La régression simple

La régression simple

Étude de la liaison linéaire entre deux variables numériques :

- une variable explicative X
- une variable à expliquer Y

Étude du lien entre deux variables X et Y

Variable X explicative

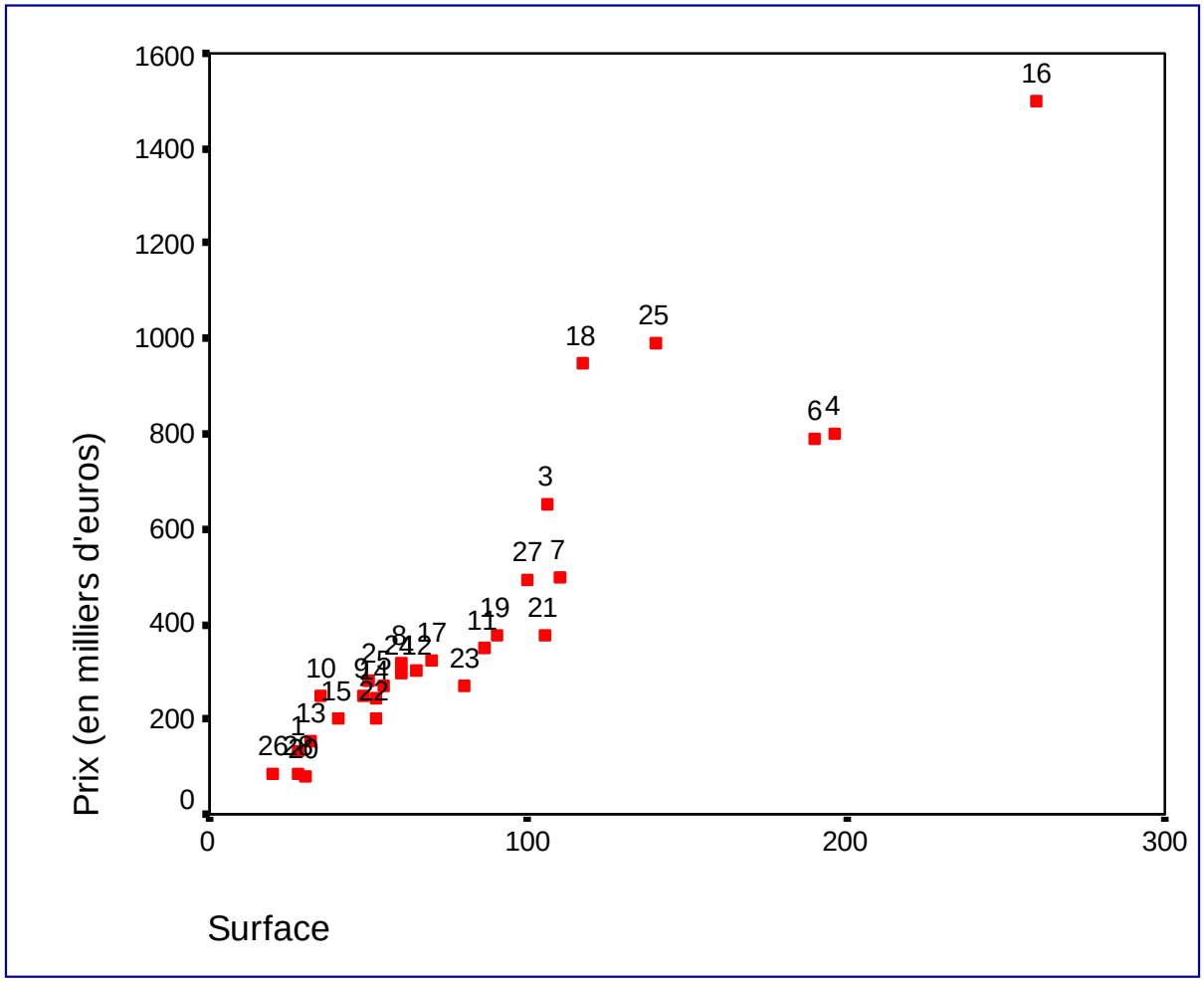
Variable Y à expliquer	Quantitatif	Qualitatif
Quantitatif	- Régression simple - Corrélation simple	Analyse de la variance à un facteur
Qualitatif	Régression logistique	Test du khi-deux d'indépendance

Cas Prix d'un appartement

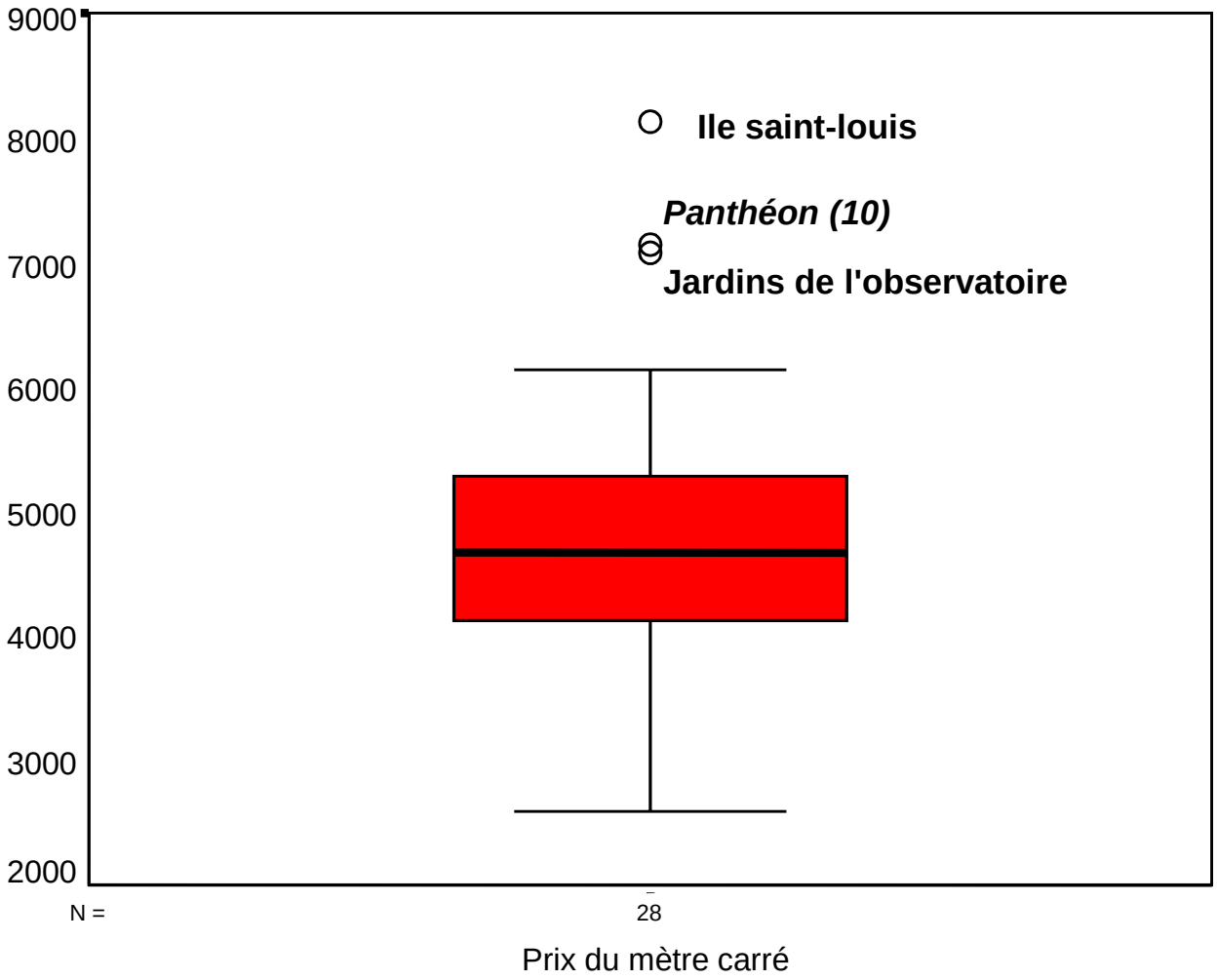
1. CENSIER, bas de R. Mouffetard, pied-à-terre, 28m^2, tt confort. Visite vendredi, samedi, dim. 130.000 € à discuter. Facilités	2. CONTRESCARPE, imm. Ancien, pierre de taille, beau duplex caractère, 50m^2, poutres, refait neuf, 280.000 €
3. R. St-Simon, en pleine verdure, calme, plein soleil, Superbe appt 4p., 106m^2, cuis. aménagée, s. de bains moderne, chff. cent. Parfait état. Px 650.000 à discuter. Agence s'abstenir. Direct. Propriétaire.	4. RAPP 7P., 196m^2 standing, 9 fenêtres plein soleil, 800.000 €.
5. R. St André-des-Arts, beau liv + chbre, imm. XVIII^e siècle, 55m^2, 268.000 €.	6. 5^e PRES QUAIS, 7 pces, 190m^2 caractère, standing, 790.000 €
7. GOBELINS, Beau 5p., 110m^2, gd cft, soleil, 500.000 €	8. GOBELINS, et. élevé, calme, asc., 2 pièces, 60m^2, 320.000 €
9. CENSIER, très grand studio + entrée 48m^2, tt cft, ensoleillé, calme, bel imm., 250.000 €	10. PANTHEON, 7^e étage, ascenseur, grand studio 35m^2 + terrasse. Vue. 250.000 €.
11. RUE MADAME, 3P. + Serv., 86m^2, 350.000 €.	12. RUE DE SEINE, 3P., tt cft, 65m^2, calme, soleil, 300.000 €.
13. PANTHEON, bel imm., verdure, magnifique studio 32m^2, caractère, 155.000 €.	14. SEVRES BAB, 1er ét., 2P., gde cuis., bns, 52m^2, état neuf, 245.000 €.
15. MONTPARNASSE, Part. vend atelier d'artiste 40m^2, duplex, vue imprenable, tout confort, Prix 200.000 €.	16. RUE D'ASSAS, imm. gd standing, bel appart 260m^2, triple récept. + 5 ch., tt cft (travaux) 2 park., 2 ch. Serv., Prix 1.500.000 € à déb.
17. BD St-GERMAIN, 4P., 70m^2, à amén., 4^e ét., 325.000 €.	18. ILE St-LOUIS, Lux. appt., 117m^2, en duplex, gde récept., gde chambre, 2 sdb, Terras., parf. et., décor tr. bon goût, 950.000 €.
19. JUSSIEU, Charme, gd 3 pces, 90m^2, 378.000 €.	20. QUARTIER LATIN, 30m^2 à aménager, prix 78.000 €.
21. MONTPARNASSE, Imm. p.d.t., 4-5 P., 105m^2, bon état, 375.000 €.	22. RUE MAZARINE, 4^e ét., sans ascens., 52m^2 à rénover. Prix total 200.000 €.
23. CENSIER, Bel imm., 4P. 80m^2, tt cft, petits travaux, 270.000 €.	24. ASSAS LUXEMBOURG, 3P. 60m^2 s/arbres, imm. caractère, 295.000 €.
25. SUR JARDINS OBSERVATOIRE, 140m^2, grand charme, 990.000 €.	26. RUE DE SAVOIE, 4^e ét., Studio 20m^2, dche, 85.000 €. crédit possible.
27. PRES LUXEMBOURG, Bel imm., pierre de taille, Appartement 100m^2, salon, sal. à manger, 2 chbres, office, cuis., bains, chf. cent., asc., prix : 495.000 €.	28. Mo GOBELINS, studio, cuis., s. de bains, 28m^2, calme. Prix 85.000 €.

