170 Régression sur un critère ordonné

170.1 Utilité

La régression est un modèle qui permet de qualifier et de quantifier la relation entre une variable Y, mesurée dans l'expérience, et une variable quantitative X qui détermine les conditions de l'expérience.

X serait par exemple le temps, la latitude, la profondeur, le pH, des catégories de revenus…

Nous le voyons ici dans le contexte de l'ANOVA I, mais la régression s'intégrer à des analyses à plusieurs critères, pour autant qu'elle concerne un critère fixe ordonné.

La régression permet de bénéficier de la puissance maximale de l'analyse en prenant en compte la variance globale (CMR) mesurée dans l'ensemble des échantillons et de l'information suivant laquelle on attend $\mu 1 > \mu 2 > \mu 3 > \mu 4$ ou $\mu 1 < \mu 2 < \mu 3 < \mu 4$.

Ceci limite le nombre d'hypothèses nulles que l'on peut énoncer par rapport au modèle non ordonné.

$$H_1 : \mu_1 < \mu_2 = \mu_3$$
$$H_2 : \mu_1 = \mu_2 < \mu_3$$
$$H_3 : \mu_1 = \mu_3 > \mu_2$$
$$H_4 : \mu_1 < \mu_2 < \mu_3$$

Le modèle de régression combiné à l'ANOVA permet également de répondre quantitativement à la question de savoir si la régression peut être considérée comme linéaire, ou non. Démontrer qu'une relation est non linéaire est important pour limiter les erreurs que l'on peut faire dans les estimations basées sur ce modèle, particulièrement en extrapolation.

170.2 Hypothèses

Un gestionnaire communal cherche à mettre en évidence la relation entre la production de déchets domestiques (Y) et le revenu des ménages (X). Il constitue 4 catégories de revenus annuels croissants et suit la production de déchets de 8 ménages pris au hasard dans chaque catégorie de revenus.

Le revenu annuel constitue un critère fixe à 4 niveaux, les ménages de chaque catégorie sont indépendants, le modèle correspond bien à celui d'une ANOVA I fixe avec 8 réplicats par niveau du critère fixe.

Ce modèle s'écrit :

$y_{(i)j} = \mu + a_i + E_{(i)j}$ avec δ^2_i qui représente la dispersion des a_i

Sous Ho : $\delta^2_i = 0$ ce qui s'exprime en termes de moyennes :

Ho : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

μ_i représentant les moyennes de chaque temps.

L'hypothèse alternative est une évolution de la production de déchets en fonction du revenu, suivant un modèle linéaire ou non linéaire, de pente positive ou négative suivant que le revenu augmente la consommation et la production de déchet (H1), ou diminue le gaspillage et la production de déchet (H2).

$H_1 : \mu_1 < \mu_2 < \mu_3 < \mu_4$.
$H_2 : \mu_1 > \mu_2 > \mu_3 > \mu_4$.

170.3 Modèles

La prise en compte du modèle de régression donne un test plus puissant que celui de l'ANOVA simple, qui teste globalement toutes les alternatives.

Si Ho est vraie, le modèle est $y_{(i)j} = \mu + E_{(i)j}$

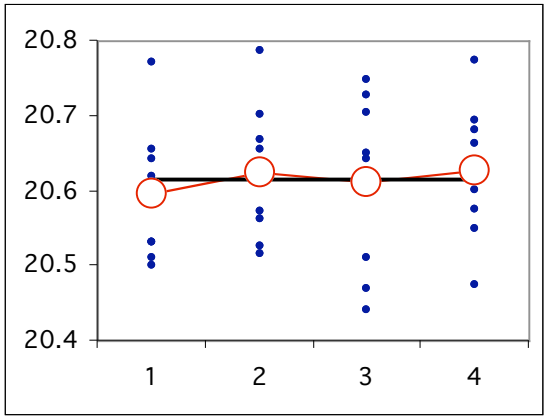

Figure 170-1 Représentation d'une croissance nulle des déchets en fonction des revenus annuels. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) et moyenne générale (traits).

