E.I.S.T.I. – Département de Mathématiques
1ère année d'ingénieurs

STATISTIQUES DESCRIPTIVES
TD N°4 : statistiques bivariées
Quantitatif-Qualitatif / Quantitatif-Quantitatif

Quantitatif-Qualitatif

Dans une population Ω de taille n, on observe deux caractéristiques :

- une qualitative, $x=\{x_k\}_{k=1,\dots,n,}$ à p modalités notées, $m_1,\dots,m_l,\dots,m_p$
- une quantitative continue $y=\{y_k\}_{k=1,\dots,n}$ de moyenne $\bar{y}$ et de variance s_y^2.

On suppose que les modalités de la série x définissent des sous-populations

$$\Omega = \Omega_1 \cup \dots \cup \Omega_p \quad \text{où } \Omega_l \cap \Omega_{l'} = \varnothing,$$

On peut alors considérer les restrictions de la caractéristiques y sur chacune des sous-populations et calculer les indicateurs numériques usuels pour chaque modalités de x,

- moyennes : $\overline{y_l}$, l=1,…,p
- variances : s_l^2, l=1,…,p

….

Exercice 1 : Formules de décomposition

Montrer que

$$\bar{y} = \frac{1}{n}\sum_{l=1}^{p} n_l \overline{y_l}$$

$$s_y^2 = \frac{1}{n}\sum_{l=1}^{p} n_l(\overline{y_l} - \bar{y})^2 + \frac{1}{n}\sum_{l=1}^{p} n_l s_l^2 = s_E^2 + s_R^2$$

Calculs longs – A faire par le prof au tableau

A quoi correspondent les termes s_E^2 et s_R^2 ?

On définit un indice de liaison entre les deux caractéristiques x et y par le rapport de corrélation

$$s_{y/x} = \sqrt{\frac{s_E^2}{S_y^2}}$$

Donner un encadrement de $S_{y/x}$. A quoi correspondent les cas $S_{y/x}=0$ et $S_{y/x}=1$?

Exercice 2

Dans une entreprise on étudie le lien entre le salaire mensuel, l'âge et le nombre d'enfants des cadres supérieurs.

On effectue une première étude entre le nombre d'enfants et le salaire.

1) Sur un même graphique, tracer les boites à moustaches. Commenter.
2) Calculer le rapport de corrélation. Qu'en pensez-vous ?

individus	Age	Salaire (k€)	nb enfants
1	junior	35	1
2	senior	51	3
3	expert	39	3
4	expert	40	1
5	senior	50	3
6	expert	37	3
7	junior	33	0
8	junior	36	0
9	expert	42	2
10	senior	44	0
11	expert	46	1
12	expert	36	2
13	expert	41	2
14	senior	49	1
15	junior	36	0
16	junior	34	2
17	junior	36	1
18	expert	36	3
19	junior	32	2
20	junior	35	2

Moyenne	39,4
Variance	33,73

salaires (k€)

nb enfants			
0	1	2	3
33	35	41	51
36	40	34	39
44	46	32	50
36	49	42	37
	36	36	36
		35	

Effectif	4	5	6	5
Moyenne	37,25	41,20	36,67	42,60
Mediane	36,00	40,00	35,50	39,00

Variance	22,25	37,70	15,87	53,30
Ecart-type	4,72	6,14	3,98	7,30
CV	0,13	0,15	0,11	0,17
Q1	35,25	36,00	34,25	37,00
Q3	38,00	46,00	39,75	50,00
IIQ	2,75	10,00	5,50	13,00
m	31,13	21,00	26,00	17,50
M	42,13	61,00	48,00	69,50

Quantitatif-Quantitatif

Dans une population Ω de taille n, on observe deux caractéristiques quantitatives continues, $x=\{x_k\}_{k=1,\dots,n}$, et $y=\{y_k\}_{k=1,\dots,n}$, de moyennes $\bar{x}$ et $\bar{y}$ et de variances s_x^2 et s_y^2.