Cas Prix d'un appartement

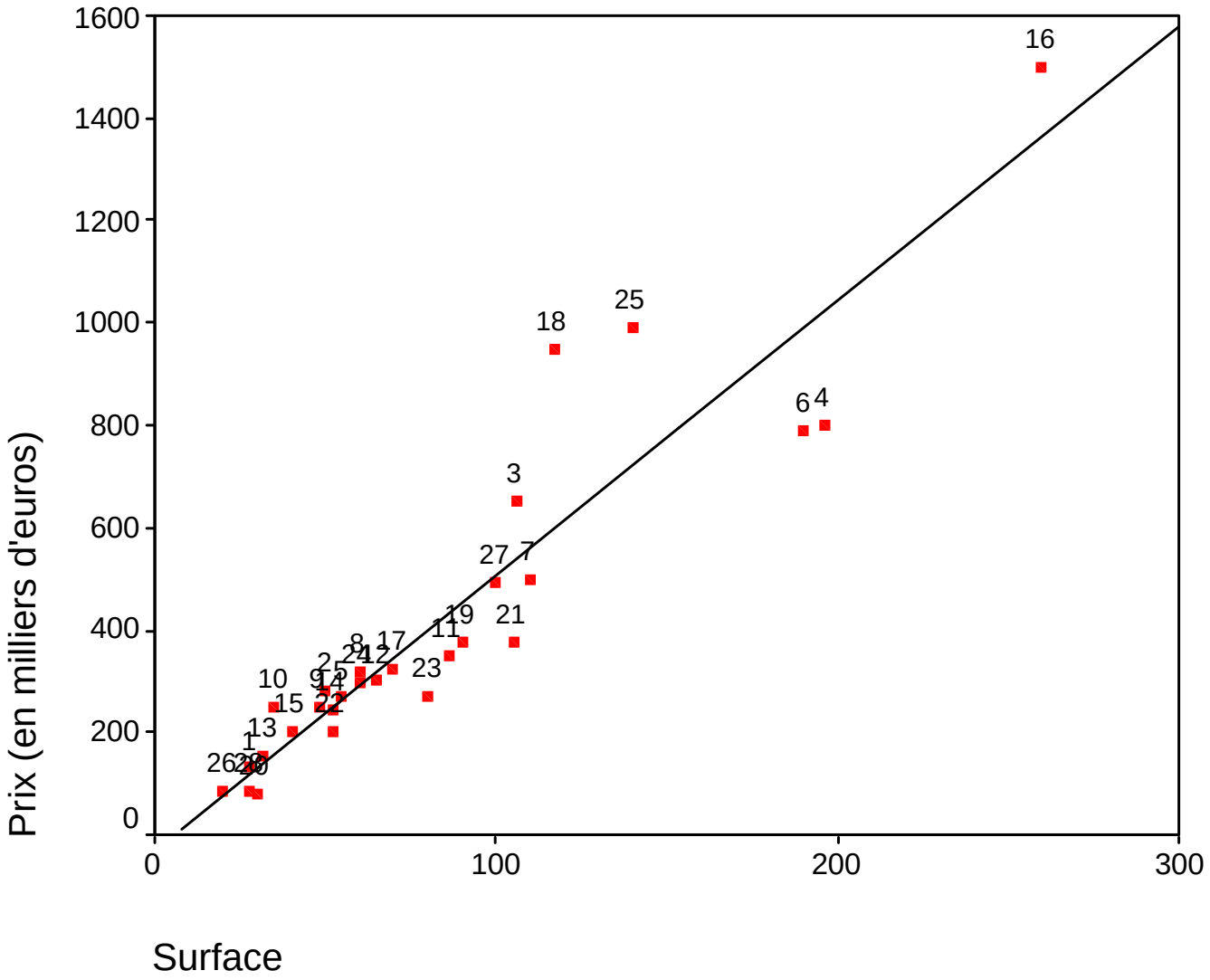
	Localisation	Surface	Prix (en milliers d'euros)
1	censier	28	130
2	contrescarpe	50	280
3	rue saint-simon	106	650
4	rapp	196	800
5	saint-andré des arts	55	268
6	5-ième, près quais	190	790
7	gobelins	110	500
8	gobelins	60	320
9	censier	48	250
10	panthéon	35	250
11	rue madame	86	350
12	rue de seine	65	300
13	panthéon	32	155
14	sèvres-babylone	52	245
15	montparnasse	40	200
16	rue d'assas	260	1500
17	saint-germain	70	325
18	ile saint-louis	117	950
19	jussieu	90	378
20	quartier-latin	30	78
21	montparnasse	105	375
22	rue mazarine	52	200
23	censier	80	270
24	assas luxembourg	60	295
25	jardins de l'observatoire	140	990
26	rue de savoie	20	85
27	près luxembourg	100	495
28	gobelins	28	85