Si la croissance suit une régression linéaire, le modèle est le suivant :

$$y_{(i)j} = \beta o + \beta 1 X_i + E_{(i)j} \quad \textbf{Équation 170-1}$$

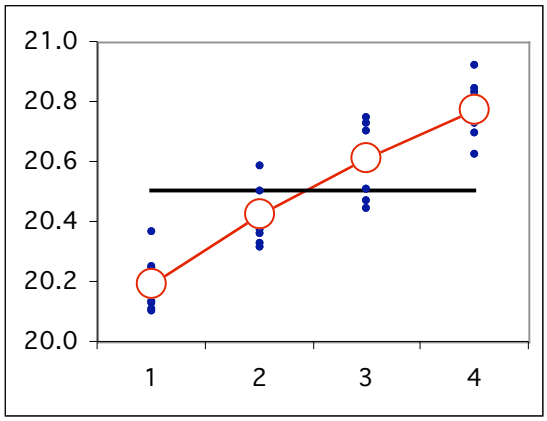

Figure 170-2 Représentation d'une croissance linéaire des déchets en fonction des revenus annuels. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) et moyenne générale (traits).

Si la croissance suit une régression non linéaire, le modèle est le suivant :

$$y_{(i)j} = \beta o + \beta 1 X_i + \omega_i + E_{(i)j} \quad \textbf{Équation 170-2}$$

Les constantes ω_i représentent les écarts entre les moyennes réelles et leur position théorique sur une droite de régression.

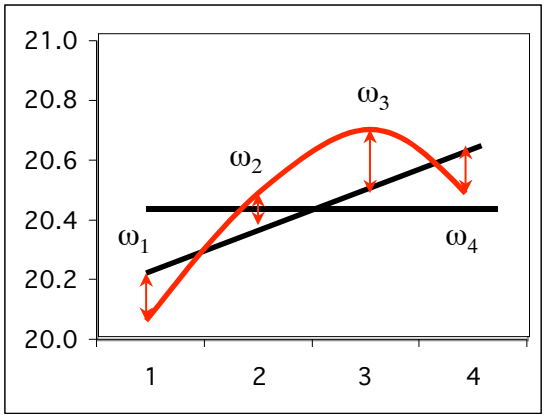

Figure 170-3 Représentation d'une croissance non linéaire théorique des déchets en fonction des revenus annuels. Fonction théorique (trait courbe), droite de régression (trait oblique) et moyenne générale (trait horizontal). Les constantes ω_i représentent les écarts entre les moyennes réelles et leur position théorique sur une droite de régression.

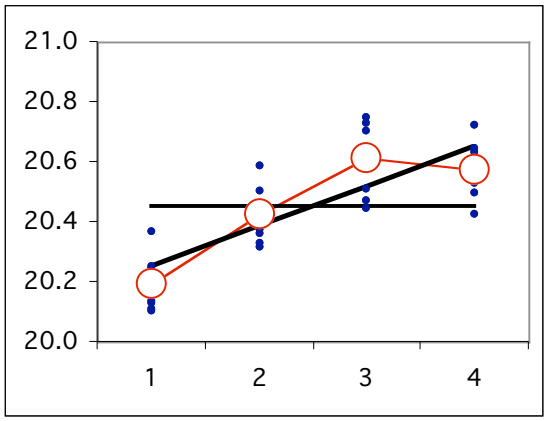

Figure 170-4 Représentation d'une croissance non linéaire observée des déchets en fonction des revenus annuels. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches), droite de régression (trait oblique) et moyenne générale (trait horizontal).

170.4 Equations

Différents types d'écarts sont pris en considération pour établir les équations. Pour la clarté de la représentation schématique, nous n'envisageons qu'une seule valeur individuelle, une seule tranche de revenus, la droite et la moyenne générale.

