Chapitre 1

Régression linéaire simple

GALHARRET Jean-Michel

Année 2020-21

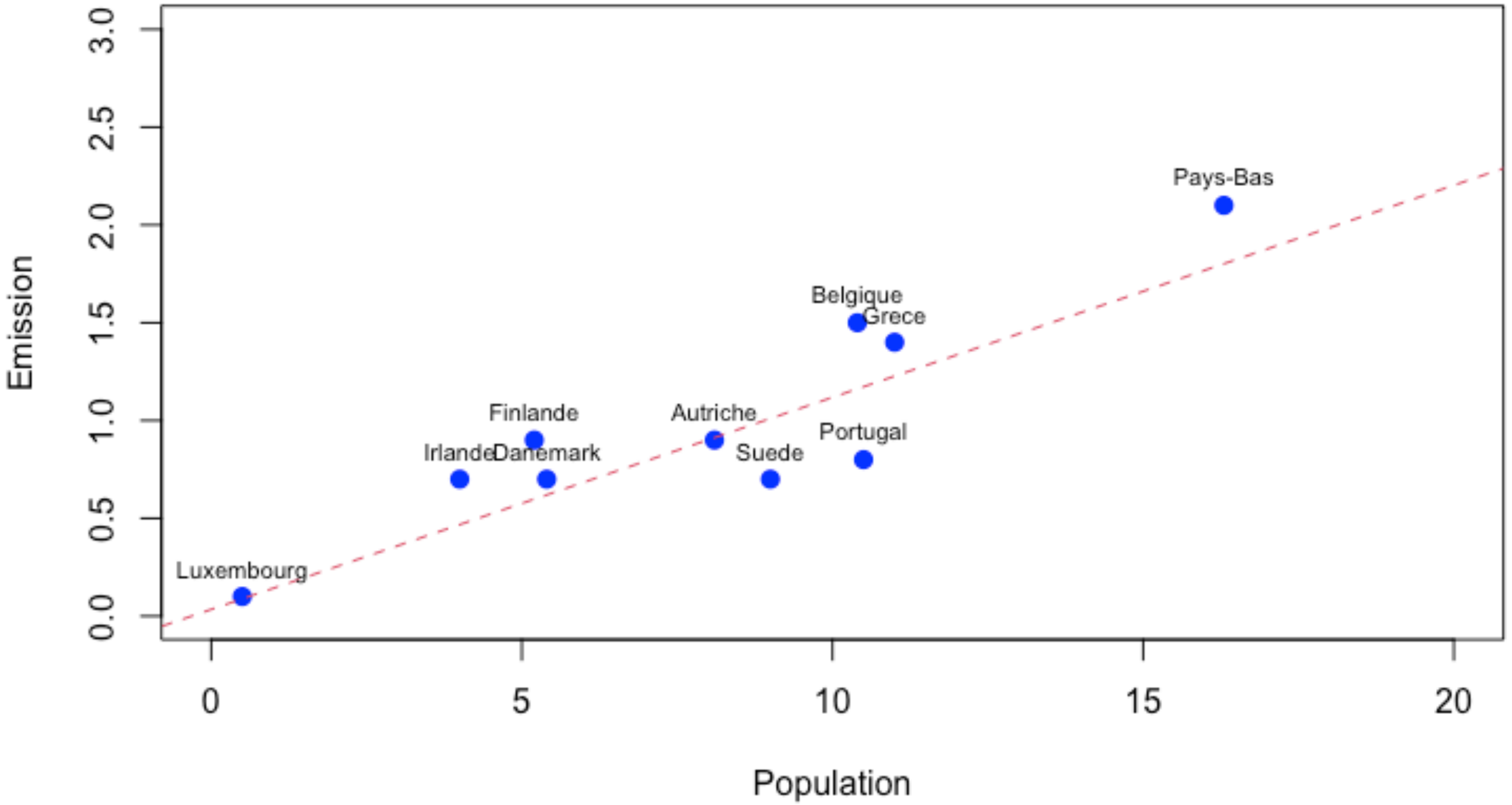
Introduction

- On considère deux variables quantitatives X et Y
- On sait déjà tester si il existe un lien entre ces deux variables.
- On veut aller plus loin et en particulier prévoir Y en fonction de X.