Identification des outliers au niveau du Prix au mètre carré

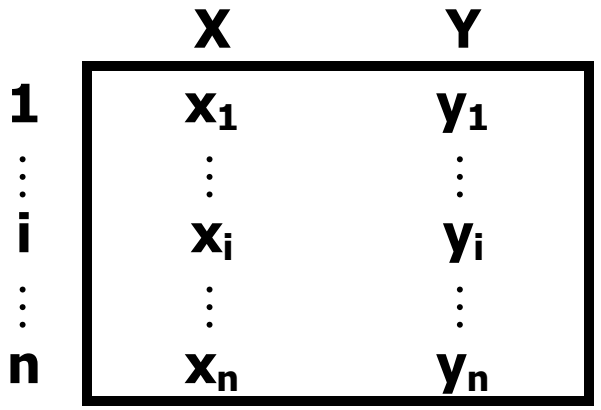


La droite des moindres carrés



Les données

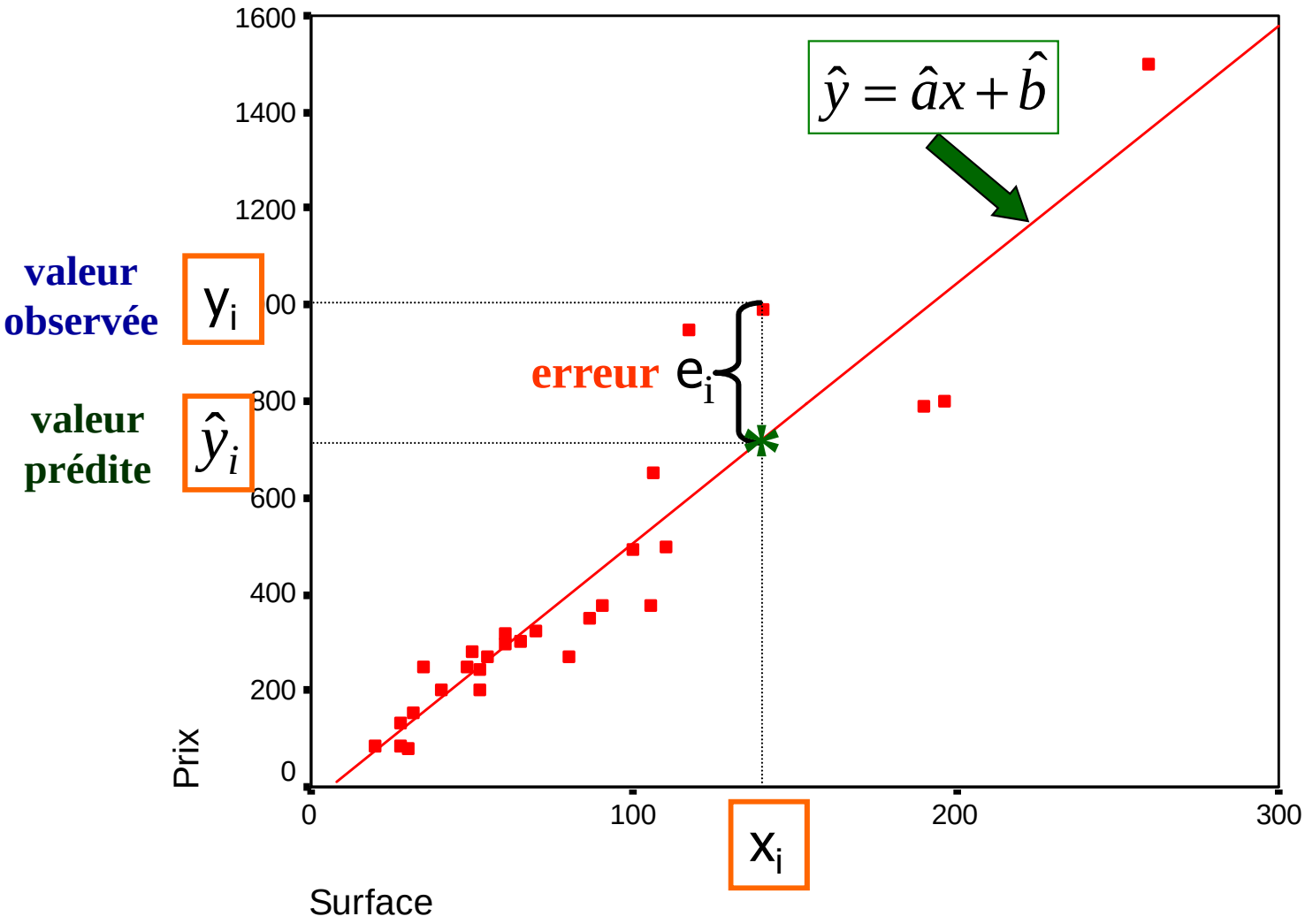
- Y = Variable à expliquer numérique (ou dépendante)
- X = Variable explicative numérique ou binaire (ou indépendante)



	X	Y
1	x_1	y_1
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

Le tableau des données

La droite des moindres carrés



On cherche

$\hat{a}$ et $\hat{b}$

minimisant

$$\sum_{i=1}^{n} e_i^2$$

Résultats SPSS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-29.466	41.246		-.714	.481
	Surface	5.353	.414	.930	12.931	.000

a. Dependent Variable: Prix (en milliers d'euros)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2527208	1	2527207.505	167.210	.000[a]
	Residual	392963.2	26	15113.970		
	Total	2920171	27			

a. Predictors: (Constant), Surface

b. Dependent Variable: Prix (en milliers d'euros)

Résultats SPSS

Numéro	Résidu standardisé	Prix (en milliers €)	Valeur prédite	Résidu
1	0.078	130	120.42	9.58
2	0.340	280	238.19	41.81
3	0.911	650	537.97	112.03
4	-1.788	800	1019.75	-219.75
5	0.025	268	264.96	3.04
6	-1.608	790	987.64	-197.64
7	-0.483	500	559.38	-59.38
8	0.230	320	291.72	28.28
9	0.183	250	227.49	22.51
10	0.749	250	157.89	92.11
11	-0.658	350	430.91	-80.91
12	-0.150	300	318.49	-18.49
13	0.107	155	141.84	13.16
14	-0.032	245	248.90	-3.90
15	0.125	200	184.66	15.34
16	1.120	1500	1362.36	137.64
17	-0.165	325	345.26	-20.26
18	2.873	950	596.85	353.15
19	-0.605	378	452.32	-74.32
20	-0.432	78	131.13	-53.13
21	-1.282	375	532.62	-157.62
22	-0.398	200	248.90	-48.90
23	-1.048	270	398.79	-128.79
24	0.027	295	291.72	3.28
25	2.196	990	719.98	270.02
26	0.060	85	77.60	7.40
27	-0.088	495	505.85	-10.85
28	-0.288	85	120.42	-35.42

Modèle :

Prix calculé

= -29.466 + 5.353 Surface

$$\sum_{i=1}^{n} e_i^2 = 392963.2$$

Pour le modèle avec constante on a aussi :

$$\sum_{i=1}^{n} e_i = 0$$

Coefficient de détermination R^2, Coefficient de corrélation R

A) Formule de décomposition

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{Somme des carrés totale (Total Sum of Squares)}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{Somme des carrés expliquée (Regression Sum of Squares)}} + \underbrace{\sum e_i^{\,2}}_{\text{Somme des carrés résiduelle (Residual Sum of Squares)}}$$