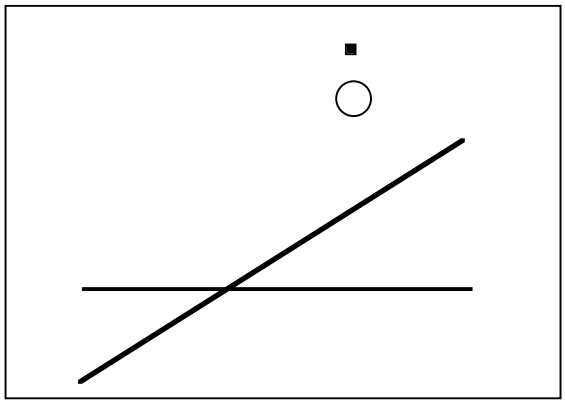

Figure 170-5 Représentation schématique d'une croissance des déchets en fonction des revenus annuels : valeur individuelle (carré), moyenne d'une tranche de revenus (rond), droite de régression (trait oblique) et moyenne générale (trait horizontal).

L'équation de l'ANOVA décompose l'écart total (1) en deux parties additives : l'écart factoriel (2) et l'écart résiduel à la moyenne (3).

$Y_{(i)j} - My = (My_i - My) + (Y_{(i)j} - My_i)$

Soit, sous forme de SCE :

SCET = SCEF +SCER

> *Les écarts sont calculés entre l'observation et la moyenne générale (1), entre la moyenne d'une tranche de revenus et la moyenne générale (2), entre l'observation et la moyenne journalière (3).*

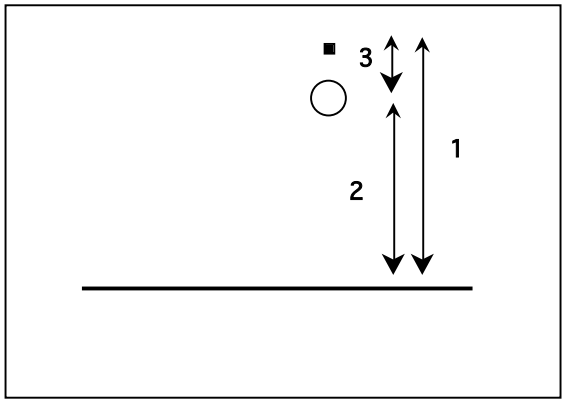

Figure 170-6 Représentation schématique de l'écart total (1) factoriel (2) et résiduel (3) entre valeur individuelle (carré), moyenne d'une tranche de revenus (rond) et moyenne générale (trait horizontal).

L'équation de la régression décompose SCEF, l'écart factoriel (1) en deux parties additives : SCEL, l'écart du modèle à la moyenne générale (2) et SCEN, l'écart de la moyenne d'une tranche de revenus au modèle linéaire (3).

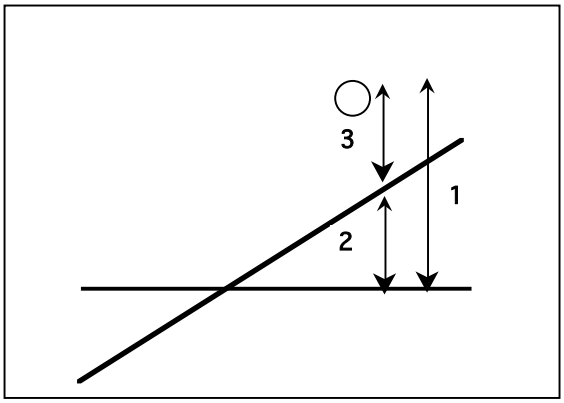

Figure 170-7 Représentation schématique de l'écart factoriel (1), l'écart expliqué par la régression linéaire (2) et l'écart non expliqué par la régression linéaire (3). Moyenne d'une tranche de revenus (rond), moyenne générale (trait horizontal), régression linéaire (trait oblique).

$My_i - My = (Y_{m\,i} - My) + (My_i - Y_{mi})$

Avec $Y_{m\,i} = B_o + B_1 x_i$

Soit, sous forme de SCE :

SCEF = SCEL + SCEN

SCEL : variabilité expliquée par la régression linéaire (2)
SCEN : variabilité non expliquée par la régression linéaire (3)

> *Les écarts sont calculés entre la moyenne d'une tranche de revenus et la moyenne générale (1), entre l'estimation de la régression linéaire au point x_i et la moyenne générale (2),et entre l'estimation de la régression linéaire au point x_i et la moyenne d'une tranche de revenus (3).*

Les deux équations donnent ensemble :

SCET = SCEL +SCEN +SCER Équation 170-3

Dans notre modèle, SCEL mesure $\beta_1 X_i$, SCEN mesure ω_i et SCER mesure $E_{(i)j}$

SCER a N - n_a degrés de liberté.