Exercice 3 : covariance et coefficient de corrélation linéaire et droite de régression

La *covariance* entre les deux caractéristiques est définie par

$$c_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\,\bar{y}$$

C'est une forme bilinéaire symétrique, à valeurs réelles, telle que

$$c_{xx} = s_x^2$$

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy}$$

$$c_{xy}^2 \le s_x^2 s_y^2 \text{ (inégalité de Cauchy-Schwartz)}$$

On définit alors le *coefficient de corrélation linéaire* par

$$r_{xy} = \frac{c_{xy}}{s_x s_y}.$$

1) Montrer que le coefficient de corrélation linéaire est symétrique, à valeurs dans [-1,1] et correspond à la covariance des observations réduites et centrées. A quoi correspondent les valeurs -1 et +1 ?

très rapide

La droite de régression de y sur x est construite en minimisant,

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

C'est-à-dire en minimisant la moyenne des écarts au carré entre l'observation y_i et la valeur de la droite au point x_i. On obtient alors

$$\hat{a} = \frac{c_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

La série des valeurs prédites par la droite de régression est donnée par

$$\hat{y}_i = \hat{a}x_i + \hat{b},$$

et les *résidus* par

$$\hat{e}_i = y_i - (\hat{a}x_i + \hat{b}).$$

2) Montrer que la moyenne des valeurs prédites est égale à la moyenne de la série observée.

très rapide

3) Montrer que les résidus sont de moyenne nulle et sont non corrélés avec la série X. Qu'est-ce que cela signifie ?

rapide

4) Montrer la formule de décomposition de la variance

$$s_y^2 = s_E^2 + s_R^2$$

où s_E^2 est la *variance expliquée* par la droite de régression, et s_R^2 est la *variance résiduelle.*

Calculs longs – A faire par le prof au tableau

On peut alors montrer que le *coefficient de détermination*

$$R^2 = \frac{s_E^2}{s_y^2},$$

qui donne le taux de variance expliquée par la droite de régression, est égale au coefficient de corrélation linaire au carré, $R^2 = r_{xy}^2$.

Exercice 4

On donne pour les six premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant des emplois durables à temps plein) et de demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles, à la recherche d'un emploi durable à plein temps). Les nombres sont exprimés en milliers.

Offres (x_i)	61	66,7	75,8	78,6	82,8	87,2
Demandes (y_i)	2034	2003,8	1964,5	1928,2	1885,3	1867,1

On a les résultats suivants

$\bar{x} = 75{,}35$ $\bar{y} = 1947{,}15$ $s_x^2 = 97{,}49$ $s_y^2 = 4329{,}14$ $c_{xy} = -639{,}90$

1) Tracer le nuage de points des demandes en fonction des offres.
2) Calculer le coefficient de corrélation linéaire. Conclusion
3) Déterminer la droite de régression. Tracer la droite sur le graphique.
4) Calculer les prévisions et les résidus. Vérifier les hypothèses sur les résidus.

Exercice 5 (PY Bernard, exercices corrigés de statistique descriptive, ed. economica)

Une étude a été menée auprès d'entreprises afin d'établir le lien entre les quantités commandés d'un bien, Y, et son prix, X et on obtient les observations suivantes.

Prix de vente (€)	Quantités commandées		log(Prix)	log(Quantités)
95	104		4,55	4,64
130	58		4,87	4,06
148	37		5,00	3,61
210	22		5,35	3,09
250	12		5,52	2,48
330	9		5,80	2,20
		Moyenne	**5,18**	**3,35**
		Variance	**0,21**	**0,88**
		Covariance	**-0,43**	

1) Tracer le nuage de points. Conclusion.
2) On pose u=log(x) et v=log(y). Quelle est la relation entre u et v ?
3) Calculer le coefficient de corrélation linéaire entre u et v.
4) Trouver la droite de régression de v sur u.
5) En déduire la quantité qui serait commandée si le prix était fixé à 75€.

Corrections

Exercice 1

Notons $y_{k,l}$,ki=1,…,n_l, la k^{ème} observation de la caractéristique Y appartenant à la sous-population Ω_l .

- Décomposition de la moyenne

$$\overline{y_l} = \frac{1}{n_l}\sum_{k=1}^{n_l} y_{k,l} \Rightarrow n_l\overline{y_l} = \sum_{k=1}^{n_l} y_{k,l} \Rightarrow \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l} = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l} = \frac{1}{n}\sum_{k=1}^{n} y_k = \overline{y}$$