B) $$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

C) $$R = \text{signe}(\hat{a})\sqrt{R^2}$$

Résultats SPSS

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.930[a]	.865	.860	122.939

a. Predictors: (Constant), Surface

|R|

Correlations

		PRIX
SURFACE	Pearson Correlation	.930**
	Sig. (2-tailed)	.000
	N	28

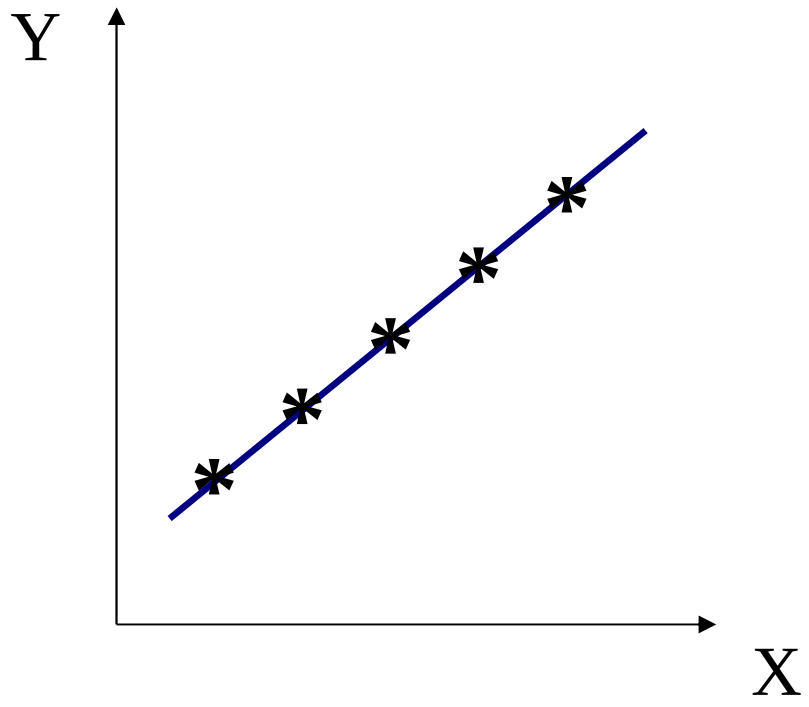
R

**. Correlation is significant at the 0.01 level

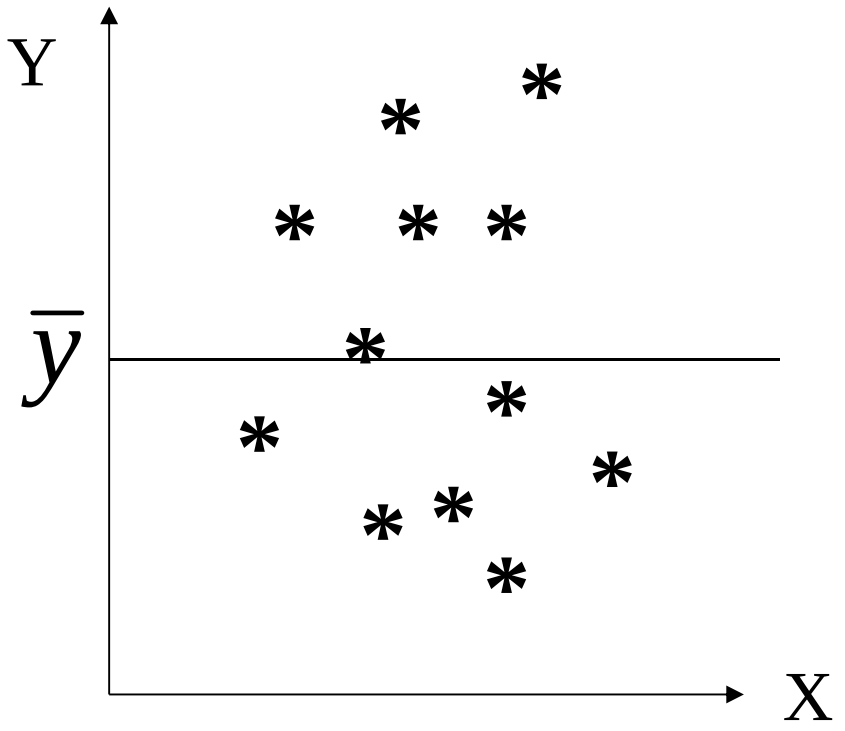
Le R^2 mesure la force de la liaison linéaire entre X et Y

1) $0 \leq R^2 \leq 1$

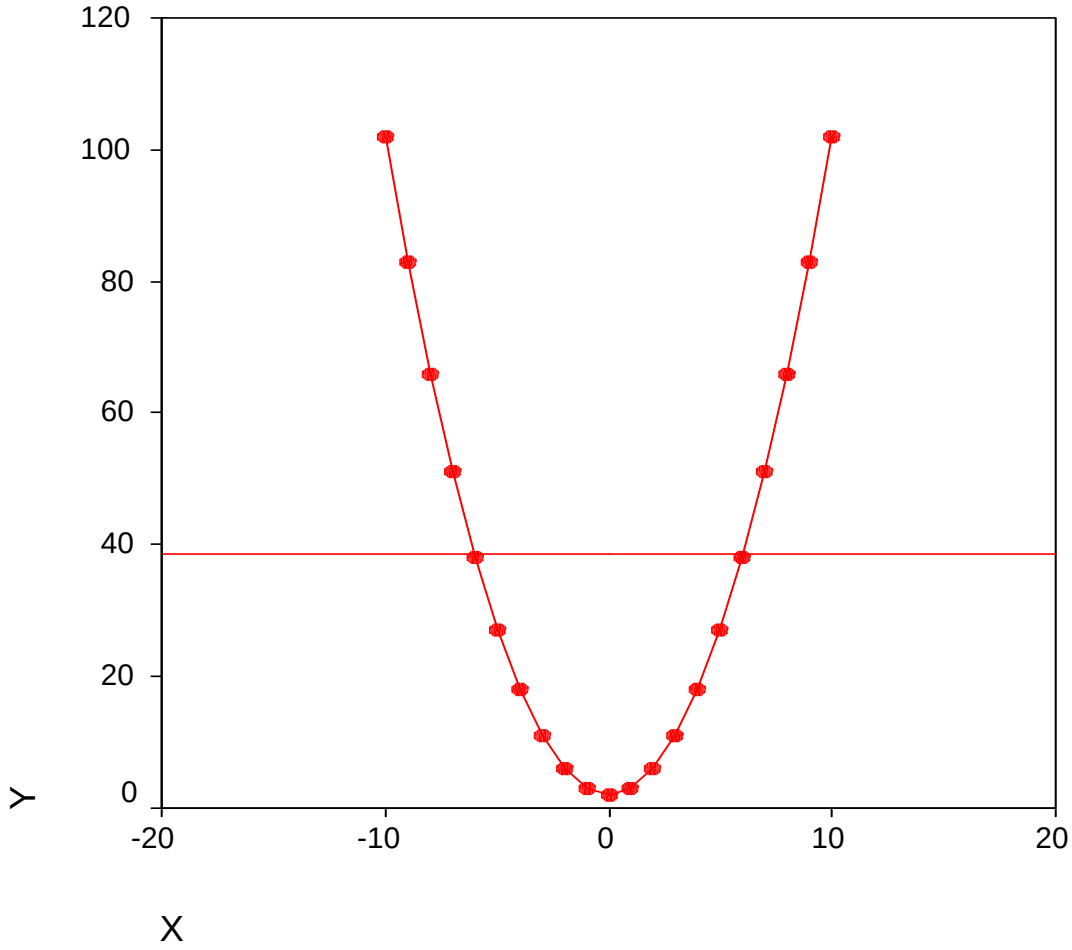
2) $R^2 = 1$



3) $R^2 = 0$



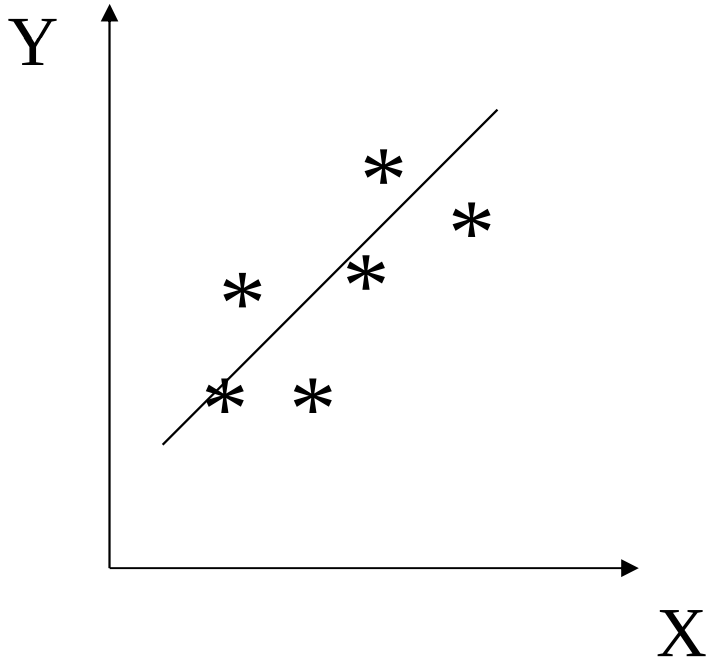
Le R^2 mesure la force de la liaison linéaire entre X et Y