SCEF a n_a -1 degrés de liberté.

SCEL a 1 seul degré de liberté

> *Cela s'explique par le fait que la régression linéaire dépend de 2 paramètres : βo et $\beta 1$.*

Il reste pour SCEN = SCEF -SCEL n_a -2 degrés de liberté.

En termes de carrés moyens :

CML = SCEL
CMN = SCEN/ n_a -2

170.5 Tests

Deux questions se posent, dont les réponses complètent 4 scénarios :

1. les différences entre les moyennes sont -elles – au moins en partie - expliquées par une régression linéaire?

$$E(CML) = \sigma^2 + \beta_1^2 SCE_X$$

Ho : $\beta_1 = 0$

Cette hypothèse se reformule :

Ho : E(CML / CMR) = 1
Ho : CML / CMR = $F_{1;N-\nu\alpha}$

2. les différences entre les moyennes sont -elles – au moins en partie - expliquées par des écarts à la régression linéaire?

$$E(CMN) = \sigma^2 + \frac{1}{n_a - 2}\sum \omega_i^2$$

Ho : $\omega_\iota = 0$ pour tout i

Cette hypothèse se reformule :

Ho : E(CMN / CMR) = 1
Ho : CMN / CMR = $F_{\nu\alpha-2;N-\nu\alpha}$

Les quatre scénarios de réponse sont les suivants :

170.5.1 Scénario 1

AHo : $\beta 1 = 0$ et AHo : $\omega_i = 0$ pour tout i

Cela signifie qu'il n'y a aucune différence significative entre les moyennes.

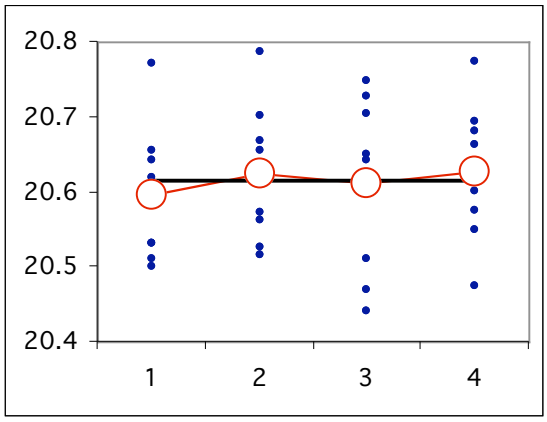

Figure 170-8 Représentation d'une évolution nulle des déchets en fonction des revenus annuels. AHo : $\beta_1 = 0$. AHo : $\omega_i = 0$ pour tout i. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) et moyenne générale (traits).

Sauf cas exceptionnel, l'ANOVA1 simple donnerait dans ce cas AHo.

170.5.2 Scénario 2

RHo : $\beta 1 = 0$ et AHo : $\omega_i = 0$ pour tout i

Cela signifie que les différences significatives entre les moyennes s'expliquent par une régression linéaire, sans écarts à la linéarité.

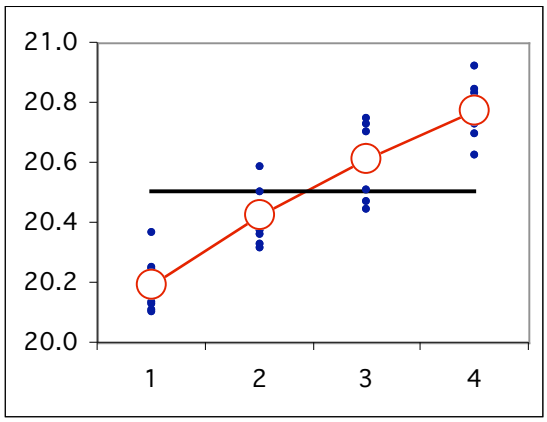

Figure 170-9 Représentation d'une croissance linéaire des déchets en fonction des revenus annuels. RHo : $\beta_1 = 0$. AHo : $\omega_i = 0$ pour tout i. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) et moyenne générale (traits).

