Séances 9

Liaison entre variables
la régression linéaire simple

Introduction : différence entre corrélation linéaire et régression linéaire

Corrélation linéaire :

- **On cherche à savoir s'il existe un lien entre 2 variables quantitatives X et Y et à en mesurer l'intensité.**
- **Les deux variables jouent un rôle symétrique (on peut permuter X et Y) ou asymétrique.**

Dans une liaison symétrique :

- X est liée à Y, et Y est liée à X;
- Mais X ne dépend pas de Y et Y ne dépend pas de X
- On peut permuter X et Y en abscisses et en ordonnées

Régression linéaire :

- **Etude de l'influence d'une variable X sur une variable Y**
- **Les deux variables jouent un rôle asymétrique uniquement (on ne peut pas permuter X et Y) :**
  - **X = variables explicative / Y = variable expliquée**
- **On pourra éventuellement chercher à prédire Y en fonction de X**

Liaison asymétrique :

- Ex : la taille dépend de l'âge mais l'âge ne dépend pas de la taille,
- On ne peut pas peut pas permuter X et Y en abscisses et en ordonnées.

Introduction : différence entre corrélation linéaire et régression linéaire

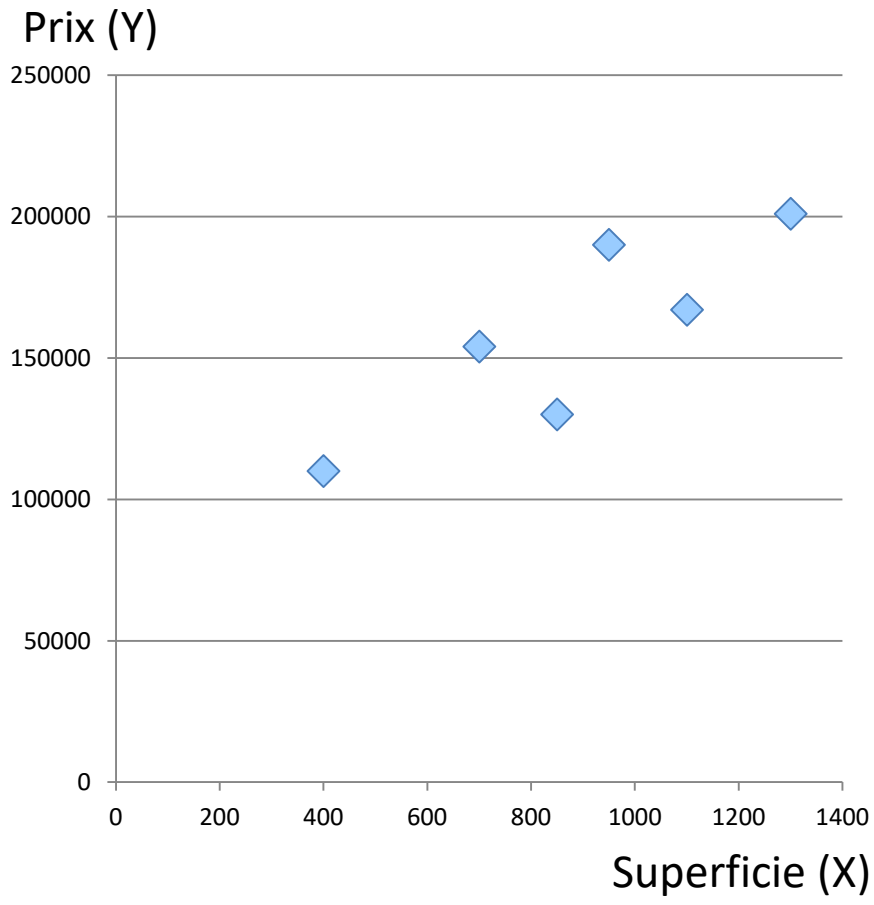
	Corrélation	Régression
Variables	X = quantitative Y = quantitative	X = quantitative Y = quantitative
Symétrie de la liaison	Oui/Non Y liée à X X liée à Y	Non Y dépend de X
Exemples	Y = vente de lunettes en été X = vente de glaces en été	Y = taille X = âge
Prédiction	Non	Oui (équation)

Une régression linéaire simple consiste à **modéliser l'influence de la variable X sur la variable Y par une fonction *Y = f(X)*** que l'on peut représenter graphiquement par une droite sur un nuage de points : la **droite de régression**

Selon les disciplines X et Y portent différents noms :

X	Y
Variable explicative	Variable expliquée
Variables indépendante	Variable dépendante
Facteur	Variable-réponse
Régresseur	

On considère la relation qu'il existe dans une région donnée entre le prix de vente des terrains à bâtir et la superficie de ces terrains. On cherche à savoir dans quelle mesure varie le prix des terrains en fonction de leur superficie.