- Décomposition de la variance

$$s_E^2 = \frac{1}{n}\sum_{l=1}^{p} n_l(\overline{y_l} - \overline{y})^2 = \frac{1}{n}\sum_{l=1}^{p} n_l(\overline{y_l}^2 - 2\overline{y_l}\overline{y} + \overline{y}^2) = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \frac{2}{n}\overline{y}\underbrace{\sum_{l=1}^{p} n_l\overline{y_l}}_{=n\overline{y}} + \frac{1}{n}\overline{y}^2\underbrace{\sum_{l=1}^{p} n_l}_{=n}$$

$$= \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - 2\overline{y}^2 + \overline{y}^2 = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \overline{y}^2$$

Nous avons

$$s_l^2 = \frac{1}{n_l}\sum_{k=1}^{n_l}(y_{k,l} - \overline{y_l})^2 ,$$

d'où

$$s_R^2 = \frac{1}{n}\sum_{l=1}^{p} n_l s_l^2 = \frac{1}{n}\sum_{l=1}^{p} n_l \frac{1}{n_l}\sum_{k=1}^{n_l}(y_{k,l} - \overline{y_l})^2 = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l}(y_{k,l}^2 - 2y_{k,l}\overline{y_l} + \overline{y_l}^2)$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{2}{n}\sum_{l=1}^{p}\overline{y_l}\underbrace{\sum_{k=1}^{n_l} y_{k,l}}_{=n_l\overline{y_l}} + \frac{1}{n}\sum_{l=1}^{p}\overline{y_l}^2\underbrace{\sum_{k=1}^{n_l} 1}_{=n_l}$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{2}{n}\sum_{l=1}^{p}\overline{y_l}n_l\overline{y_l} + \frac{1}{n}\sum_{l=1}^{p}\overline{y_l}^2 n_l = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2$$

Donc

$$s_E^2 + s_R^2 = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \overline{y}^2 + \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \overline{y}^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - \overline{y}^2 = s_y^2$$

Le premier terme correspond à la *variance expliquée* par la partition de la série observée X et le deuxième terme est un reste appelé *variance résiduelle*.

- Rapport de corrélation

$$s_y^2 = s_E^2 + s_R^2 \Rightarrow 0 \le s_E^2 \le s_y^2 \Rightarrow 0 \le \frac{s_E^2}{s_y^2} \le 1 \Rightarrow 0 \le s_{y/x} \le 1$$

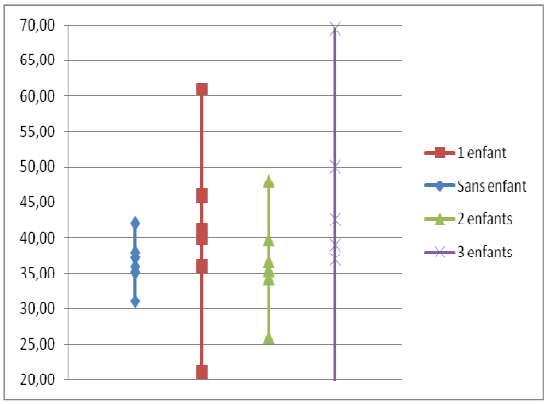
Cela signifie que s_E^2 représente le pourcentage de variabilité de y expliquée par x.

- Si $s_{y/x}=0$ alors la variance expliquée est nulle, il n'y a donc aucun lien entre y et x
- Si $s_{y/x}=1$ alors la variance expliquée est égale à la variance de y donc y est entièrement expliquée par x.

Attenttion : $s_{y/x} \neq s_{x/y}$. Cela signifie que si $s_{y/x}$ est proche de 1 alors x permet d'expliquer y mais pas réciproquement.

Exercice 2

1) Il ne semble pas y avoir de lien entre nb enfants et salaire moyen. Plus grande dispersion pour 1 enfant et 3 enfants …



2)

variance résiduelle	31,96
variance expliquée	6,54
rapport de corrélation	0,9735

Le rapport de corrélation est très fort et pourtant il semble absurde de conclure que le nombre d'enfants conditionne le salaire.. Il n'y a pas de lien direct réel entre les deux caractéristiques mais la liaison peut venir d'une troisième variable cachée, par exemple ici l'âge. Les calculs des coefficients de corrélation montreraient certainement un lien très important entre le salaire et l'âge et entre l'âge et le nombre d'enfants. **En conclusion, il ne faut pas oublier que la corrélation n'explique pas la causalité**.

