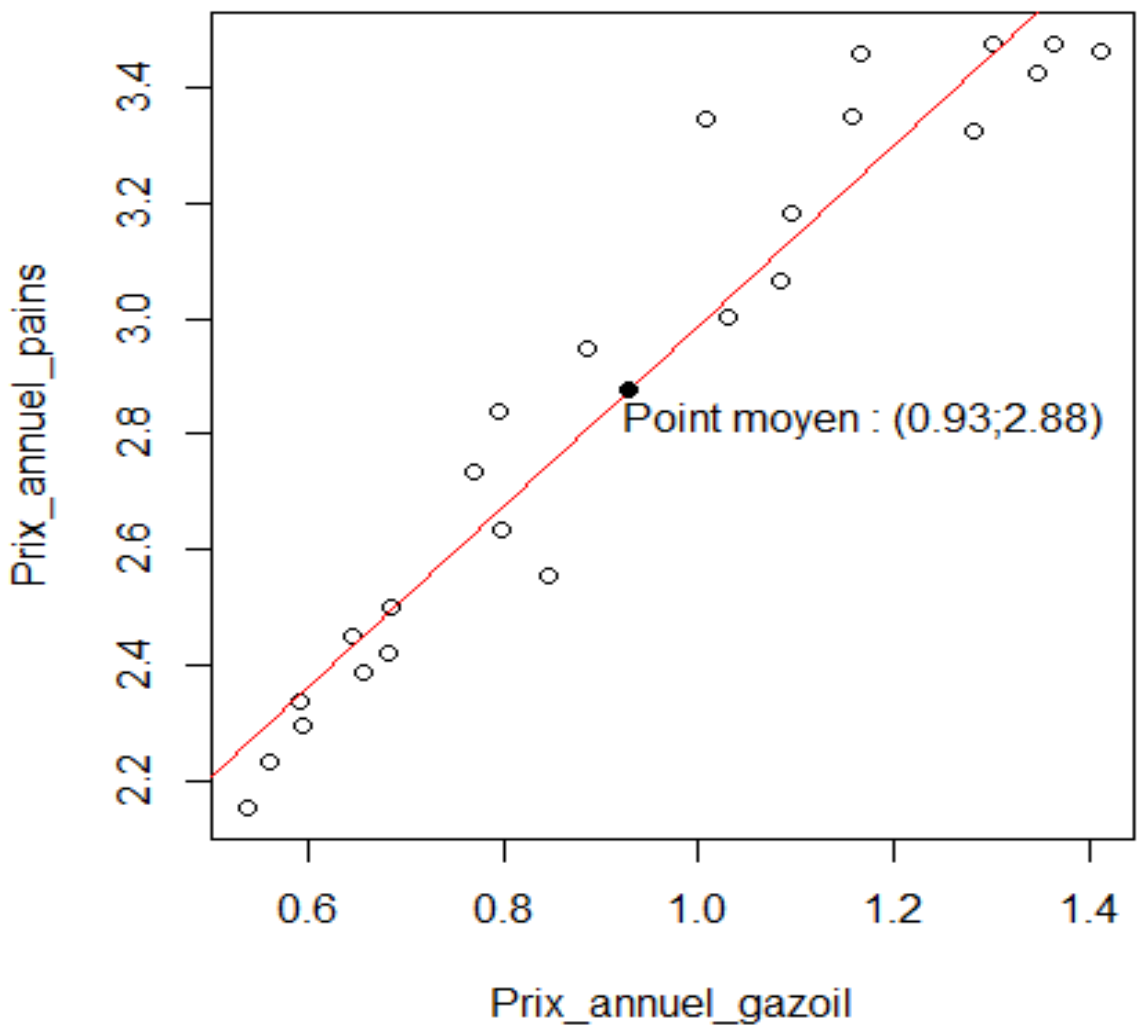
	X	Y	XY	X^2	Y^2
1992	0.54	2.15	1.16	0.29	4.62
1993	0.56	2.23	1.25	0.31	4.97
1994	0.60	2.30	1.38	0.36	5.29
1995	0.59	2.34	1.38	0.35	5.48
1996	0.66	2.39	1.58	0.44	5.71
1997	0.68	2.42	1.65	0.46	5.86
1998	0.64	2.45	1.57	0.41	6.00
1999	0.69	2.50	1.72	0.48	6.25
2000	0.85	2.56	2.18	0.72	6.55
2001	0.80	2.63	2.10	0.64	6.92
2002	0.77	2.73	2.10	0.59	7.45
2003	0.80	2.84	2.27	0.64	8.07
2004	0.89	2.95	2.63	0.79	8.70
2005	1.03	3.00	3.09	1.06	9.00
2006	1.08	3.07	3.32	1.17	9.42
2007	1.10	3.18	3.50	1.21	10.11
2008	1.28	3.32	4.25	1.64	11.02
2009	1.01	3.35	3.38	1.02	11.22
2010	1.16	3.35	3.89	1.35	11.22
2011	1.34	3.42	4.58	1.80	11.70
2012	1.41	3.46	4.88	1.99	11.97
2013	1.36	3.47	4.72	1.85	12.04
2014	1.30	3.48	4.52	1.69	12.11
2015	1.17	3.46	4.05	1.37	11.97
Total	22.31	69.05	67.15	22.63	203.65