Terrain	Prix (euros)	Superficie (m^2)
A	167 000	1100
B	130 000	850
C	154 000	700
D	190 000	950
E	201 000	1300
F	110 000	400

Le prix semble croître avec la superficie mais on voudrait :

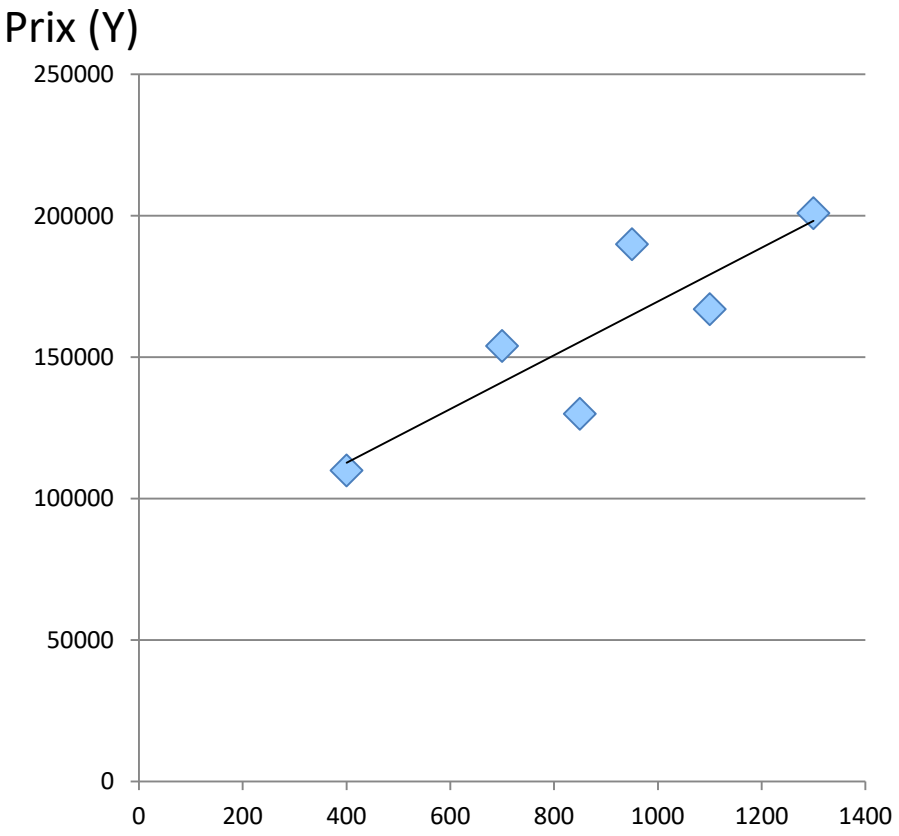
1) Disposer d'un résumé de ces données qui permette d'en décrire rapidement les caractéristiques principales et d'en dégager la structure;
2) Pouvoir inférer la relation à d'autres terrains qui ne font pas partie de l'échantillon (ex combien vaut un terrain de 500 m^2 ?)

Nous recherchons la fonction f qui représentera au mieux la relation telle que figurée sur le nuage de point

Concrètement, on cherche à déterminer l'équation de la droite qui ajuste le mieux le nuage de points, c'est-à-dire qui, étant donné une valeur particulière de X, indiquera la valeur correspondante de Y de telle sorte que cet Y estimé soit le plus proche possible de la valeur de Y observée pour tout individu i.