Exercice 3

1) D'après l'inégalité de Cauchy-Schwartz, on a

$$0 \le c_{xy}^2 \le s_x^2 s_y^2 \Rightarrow 0 \le \frac{c_{xy}^2}{s_x^2 s_y^2} \le 1 \Rightarrow 0 \le r_{xy}^2 \le 1 \Rightarrow -1 \le r_{xy} \le 1$$

Les valeurs +1 et -1 correspondent à une relation linéaire parfaite entre x et y, c'est-à-dire à l'existence de deux coefficients a et b tels que : y=ax+b.

Notons $x'=\{\frac{x_k-\bar{x}}{s_X}\}_{k=1,\dots,n}$ et $y'=\{\frac{y_k-\bar{y}}{s_Y}\}_{k=1,\dots,n}$ les séries observées centrées et réduites.

$$c_{x'y'} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k-\bar{x}}{s_X}\right)\left(\frac{y_k-\bar{y}}{s_Y}\right) = \frac{1}{s_X}\frac{1}{s_Y}\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \frac{1}{s_x}\frac{1}{s_y}c_{xy}$$

2)

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{a}x_i + \hat{b} = \hat{a}\frac{1}{n}\sum_{i=1}^{n}x_i - \hat{b} = \hat{a}\bar{x} - \hat{b} = \bar{y}$$

3)

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = \frac{1}{n}\sum_{i=1}^{n}[y_i - (\hat{a}x_i + \hat{b})] = \frac{1}{n}\sum_{i=1}^{n}y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n}x_i - \hat{b} = \bar{y} - \hat{a}\bar{x} - \hat{b}$$
$$= \bar{y} - \hat{a}\bar{x} - (\bar{y} - \hat{a}\bar{x}) = 0$$

Pour montrer qu'ils sont non corrélés, il suffit de montrer que la covariance est nulle,

$$c_{xe} = \frac{1}{n}\sum_{i=1}^{n}x_i\hat{e}_i - \bar{x}\,\bar{e} = \frac{1}{n}\sum_{i=1}^{n}x_i\hat{e}_i = \frac{1}{n}\sum_{i=1}^{n}x_i(y_i - \hat{a}x_i - \hat{b}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \hat{b}\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$= (c_{xy} + \bar{x}\,\bar{y}) - \hat{a}(s_x^2 + \bar{x}^2) - \hat{b}\bar{x} = (c_{xy} + \bar{x}\,\bar{y}) - \hat{a}(s_x^2 + \bar{x}^2) - (\bar{y} - \hat{a}\bar{x})\bar{x}$$

$$= c_{xy} - \hat{a}s_x^2 = c_{XY} - \frac{c_{xy}}{s_x^2} s_x^2 = 0$$

Cela signifie qu'il ne reste plus « d'information » pour expliquer y par x dans les résidus.

4)

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

- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = s_e^2$ car $\bar{e} = 0$
- $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = s_y^2$ car $\bar{\hat{y}} = \bar{y}$
- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} e_i\hat{y}_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n} e_i\hat{y}_i$ car $\bar{e} = 0$

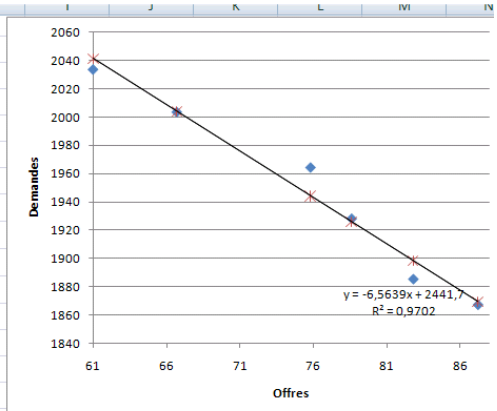
  $= \frac{1}{n}\sum_{i=1}^{n} e_i(\hat{a}x_i + \hat{b}) = \hat{a}\frac{1}{n}\sum_{i=1}^{n} e_i x_i + \hat{b}\bar{e} = 0$ d'après 3)