Dans de nombreux cas l'ANOVA1 simple donnerait dans ce cas AHO, les différences entre les moyennes ne seraient pas détectées, car le test de régression est plus puissant.

170.5.3 Scénario 3

RHo : $\beta 1 = 0$ et RHo : $\omega_i = 0$ pour tout i

Cela signifie que les différences significatives entre les moyennes s'expliquent en partie par une régression linéaire de pente non nulle, mais que des écarts à la linéarité sont nécessaires pour expliquer les écarts entre les moyennes et le modèle.

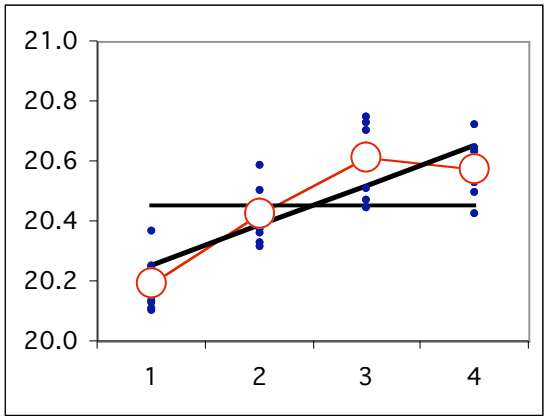

Figure 170-10 Représentation d'une croissance non linéaire, avec une tendance linéaire, des déchets en fonction des revenus annuels.
RHo : $\beta_1 = 0$. RHo : $\omega_i = 0$ pour tout i. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) droite de régression (trait oblique) et moyenne générale (trait horizontal).

Dans tous les cas l'ANOVA1 simple donnerait dans ce cas Rho, mais nous obtenons ici un test objectif de non linéarité (irréalisable s'il n'y a pas de réplicats pour chaque valeur de x_i)

170.5.4 Scénario 4

AHo : $\beta_i = 0$ et RHo : $\omega_i = 0$ pour tout i

Ceci signifie que les différences significatives entre les moyennes ne s'expliquent pas par une régression linéaire (la pente de la droite n'est pas significativement différente de 0). Ne subsistent que des écarts à la linéarité.

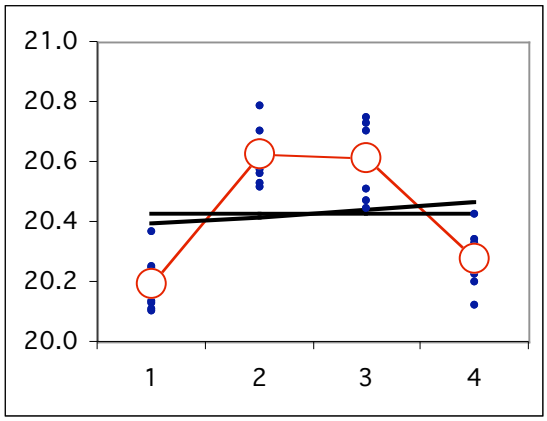

Figure 170-11 Représentation d'une croissance non linéaire sans aucune tendance linéaire des déchets en fonction des revenus annuels. AHo : $\beta_1 = 0$. RHo : $\omega_i = 0$ pour tout i. Valeurs individuelles (noires), moyennes par tranche de revenus (blanches) droite de régression (trait oblique) et moyenne générale (trait horizontal).

Dans ce cas l'analyse est équivalente à une ANOVA1 simple : la régression linéaire n'apporte aucune information et le modèle peut se simplifier :

$y_{(i)j} = \beta o + \omega_i + E_{(i)j}$

ce qui correspond terme à terme au modèle d'une ANOVA1 non ordonnée:

$X_{(i)j} = \mu + a_i + E_{(i)j}$ avec $\mu = \beta o$ et $\omega_i = a_i$

170.6 Calculs

Soient les données de poids de déchets (Kg/an/habitant) de 32 ménages prélevés aléatoirement dans une commune, classés par 8 en 4 catégories de revenus.