La forme la plus simple que l'on peut supposer est une fonction linéaire du type :

$y = ax + b$ ➡ En statistique, on notera $Y = b_0 + b_1 X$



Y : variable dépendante (expliquée)
X : variable indépendante (explicative)
b_0: ordonnée à l'origine (valeur de Y pour X = 0)
b_1 : pente (variation moyenne de la valeur de Y pour une augmentation d'une unité de X)

➡ Il s'agit donc de déterminer la valeur des paramètres b_0 et b_1 qui permettent de construire cette droite

Pour la plupart des phénomènes que l'on étudie (*a fortiori* pour les comportements humains), une fonction déterministe est insuffisante.

Exemple de fonction déterministe : les sommes encaissées varient de manière déterministe en fonction du prix de l'article et de la quantité d'articles vendus. Connaissant le prix et les quantités vendues, on connaîtra exactement les sommes encaissées.

Exemple de fonction statistique (non déterministe) : si l'on sait que les dépenses de loisirs d'un ménage dépendent de ses revenus. Connaissant les revenus on ne pourra pas prévoir exactement ses dépenses de loisirs car d'autres facteurs rentrent en jeu.

Pour rendre compte de cette incertitude on introduit dans le modèle statistique un terme d'erreur aléatoire que l'on appelle également « résidus ». Ce terme constitue la composante aléatoire du phénomène et il représente l'ensemble des facteurs inconnus ou non pris en compte.