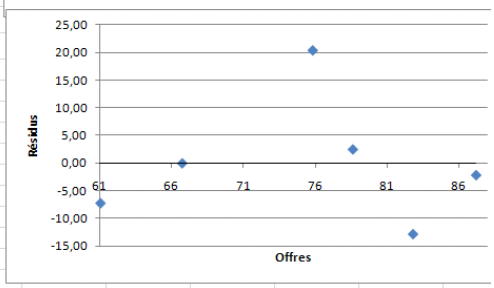
Exercice 4

	Offres	Demandes		Prévisions	Résidus
	61	2034		2041,34	-7,34
	66,7	2003,8		2003,93	-0,13
	75,8	1964,5		1944,20	20,30
	78,6	1928,2		1925,82	2,38
	82,8	1885,3		1898,25	-12,95
	87,2	1867,1		1869,37	-2,27
Moyenne	75,35	1947,15	Moy	1947,15	0,00
Variance	97,49	4329,14		ok	ok
Covariance	-639,90		coef. corr. (e,x)	0,00	
				ok	
			coef. corr. (x,y)	-0,98	
			R^2	0,97	
			a	-6,56	
			b	2441,74	

Attention!!! la variance dans Excel est sans biais mais pas la covariance, Il faut donc penser à multiplier la covariance par n/(n-1)

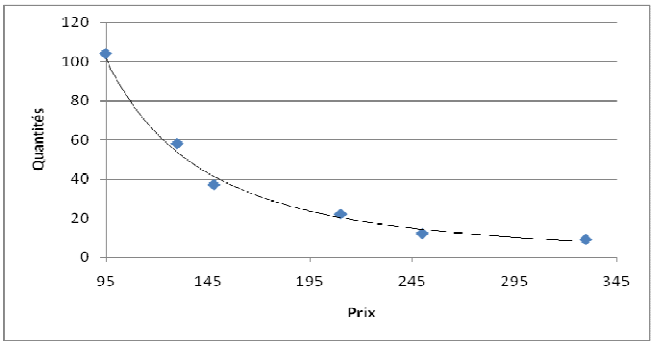
négatif car plus l'offre est forte et plus la demande est faible





Exercice 5

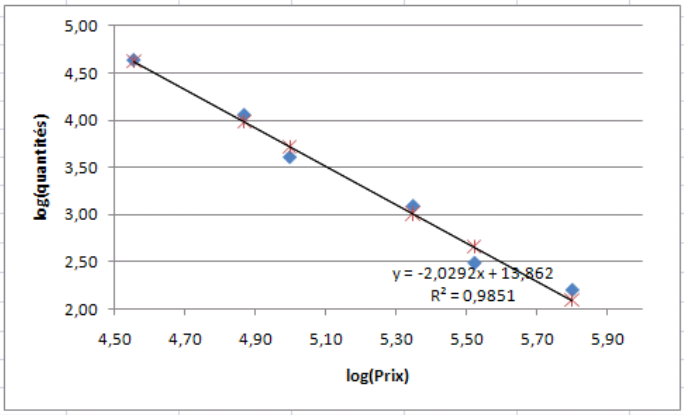
1) Le nuage de point suggère une relation puissance, $y=\alpha e^{\beta x}$, entre x et y.

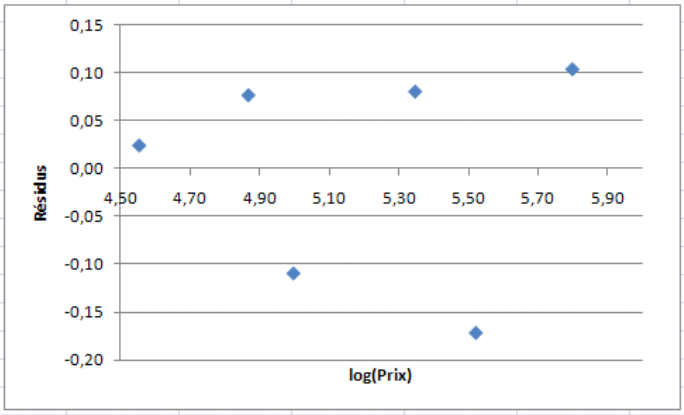


2) Si on passe au log de chaque côté on a ln(y)=ln(α)+βx, autrement dit une relation linéaire v=au+b, où a=β et b=ln(α).

3)

	Prévisions	Résidus
	4,62	0,02
	3,98	0,08
	3,72	-0,11
	3,01	0,08
	2,66	-0,17
	2,09	0,10
Moy	3,35	0,00
coef. corr. (e,x)	0,00	
coef. corr. (x,y)	**-0,99**	
R^2	**0,99**	
a	**-2,03**	
b	**13,86**	





x=75€ ⇒ u=ln(x)=4,32 ⇒ v=ln(y)=-2,03*4,32+13,86=5,09 ⇒ y=exp(v)=163,27