Modèle non linéaire:
$Y = aX^2 + bX$

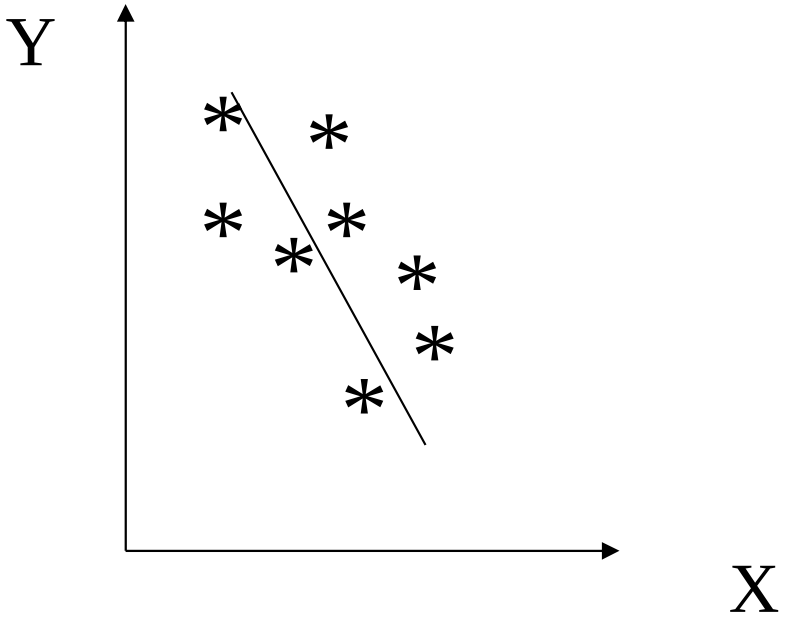
Rsq = 0.0000

La corrélation R mesure la force et le sens de la liaison linéaire entre X et Y



$\hat{a} > 0$

$R > 0$

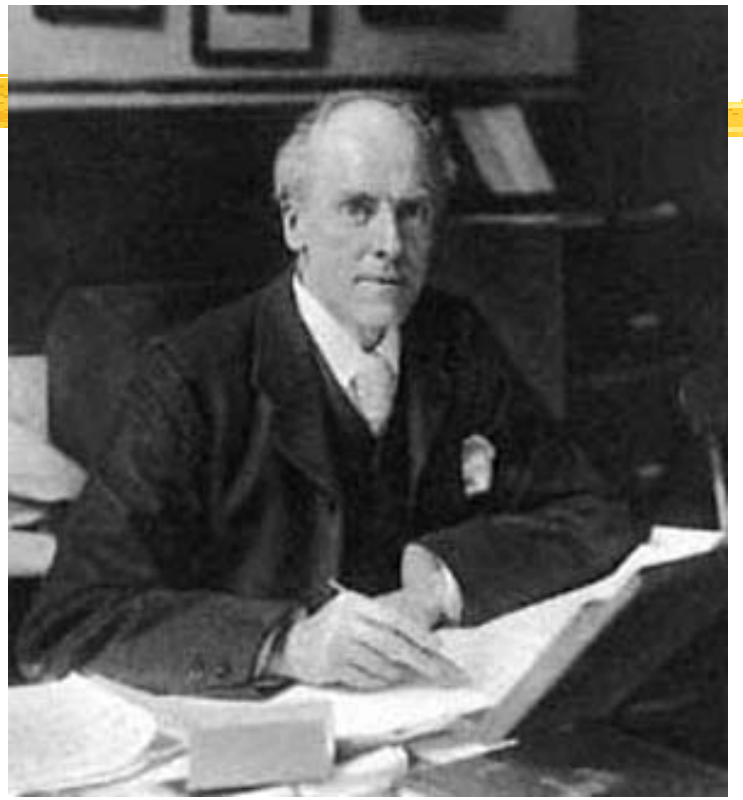


$\hat{a} < 0$

$R < 0$

Calcul direct de R

Karl Pearson



$$R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

Résultat SPSS :

Correlations

		PRIX
SURFACE	Pearson Correlation	.930**
	Sig. (2-tailed)	.000
	N	28

**. Correlation is significant at the 0.01 level

La corrélation R est-elle significative au risque $\alpha = 0.05$?

Notations

- ρ = Corrélation au niveau de la population
- R = Corrélation au niveau de l'échantillon

Test :

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Règle de décision

On rejette H_0 au risque $\alpha = 0.05$ de se tromper si

$$|R| \geq \frac{2}{\sqrt{n}}$$

(Bonne approximation pour n > 20)

La corrélation R est-elle significative au risque α ?

Notations

- ρ = Corrélation au niveau de la population
- R = Corrélation au niveau de l'échantillon

Test :

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Règle de décision

On rejette H_0 au risque α de se tromper si

$$|R| \geq \frac{t_{1-\alpha/2}(n-2)}{\sqrt{t^2_{1-\alpha/2}(n-2)+n-2}}$$

Niveau de signification

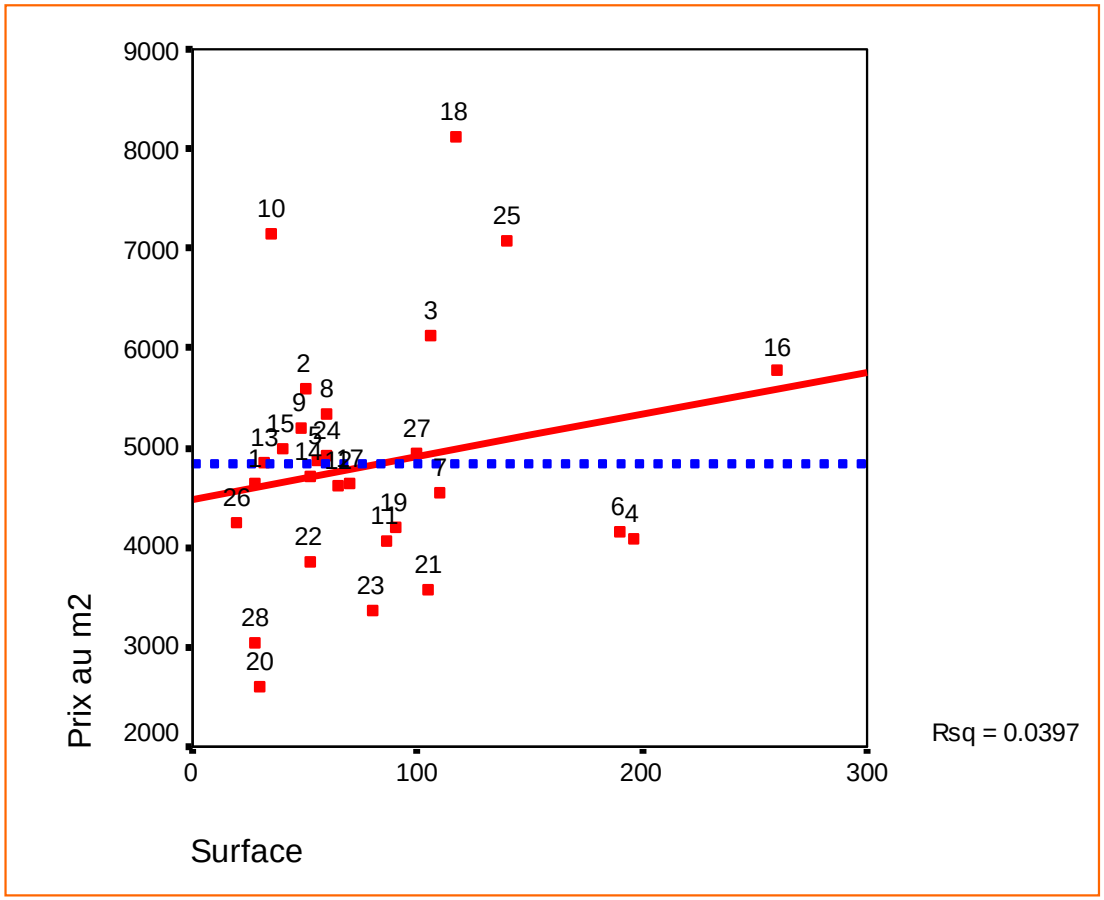
Plus petit α conduisant au rejet de H_0.