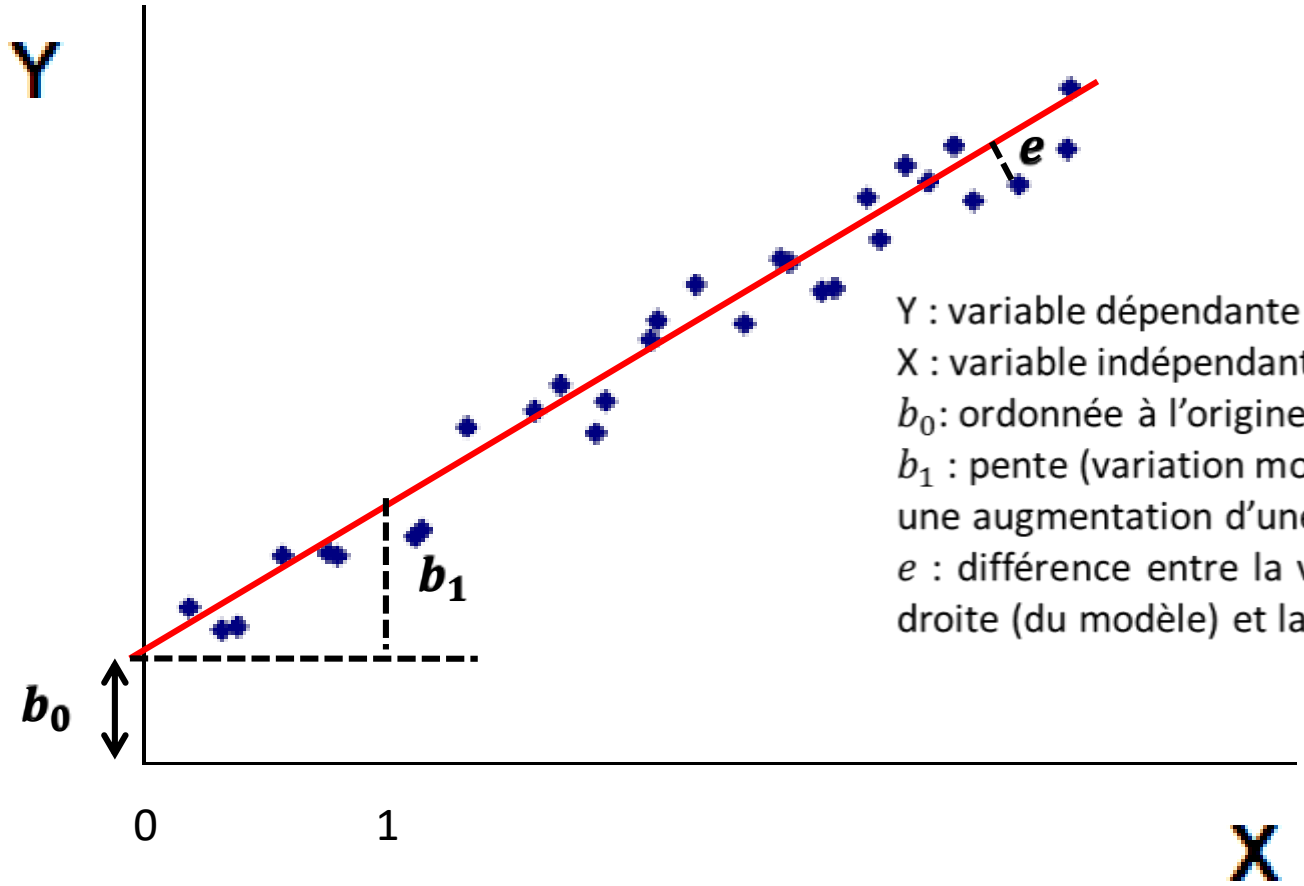
$$Y = b_0 + b_1 X + e$$

On peut donc noter :

$\hat{Y} = b_0 + b_1 X$ Où $\hat{Y}$ est la valeur prédite à partir du modèle

$Y = \hat{Y} + e$ Où Y est la valeur observée

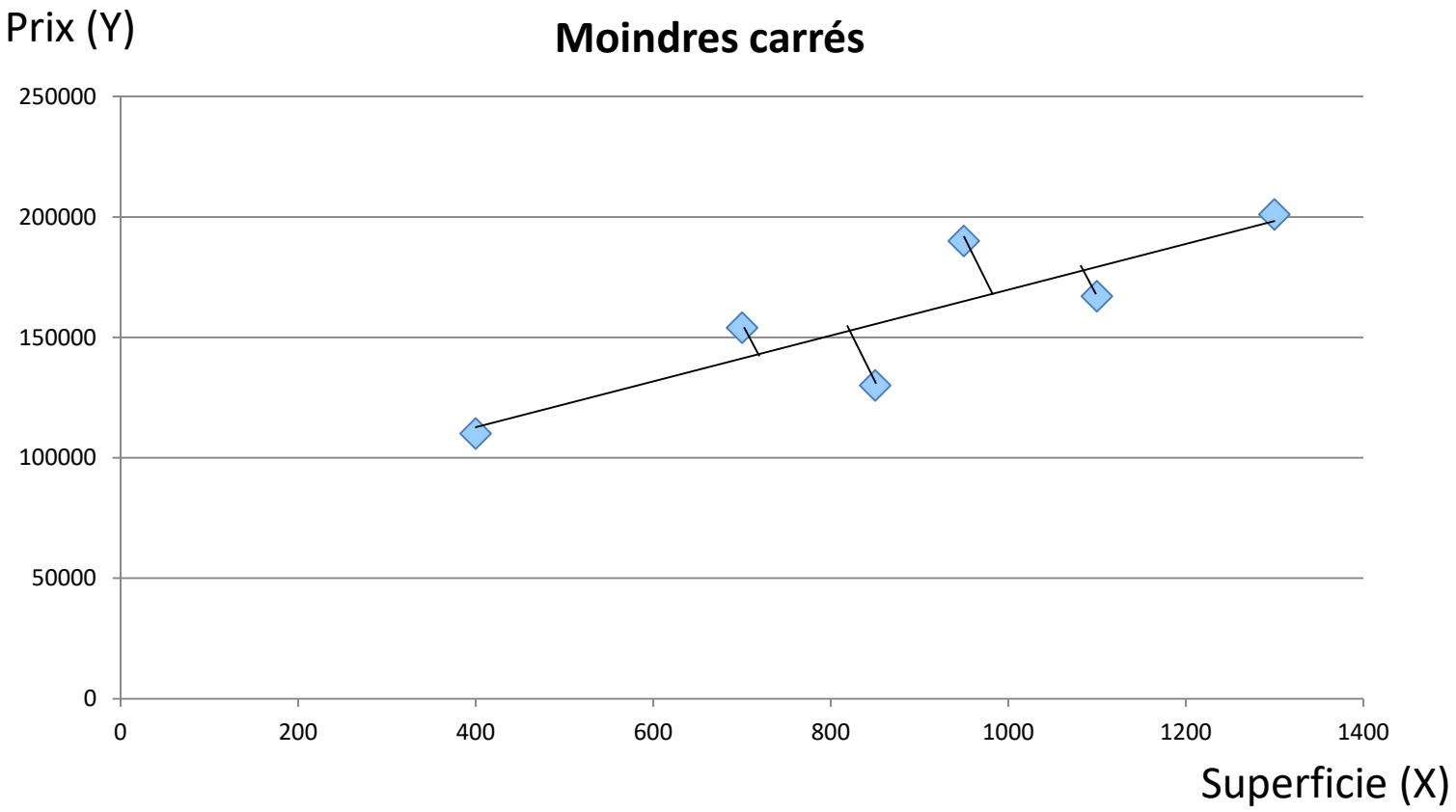
Interprétation graphique des paramètres de la droite de régression