Les ménages des quatre catégories de revenus étant indépendants les uns les autres, le plan expérimental correspond bien à une ANOVA 1 :

Ho : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

	Revenus			
ménages	1	2	3	4
1	156	154	136	255
2	211	163	139	281
3	169	56	162	254
4	141	165	169	239
5	175	107	165	252
6	149	165	92	233
7	167	91	227	262
8	132	170	164	209

Tableau 170-1 Données de poids de déchets (Kg/an/habitant) de 32 ménages, pour 4 catégories de revenus.

L'expérimentateur cherchant une augmentation continue du poids en fonction du temps, les hypothèses alternatives ad hoc sont bien
$H_1 : \mu_1 < \mu_2 < \mu_3 < \mu_4$.
$H_2 : \mu_1 > \mu_2 > \mu_3 > \mu_4$.

Les tests à réaliser sur le modèle $y_{(i)j} = \beta_{\square} + \beta_i X_i + \omega_i + E_{(i)j}$ sont :

Ho : $\sigma^2_i = \sigma^2_2$ homogénéité des variances de $E_{(i)j}$
Ho : $\beta_i = 0$ la pente de la régression de Y en fonction de X est nulle
Ho : $\omega_i = 0$ les moyennes μ_1 , μ_2 , μ_3 , μ_4 sont situées sur une droite

Catégorie	moyennes	variances
1	162.5	598.3
2	133.9	1874.4
3	156.8	1453.1
4	248.1	459.0
générale	175.3	1096.2

Tableau 170-2 Moyennes et variances des poids de déchets de 8 ménages de 4 catégories de revenus. La variance générale (1096.2) est le CMR.

Test de Hartley : homogénéité des variances σ^2_i

$$H_{7;4} = \frac{1874.4}{459} = 4.08$$

$$H_{7;4;0.95} = 8.44$$

Il y a acceptation de Ho : homogénéité des variances σ^2_i.

SCET, SCEF et SCER se calculent comme dans l'Anova 1.

SCET	90912.9
SCEF	60219.6
SCER	30693.3

Tableau 170-3 Valeurs calculées pour les SCE (données de poids de déchets de 8 ménages de 4 catégories de revenus).

Il faut y ajouter le calcul de SCEL, et SCEN qui est SCEF -SCEL.

Le calcul de SCEL est lié à celui de la régression linéaire : il implique donc SPE et SCEx :

$$SCEL = \frac{SPE^2}{SCEx}$$

Le calcul de SPE pourrait se calculer en recréant une série statistique bivariée de 32 couples de valeurs :

X	Y
1	156
1	211
…	…
2	154
…	…
4	209

Tableau 170-4 Constitution d'une série X, Y à partir des données de poids de déchets (Y) de ménages de quatre catégories de revenus (X).

SPE et SCEX étant dès lors calculé sur les 32 valeurs de X énumérées.

Dans le tableur Excel, nous pouvons réaliser l'opération plus rapidement :

*SPE = {SOMME((X : X -Mx) *(Y : Y -My))}*
*SCEX = n *somme.carres.ecarts(X : X)*

Expressions dans lesquelles

X : X représente la plage de données contenant les valeurs de x
Ici : 1,2,3,4
Mx représente moyenne(X : X)
Ici : 2.5

Y : Y représente la plage de données contenant les valeurs de Y
Ici : 8 lignes et 4 colonnes
MY représente moyenne(Y : Y)
Ici : 175.31
N représente le nombre de réplicats
Ici : 8

Soit : SPE = 1119 et SCEx = 40

SCEL = SPE^2/SCEX = 31304.025

La table complète de l'analyse de la variance est la suivante :

	SCE	d.l.	C.M.	F	p
Total	90913	31			
Factoriel	60220	3	20073	18	9E -07
Pente	31304	1	31304	29	1E -05
Ecarts à la néarité	28916	2	14458	13	9E -05
Résiduel	30693	28	1096		

Tableau 170-5 Table de l'ANOVA 1 et test de régression (production de déchets ménangers).