$$\bar{x} = \frac{22.31}{24} \approx 0.93 \quad \text{et} \quad \bar{y} = \frac{69.05}{24} \approx 2.88$$

$$\begin{aligned}
\mathtt{Cov}(X, Y) &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \\
&= \frac{67.15}{24} - 0.93 * 2.88 \approx 0.12 \\
\mathtt{Var}(X) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \\
&= \frac{22.63}{24} - 0.93^2 \approx 0.08 \\
\hat{a} &= \frac{\mathtt{Cov}(X, Y)}{\mathtt{Var}(X)} \\
&= \frac{0.12}{0.08} = 1.5 \\
\hat{b} &= \bar{y} - \hat{a}\bar{x} \\
&= 2.88 - 1.5 \times 0.93 = 1.485
\end{aligned}$$

Alors : $\hat{y} = 1.5x + 1.485$

Exemple


Prix_annuel_pains
2.2
2.4
2.6
2.8
3.0
3.2
3.4
0.6
0.8
1.0
1.2
1.4
Prix_annuel_gazoil
Point moyen : (0.93;2.88)

Plan

Coefficient de corrélation linéaire

Formule

- Ce coefficient, noté r_{XY} ou r, sert à mesurer l'alignement des points
- C'est le rapport entre la covariance et le produit des écarts-types :

$$r = \frac{\mathtt{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Intérprétation

- $r > 0$: Les deux variables ont tendance à varier dans le même sens
- $r < 0$: Les deux variables ont tendance à varier en sens opposé
- $r = 0$: Il n'y a pas de relation linéaire entre X et Y
- plus $|r|$ est proche de 1, plus la liaison linéaire est forte
- plus $|r|$ est proche de 0, plus la liaison linéaire est faible
- $r = 1$ ou $r = -1$ correspond à une liaison linéaire parfaite entre X et

Coefficient de corrélation linéaire - Proporiétés

- $r = \text{Cov}(\frac{X-\bar{x}}{\sigma_X}, \frac{Y-\bar{y}}{\sigma_Y})$
  La variable $z = \frac{X-\bar{x}}{\sigma_X}$ (ou $\frac{Y-\bar{y}}{\sigma_Y}$)) est **la variable centrée réduite associée à** X (ou Y)
  - Z est centrée car sa moyenne est nulle
  - Z est réduite car sa variance égale à un
- r_{XY} est indépendant des unités de mesure de X et Y
- r_{XY} est symétrique : $r_{XY} = r_{YX}$
- $-1 \leq r_{XY} \leq 1$

Plan

Décomposition de la variance de Y (Analyse de la variance)

On peut décomposer la variance de Y comme :

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ \text{Variance totale de } Y &= \text{Variance expliquée} + \text{Variance résiduelle} \end{aligned}$$

Plan

Coefficient de détermination R^2

Formule

- Ce coefficient, noté R^2, sert à mesurer l'adéquation entre le modèle ($\hat{y} = \hat{a}x + \hat{b}$) et les observations
- $R^2 = \frac{\text{Variance expliquée}}{\text{Variance totale}}$
- On peut montrer que $R^2 = \frac{\mathtt{Cov}^2(X,Y)}{\sigma_X^2 \sigma_Y^2} = r^2$

Interpretation

- $0 \leq R^2 \leq 1$
- Plus R^2 tend vers 1, plus le nuage de points se rapproche de la droite de régression
- Au contraire, plus R^2 se rapproche de 0, plus le nuage de points est diffus autour de la droite de régression.
- $R^2 = 1$: Le modèle est capable de déterminer 100% de la distribution de points

Exemple

- $\mathtt{Var}(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{y}^2 = \frac{203.65}{24} - (2.88)^2 \approx 0.19$
- $\sigma_X = \sqrt{0.08} \approx 0.28$
- $\sigma_Y = \sqrt{0.19} \approx 0.44$
- $r = \frac{\mathtt{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0.12}{0.28 \times 0.44} \approx 0.97$
- $R^2 = \frac{\mathtt{Cov}^2(X,Y)}{\sigma_X^2 \sigma_Y^2} = r^2 = (0.97)^2 \approx 0.94$