Y : variable dépendante (expliquée)
X : variable indépendante (explicative)
b_0: ordonnée à l'origine (valeur de Y pour X = 0)
b_1 : pente (variation moyenne de la valeur de Y pour une augmentation d'une unité de X)
e : différence entre la valeur estimée à partir de la droite (du modèle) et la valeur observée.

Pour calculer la droite de régression qui est la plus « proche » des points, on recourt en général à la méthode des moindres carrés ordinaires (MCO).

Cette méthode consiste à minimiser pour tous les individus les écarts entre la valeur observée et la valeur prédite grâce à l'équation de la droite. Plus précisément, on minimise la somme des distances au carré entre les points et la droite.



Estimation des paramètres par les MCO

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \Rightarrow b_1 = \frac{cov\,(X,Y)}{Var(X)}$$

$\sum$ Est une somme

Moyenne de X $\bar{X}$

Moyenne de Y $\bar{Y}$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$cov(X,Y) = 7811111$

$\acute{X} = 883$

$\acute{Y} = 158667$

var $(X) = 82222$

$$b_1 = \frac{cov\,(X,Y)}{Var(X)} = \frac{7811111}{82222} = 95$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$= 158667 - 95*883 = 74750$$

Equation de la droite de régression par les MCO

$$PRIX = 74750 + 95 * SUPERFICIE + e$$

Equation de la droite de régression par les MCO

$$PRIX = 74750 + 95 * SUPERFICIE + e$$

Pour β_1 :

Quand la superficie d'un terrain augmente d'1 m², son prix augmente de 95 euros

Pour β_0 :

L'interprétation de ce coefficient n'est pas toujours pertinente. Ici, il correspond au prix d'un terrain de superficie nulle : **74750 euros** pour si peu de place, c'est un peu cher!