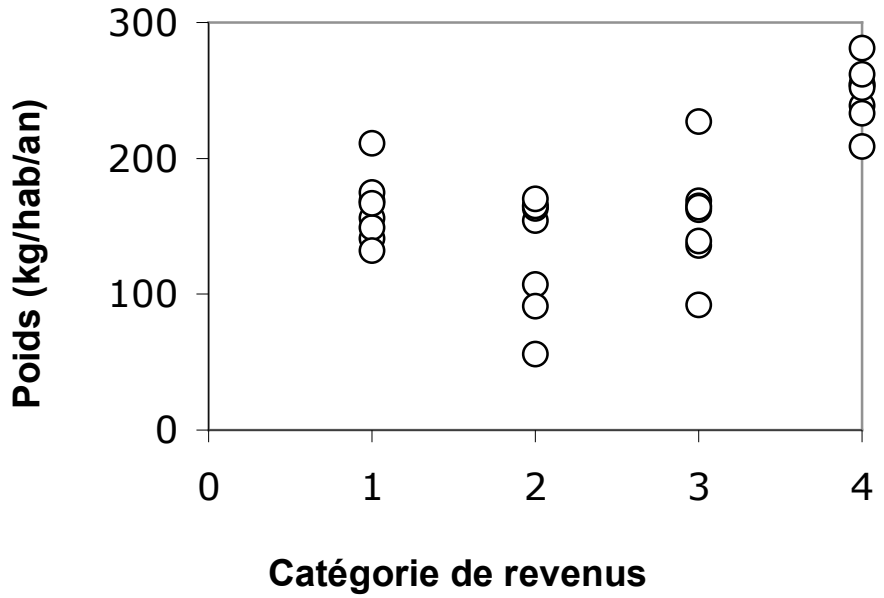

Figure 170-12 Représentation graphique des poids de déchets de 8 ménages de 4 catégories de revenus.

En conclusion, l'analyse montre sans aucun doute (α < 0.0001) que les moyennes sont différentes, que la pente de la relation linéaire entre la production de déchets et le revenu du ménage n'est pas nulle, et que cette relation n'est pas linéaire.

> *L'estimation de la pente et l'inférence sur la régression ne sont donc pas appropriées.*

170.7 Exemple

L' Institut de Conseil et d'Etudes en Développement Durable effectue une enquête en 2007 pour établir l'impact de la libéralisation de la distribution du gaz et de l'électricité sur le prix moyen au MW/h en entreprise. L'enquête différentie les entreprises suivant leur consommation moyenne et les regroupe en 4 niveaux de consommation, par tranche de 1000 MW/h. Elle répertorie ensuite le coût au kW/h facturé à 10 entreprises de chaque catégorie.

	Tranche de consommation (MWh)			
	1000	**2000**	**3000**	**4000**
Prix du Kwh (€ cents)	27,9	22,3	21,8	19,7
	26,9	21,2	20,9	18,8
	26,8	22,4	24,3	19,2
	23,4	21,7	17	19,9
	25,3	22,2	20,8	21,7
	24,9	25,8	21,8	18,4
	24,2	20,8	20,7	21,8
	24,7	22,5	18,6	20,9
	22,5	22,7	19,8	20,7
	26,8	19,7	23,7	18,3

Tableau 170-6 Prix du Kwh (€c) pour 40 entreprises réparties dans 4 catégories de consommation (données fictives).

Résultats intermédiaires :

catégories	**1000**	**2000**	**3000**	**4000**	**Global**
moyennes	25,34	22,13	20,94	19,94	22,09
variances	3,00	2,53	4,75	1,67	2,99

SCEx = 5 10^7 SPExy = -86 950

La table d'analyse est la suivante :

Effet	**SCE**	**d.l.**	**C.M.**	**F**	**p**
Total	272,62	39			
Factoriel	165,09	3	55,03		
Pente	151,21	1	151,21	50,62	0,000
Ecarts à la linéarité	13,88	2	6,94	2,32	0,112
Résiduel	107,53	36	2,99		

Tableau 170-7 Table de l'ANOVA 1 et test de régression (prix du Kwh de 40 entreprises).