Seuil critique $r(\nu = n-2, \alpha)$ pour un coefficient de corrélation simple

Table 6

ν \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Exemple de corrélation non significative



Correlations

		SURFACE
PRIXM2	Pearson Correlation	.199
	Sig. (2-tailed)	.309
	N	28

On a 30,9 chances sur 100 de se tromper en affirmant qu'il existe une liaison linéaire entre le prix au m^2 et la surface.

On considère donc que la corrélation (.199) entre le prix au m^2 et la surface n'est pas significative.

En rouge la droite des moindres carrés, en bleu la droite y = prix au m^2 moyen

Le modèle statistique de la régression simple

- Chaque valeur observée y_i est considérée comme une réalisation d'une variable aléatoire Y_i définie par :

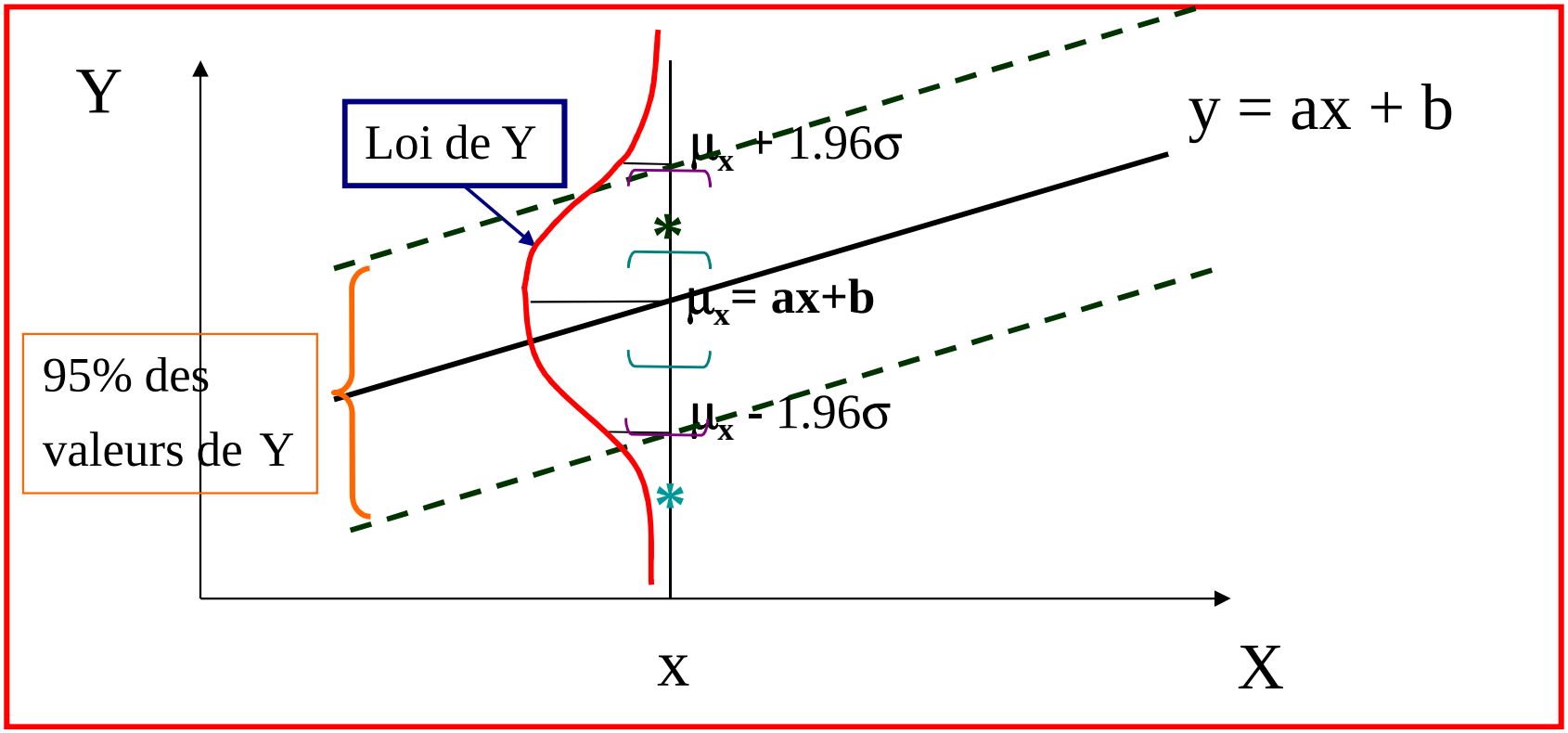
$$Y_i = ax_i + b + \varepsilon_i$$

où ε_i est un terme aléatoire suivant une loi normale $N(0, \sigma)$.

- On suppose que les les aléas ε_i sont indépendants les uns des autres.

Le modèle de la régression simple

Modèle : $Y = aX + b + \varepsilon$, avec $\varepsilon \sim N(0, \sigma)$



L'écart-type σ représente à peu près le quart de l'épaisseur du nuage

Estimation de a, b et σ

- Estimation de a et b :

$$\hat{a} = \text{estimation de } a$$

$$\hat{b} = \text{estimation de } b$$

- Estimation de σ :

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \text{ estimation de } \sigma^2$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \text{ estimation de } \sigma$$

Prévision de Y

- Modèle : $Y = aX + b + \varepsilon$, avec $\varepsilon \sim N(0, \sigma)$

  $\mu_x = E(Y \mid X = x) = ax + b$

- Problème 1 : *Calculer une estimation et un intervalle de confiance au niveau de confiance 95 % de la moyenne μ_x de Y lorsque X est fixé à x.*

- Soit y une future valeur de Y pour X fixé à x.

- Problème 2 : *Calculer une prévision et un intervalle contenant 95 % des futures valeurs de Y lorsque X est fixé à x.*

Résultat pour μ_x

- Estimation de $\mu_x = E(Y \mid X=x)$:

$$\hat{\mu}_x = \hat{a}x + \hat{b}$$

- Intervalle de confiance de μ_x au niveau 95 % :

$$\hat{y} \pm t_{0.975}(n-2) \times \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Formule approchée :