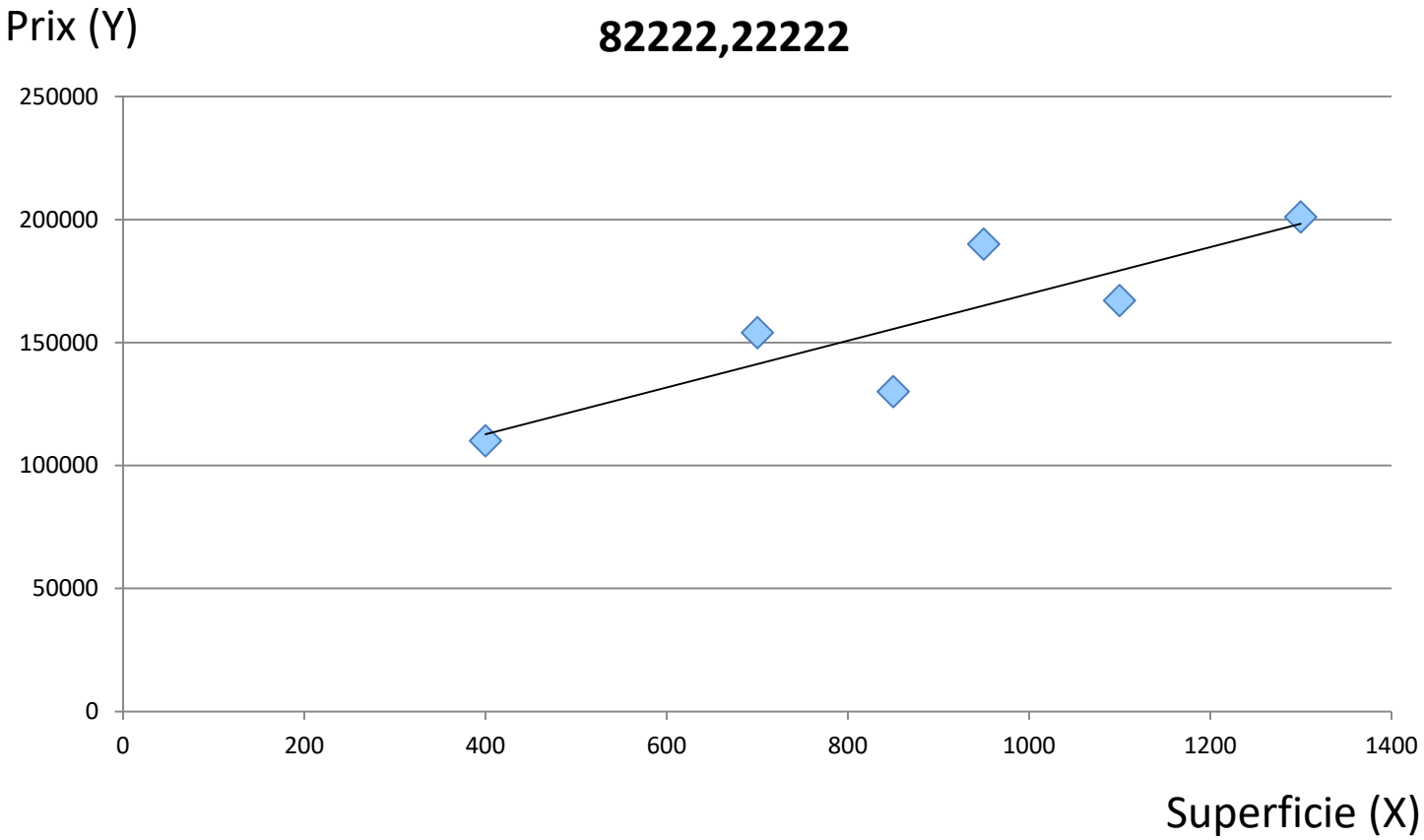
On peut aussi estimer le prix le plus probable d'un terrain de 500 m² par exemple:

74750+95*500 = 122 250 euros

Pour construire une droite, il faut au moins deux points. Ces deux points doivent être éloignés

Valeur de Y pour X = 0; $Y = 74750 + 95 * 0 = 74750$

Valeur de Y pour X = 1300; $Y = 74750 + 95 * 1300 = 198250$



Terrain	Prix (Y)	Superficie (X)	Prix estimés ()	Résidus (e)
A	167000	1100	74750+95*1100=179250	167000-179250= -12250
B	130000	850	74750+95*850= 155500	130000-155500=-25500
C	154000	700	141250	12750
D	190000	950	165000	25000
E	201000	1300	198250	2750
F	110000	400	112750	-2750

Calcul des prix prédits par le modèle

$$\hat{Y} = 74750 + 95 * X$$

Calcul des résidus : différence entre les valeurs observées et les valeurs prédites par le modèle

$$e = Y - \hat{Y}$$

Y = prix observés

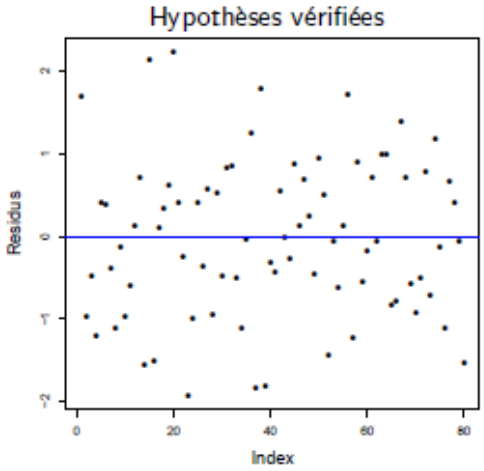
$\hat{Y}$ = prix calculés à partir du modèle