En conclusion, l'analyse montre sans aucun doute (α < 0.0001) que le prix moyen du KWh décroît linéairement en fonction de la consommation. Les écarts à la relation linéaire ne sont pas significatifs (α > 0.11). Moyennant un risque d'erreur β inconnu, nous pouvons considérer que la relation est linéaire.
La pente de la régression est de -1,74 10^{-3} €c/Kwh/MWh et le coefficient de détermination est de 0,55. (SCE pente/SCE totale).

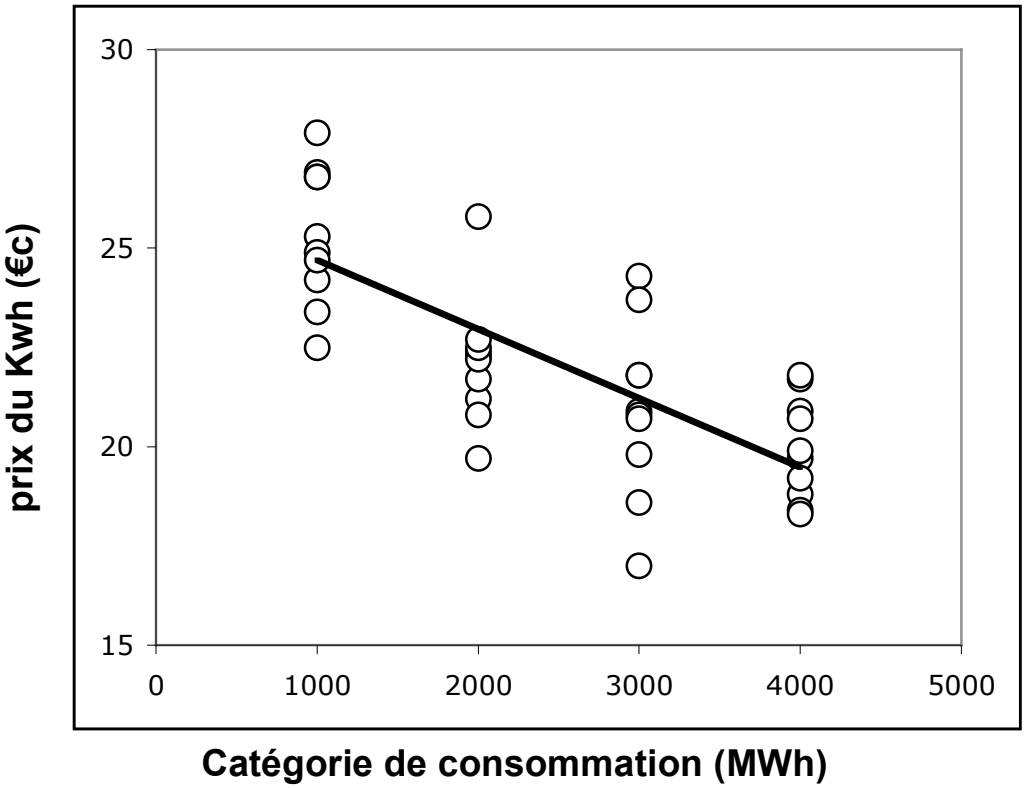

Figure 170-13 Représentation graphique de l'évolution du prix du KWh en fonction de la tranche de consommation.

Intervalle de confiance de la pente réelle β_1 :

Reprenons la formule vue pour la régression réalisée sans réplicat :

$$\varepsilon = t_{N-na;1-\alpha/2}\sqrt{\frac{CMR}{SCEx}}$$

dont les degrés de libertés sont adaptés à ceux du CMR (N -na). Nous trouvons :

$$\varepsilon = 2{,}03\sqrt{\frac{2{,}99}{5\ 10^{7}}} = 4{,}96\ 10^{-4}$$

P(-2,24 10^{-3} < β1 < -1,24 10^{-3}) = 0.95

Unités : €c/KWh/MWh ce qui signifie que le prix du KWh diminue de max 2,24 10^{-3} et min 1,24 10^{-3} €c par tranche d'augmentation de consommation de 1 MWh.