$$\hat{y} \pm 2\frac{\hat{\sigma}}{\sqrt{n}}$$

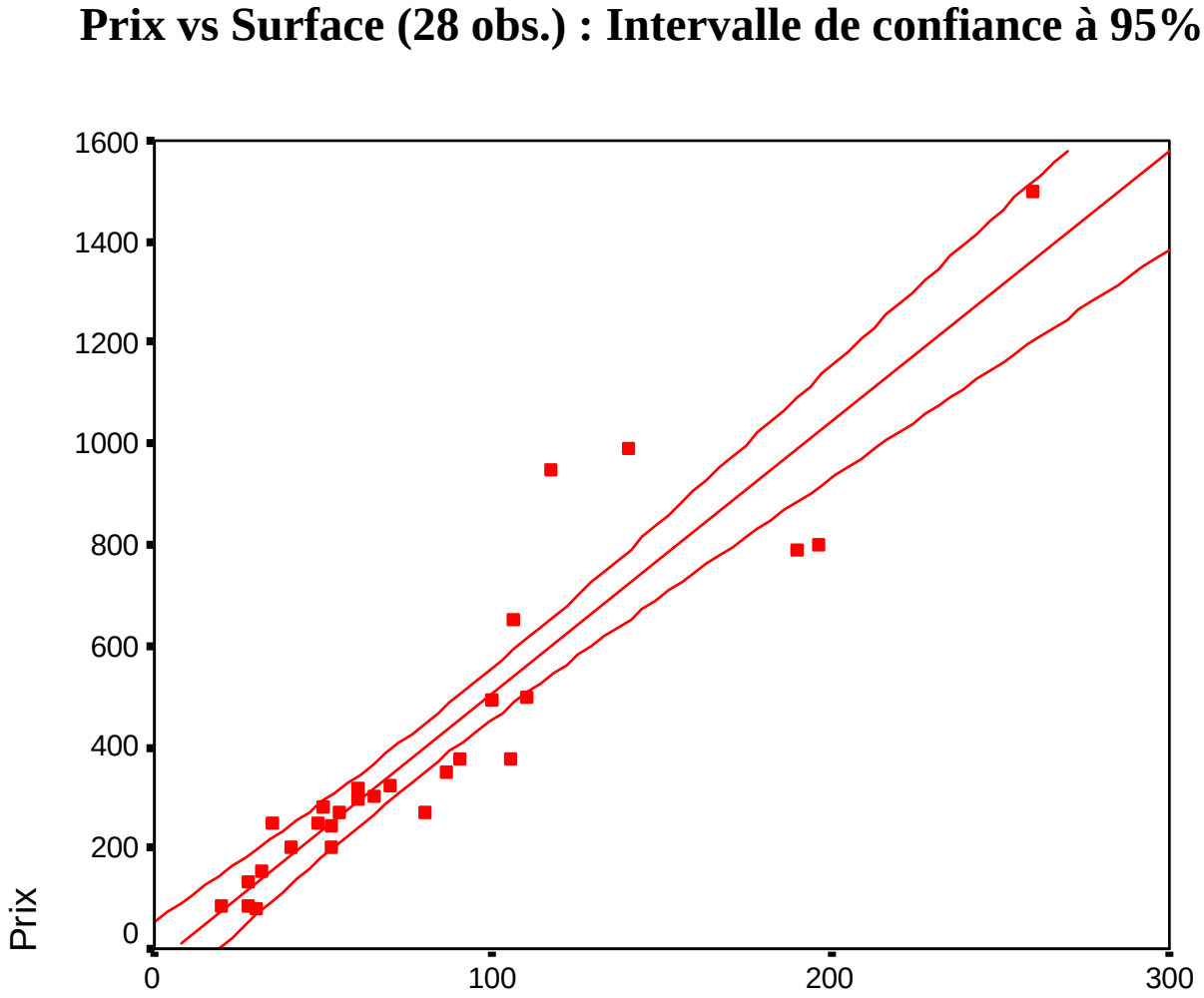
Résultats SPSS

Surface moyenne = 82.32

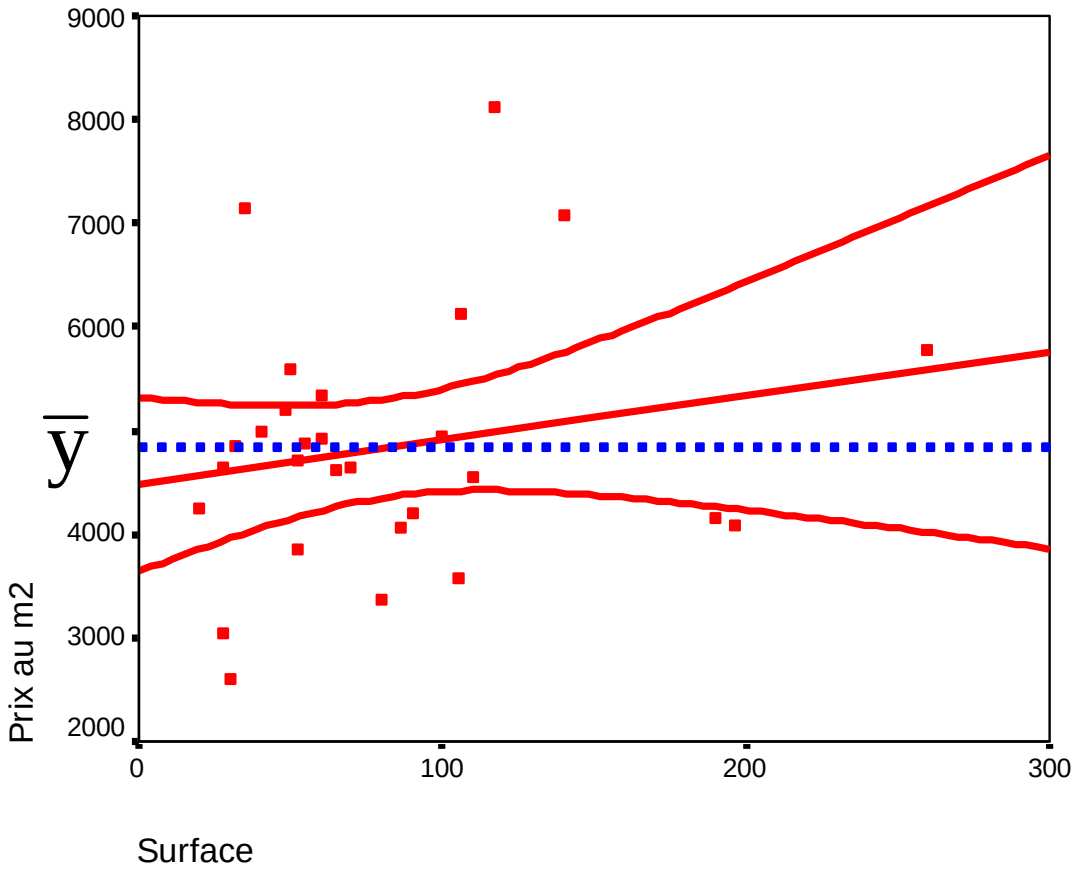
Variance de la surface = 3266.3

Numéro	Localisation	Surface	Prix (en milliers €)	Prédiction	INF95% Moyenne	SUP95% Moyenne	INF95% Individu	SUP95% Individu
1	censier	28	130	120.42	53.96	186.89	-140.88	381.72
2	contrescarpe	50	280	238.19	183.08	293.30	-20.45	496.84
3	rue saint-simon	106	650	537.97	486.14	589.80	280.00	795.94
4	rapp	196	800	1019.75	911.87	1127.64	744.99	1294.52
5	saint-andré des arts	55	268	264.96	211.84	318.07	6.73	523.18
6	5-ième, près quais	190	790	987.64	884.31	1090.96	714.62	1260.65
7	gobelins	110	500	559.38	506.13	612.63	301.13	817.64
8	gobelins	60	320	291.72	240.33	343.12	33.85	549.60
9	censier	48	250	227.49	171.51	283.47	-31.34	486.32
10	panthéon	35	250	157.89	95.43	220.36	-102.42	418.21
11	rue madame	86	350	430.91	383.05	478.77	173.71	688.10
12	rue de seine	65	300	318.49	268.51	368.47	60.89	576.09
13	panthéon	32	155	141.84	77.69	205.98	-118.88	402.55
14	sèvres-babylone	52	245	248.90	194.62	303.18	-9.57	507.37
15	montparnasse	40	200	184.66	124.85	244.47	-75.03	444.35
16	rue d'assas	260	1500	1362.36	1203.80	1520.91	1064.03	1660.69
17	saint-germain	70	325	345.26	296.36	394.15	87.86	602.65
18	ile saint-louis	117	950	596.85	540.72	652.99	337.99	855.72
19	jussieu	90	378	452.32	404.12	500.52	195.06	709.58
20	quartier-latin	30	78	131.13	65.84	196.42	-129.87	392.13
21	montparnasse	105	375	532.62	481.11	584.12	274.72	790.52
22	rue mazarine	52	200	248.90	194.62	303.18	-9.57	507.37
23	censier	80	270	398.79	350.99	446.58	141.60	655.97
24	assas luxembourg	60	295	291.72	240.33	343.12	33.85	549.60
25	jardins de l'observatoire	140	990	719.98	651.50	788.46	458.16	981.80
26	rue de savoie	20	85	77.60	6.23	148.96	-184.99	340.19
27	près luxembourg	100	495	505.85	455.78	555.92	248.23	763.47
28	gobelins	28	85	120.42	53.96	186.89	-140.88	381.72

Résultat graphique pour les intervalles de confiance



Intervalle de confiance de la moyenne
$\mu_x = ax + b$ pour une liaison non significative



La droite y = $\overline{y}$ appartient à la zone de confiance des Y moyens.
Donc la liaison entre Y et X n'est pas significative.