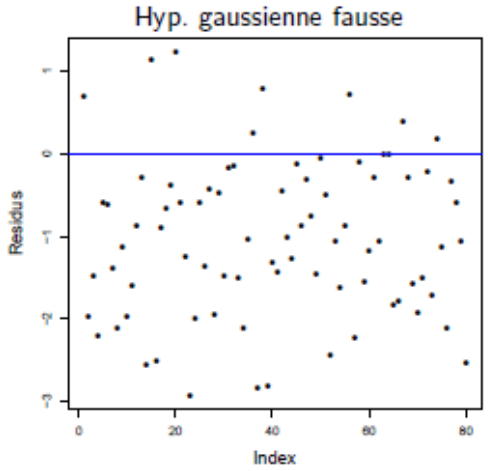
3 éléments à vérifier sur les résidus pour que les hypothèses de la régression soient respectées :

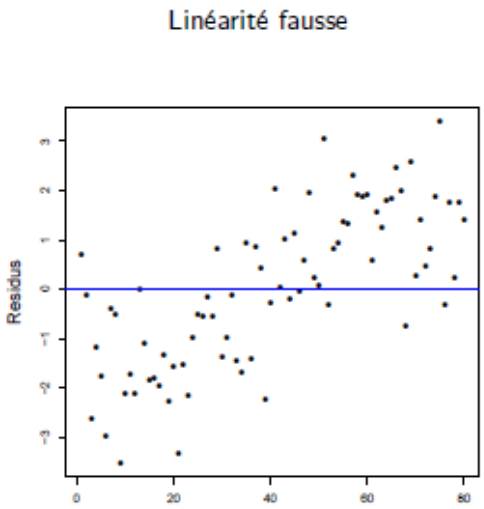
- Les résidus doivent présenter une distribution normale
- Leur variance doit être constante
- La relation doit être linéaire

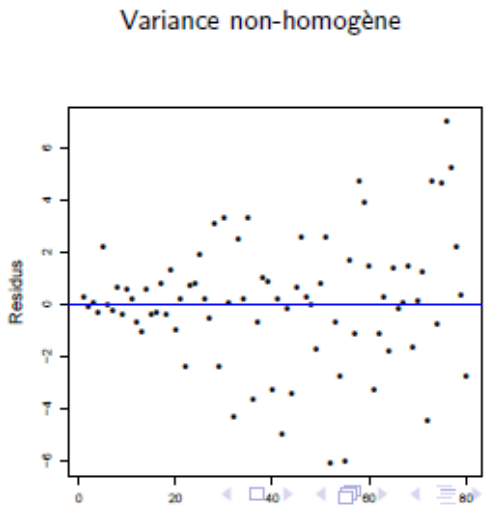
On peut vérifier ces hypothèses en plaçant sur l'axe des ordonnées les résidus et sur l'axe des abscisses les valeurs de X.

Le graphique ne doit présenter aucune structure particulière comme sur la premier graphique.









Le coefficient de détermination (R^2) nous donne la variance de Y expliquée par X. C'est une mesure du pouvoir explicatif du modèle.

Y_i

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Moyenne de Y

$\bar{Y} = 158667$

Prix Y	Superficie (X)	Prix estimés	Résidus (e)	Carré de la Somme des différences entre le Prix Y et Prix Estimé	Carré de la somme des différences entre le Prix Y et la moyenne des Prix $\bar{Y}$
167000	1100	179250	12250	(167000-179250)²=150062500	(167000-158667)²=69444444
130000	850	155500	-25500	650250000	821777778
154000	700	141250	12750	162562500	21777778
190000	950	165000	25000	625000000	981777778
201000	1300	198250	2750	7562500	1792111111
110000	400	112750	-2750	7562500	2368444444
				1603000000	**6055333333**

$$R^2 = 1 - \frac{1603000000}{6055333333} = 0{,}73$$

$$= 0{,}73$$

La surface explique 73 % de la variance du prix des terrains à bâtir dans la région observée

Test de Fischer-Sdenecor : permet de vérifier la significativité globale du modèle et de dire si l'influence de X sur Y est significative. On l'utilise lorsque l'on travaille sur un échantillon et que l'on souhaite généraliser la relation à l'ensemble de la population

Démarche :

1) On pose les hypothèses :