Résultat pour y

- Prévision de y pour x fixé :

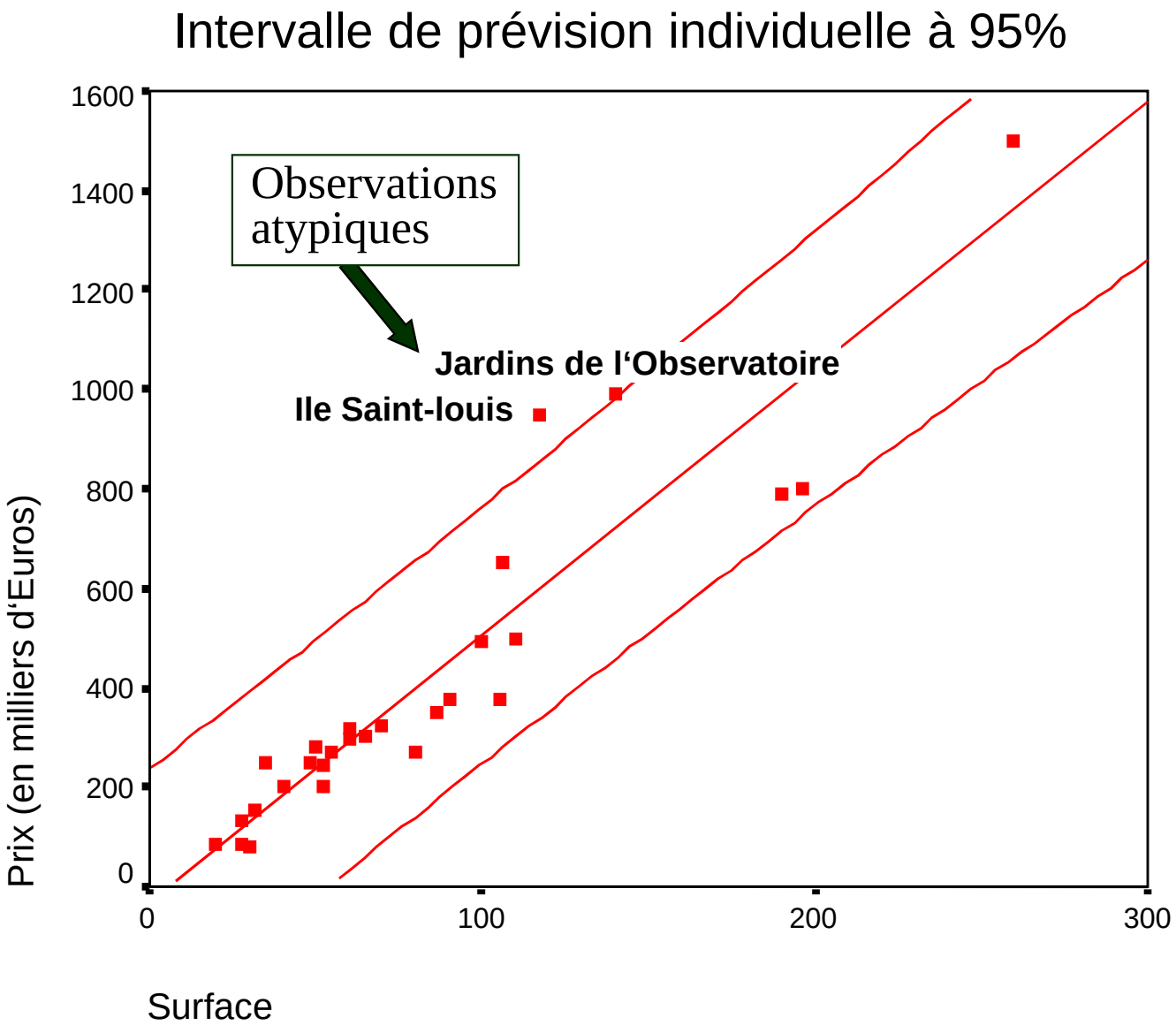
$$\hat{y} = \hat{a}x + \hat{b}$$

- Intervalle de prévision de y à 95 % pour x fixé :

$$\hat{y} \pm t_{0.975}(n-2) \times \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Formule approchée : $\hat{y} \pm 2\hat{\sigma}$

Résultat graphique pour les intervalles de prévision



Observation atypique

- Une observation est atypique (*outlier*) si elle n'appartient pas à son propre intervalle de prévision :

$$y \notin [\hat{y} \pm t_{0.975}(n-2) \times \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}]$$

- En utilisant la formule approchée :

$$y \notin [\hat{y} \pm 2\hat{\sigma}] \quad \text{soit} \quad |e| = |y - \hat{y}| > 2\hat{\sigma}$$

- <u>Conclusion :</u> Une observation i est un *outlier* si son résidu standardisé $\frac{e_i}{\hat{\sigma}}$ est supérieur à 2 en valeur absolue.

Résultats SPSS

NUMERO	Surface	Prix (en milliers d'€)	Prédiction	Résidu	Résidu standardisé	INF 95% Individuel	SUP 95% Individuel
1	28	130	120.42	9.58	0.08	-140.88	381.72
2	50	280	238.19	41.81	0.34	-20.45	496.84
3	106	650	537.97	112.03	0.91	280.00	795.94
4	196	800	1019.75	-219.75	-1.79	744.99	1294.52
5	55	268	264.96	3.04	0.02	6.73	523.18
6	190	790	987.64	-197.64	-1.61	714.62	1260.65
7	110	500	559.38	-59.38	-0.48	301.13	817.64
8	60	320	291.72	28.28	0.23	33.85	549.60
9	48	250	227.49	22.51	0.18	-31.34	486.32
10	35	250	157.89	92.11	0.75	-102.42	418.21
11	86	350	430.91	-80.91	-0.66	173.71	688.10
12	65	300	318.49	-18.49	-0.15	60.89	576.09
13	32	155	141.84	13.16	0.11	-118.88	402.55
14	52	245	248.90	-3.90	-0.03	-9.57	507.37
15	40	200	184.66	15.34	0.12	-75.03	444.35
16	260	1500	1362.36	137.64	1.12	1064.03	1660.69
17	70	325	345.26	-20.26	-0.16	87.86	602.65
18	117	950	596.85	353.15	2.87	337.99	855.72
19	90	378	452.32	-74.32	-0.60	195.06	709.58
20	30	78	131.13	-53.13	-0.43	-129.87	392.13
21	105	375	532.62	-157.62	-1.28	274.72	790.52
22	52	200	248.90	-48.90	-0.40	-9.57	507.37
23	80	270	398.79	-128.79	-1.05	141.60	655.97
24	60	295	291.72	3.28	0.03	33.85	549.60
25	140	990	719.98	270.02	2.20	458.16	981.80
26	20	85	77.60	7.40	0.06	-184.99	340.19
27	100	495	505.85	-10.85	-0.09	248.23	763.47
28	28	85	120.42	-35.42	-0.29	-140.88	381.72

Élimination des observations atypiques

- Pour rendre la prévision plus opérationnelle, on peut restreindre le champ d'application du modèle en éliminant des observations atypiques, mais en le justifiant par des considérations extra-statistiques.
- Compléter le tableau suivant jusqu'à élimination de toutes les données atypiques

Nombre d'observations	Corrélation R	Écart-type du résidu	Observations atypiques	Intervalle de prévision à 95 % du prix d'un 100 m^2
28				
26				
⋮				