H_0 (hypothèse nulle): b1= 0, X n'a aucun influence sur Y

H_1 (hypothèse alternative): b1≠ 0, X a une influence sur Y

2) On choisit un risque d'erreur

$risque\ d'erreur: \alpha = 5\ \%$

3) On lit la valeur théorique dans la table de Fischer

V1= nombre de variable explicatives (ici 1) : on lit la première colonne

V2=N(taille de la population) -2 = 6-2 = 4 : on lit la 4ème ligne

4) On calcule la statistique F de Fischer

$$F = \frac{\left(\frac{r^2}{k}\right)}{\left(\frac{(1-r^2)}{(N-k-1)}\right)}$$

r²= coefficient de détermination

k= nombre de variables explicatives

N= taille de la population

5) On confronte F à la valeur lue dans la table

Si F > valeur lue dans la table : on rejette H_0, le modèle est significatif à 95 % , il existe une relation en Y et X

Table de Fisher-Snedecor, α = 5 %

	v1				
v2	1	2	3	4	5
1	161,44	199,50	215,70	224,58	230,16
2	18,51	19,00	19,16	19,24	19,29
3	10,12	9,55	9,27	9,11	9,01
4	7,71	6,94	6,59	6,39	6,26
5	6,61	5,79	5,41	5,19	5,05
6	5,99	5,14	4,76	4,53	4,39
7	5,59	4,74	4,35	4,12	3,97
8	5,32	4,46	4,07	3,84	3,69
9	5,12	4,26	3,86	3,63	3,48
10	4,96	4,10	3,71	3,48	3,33
11	4,84	3,98	3,59	3,36	3,2
12	4,75	3,89	3,49	3,26	3,11
13	4,67	3,81	3,41	3,18	3,02
14	4,60	3,74	3,34	3,11	2,96
15	4,54	3,68	3,29	3,06	2,9
16	4,49	3,63	3,24	3,01	2,85
17	4,45	3,59	3,20	2,96	2,81
18	4,41	3,56	3,16	2,93	2,77
19	**4,38**	3,52	3,13	2,9	2,74
20	4,35	3,49	3,10	2,87	2,71
21	4,32	3,47	3,07	2,84	2,68

$$F = \frac{\left(\frac{r^2}{k}\right)}{\left(\frac{(1-r^2)}{(N-k-1)}\right)}$$

$$= \frac{\left(\frac{0,73}{1}\right)}{\left(\frac{(1-0,73)}{(6-1-1)}\right)}$$

F= 10,8

F est inférieure à la valeur lue dans la table (7,71). On ne peut pas rejeter l'hypothèse H_0. L'effet de la surface sur les prix n'est pas significatif au risque de 5 %. On ne peut pas généraliser le résultat à l'ensemble des terrains à bâtir de la région étudiée.

Reprenez le tableau de la séance dernière

1) Poser l'hypothèse
2) Calculer la droite de régression, placer la sur le nuage de points et interpréter les paramètres de cette droite.
3) Calculer les valeurs prédites par le modèle puis les résidus. Commenter ces résultats.
4) Faites le test de Fisher. Qu'en déduisez-vous ?

EXERCICE

Pays	Conso (Y)	PNB (X)	Conso prédite	Résidus (e)
All	4128	27620	5298,95	-1171,05
Aut	3301	26730	5135,19	-1834,29
Bel	5120	24820	4783,75	335,85
Dan	3977	30300	5792,07	-1815,17
Esp	2458	13740	2745,03	-287,33
Fin	5997	20740	4033,03	1964,17
Fra	4042	25060	4827,91	-785,61
Gre	2260	8180	1721,99	537,71
Irl	3136	15050	2986,07	150,33
Ita	2707	19200	3749,67	-1043,07
Lux	9356	33480	6377,19	2979,21
P-B	4580	23700	4577,67	2,03
Por	1827	9780	2016,39	-189,49
R-U	3772	18950	3703,67	68,43
Suè	5723	23120	4470,95	1251,65

Certains pays présentent des résidus négatifs. Les résultats prédits par le modèle surestime leur consommation réelle. Il s'agit souvent de pays du sud de l'Europe bénéficiant d'un climat méditerranéen qui réduit la facture énergétique.

D'autres pays présentent des résidus positifs, le modèle sous-estime leur consommation. La Suède et la Finlande font par exemple partie de ce groupe.