Exemple de problème

	Pays	Population	Emissions.2003
1	Allemagne	82.50	10.20
2	Autriche	8.10	0.90
3	Belgique	10.40	1.50
4	Danemark	5.40	0.70
5	Espagne	41.00	4.00
6	Finlande	5.20	0.90
7	France	59.90	5.60
8	Grece	11.00	1.40
9	Irlande	4.00	0.70
10	Italie	57.50	5.70
11	Luxembourg	0.50	0.10
12	Pays-Bas	16.30	2.10
13	Portugal	10.50	0.80
14	Royaume-Uni	59.50	6.50
15	Suede	9.00	0.70
16	USA	291.00	69.00

Nuage de points



La droite des MCO

- Une droite a pour équation $Y = b_0 + b_1 X$ où b_0 s'appelle l'intercept ou l'ordonnée à l'origine et b_1 le slope ou la pente.
- Les valeurs prédites par le modèle sont : $\hat{Y}_i = b_0 + b_1 X_i$
- Les erreurs de modélisation sont : $\varepsilon_i = Y_i - \hat{Y}_i$
- La droite des moindres carrés est la droite qui minimise $S = \sum_{i=1}^{n} \varepsilon_i^2$

Les estimations des coefficients

Equation de la droite de régression

Pour la méthode des moindres carrés ordinaires, la droite de régression a :

- pour pente $b_1 = \frac{\text{cov}(X, Y)}{\sigma_X^2}$.
- pour ordonnée à l'origine $b_0 = \overline{Y} - b_1\overline{X}$

Tester si il existe un lien linéaire entre X et Y revient à tester $H_0 : r = 0$ par le test de Bravais Pearson. Ceci est équivalent au fait de tester $H_0 : b_1 = 0$ (i.e. pente=0)

Détermination de b_0, b_1

- On ne va pas utiliser la formule mais obtenir ces paramètres soit en utilisant la calculatrice soit en utilisant JAMOVI. (Voir la vidéo calculs_JAMOVI)
- On obtient ici $b_0 = -2.76$,$b_1 = 0.231$

Valeurs prédites et Résidus

	i (Pays)	X_i (Population)	Y_i (Emissions)	$\widehat{Y}_i$	ε_i
1	Allemagne	82.50	10.20	16.27	-6.09
2	Autriche	8.10	0.90	-0.89	1.80
3	Belgique	10.40	1.50	-0.36	1.87
4	Danemark	5.40	0.70	-1.51	2.22
5	Espagne	41.00	4.00	6.70	-2.70
6	Finlande	5.20	0.90	-1.56	2.47
7	France	59.90	5.60	11.06	-5.47
8	Grece	11.00	1.40	-0.22	1.63
9	Irlande	4.00	0.70	-1.84	2.55
10	Italie	57.50	5.70	10.51	-4.81
11	Luxembourg	0.50	0.10	-2.64	2.75
12	Pays-Bas	16.30	2.10	1.00	1.10
13	Portugal	10.50	0.80	-0.34	1.14
14	Royaume-Uni	59.50	6.50	10.97	-4.47
15	Suede	9.00	0.70	-0.68	1.39
16	USA	291.00	69.00	64.37	4.55

Table d'ANOVA du modèle

	SCE	ddl	s^2	F
Régression	$(n-1)s^2_{\widehat{Y}}$	p	$\frac{SCE_{Reg}}{p}$	$\frac{s^2_{Reg}}{s^2_{Res}}$
Résiduel	$(n-1)s^2_{\varepsilon}$	$n-p-1$	$\frac{SCE_{Res}}{n-p-1}$	
Total	$(n-1)s^2_{Y}$	$n-1$	$\frac{SCE_{Total}}{n-1}$	

Pourcentage de variance expliqué

- Il s'agit du coefficient de détermination

$$R^2 = \frac{SCE_{Reg}}{SCE_Y}$$

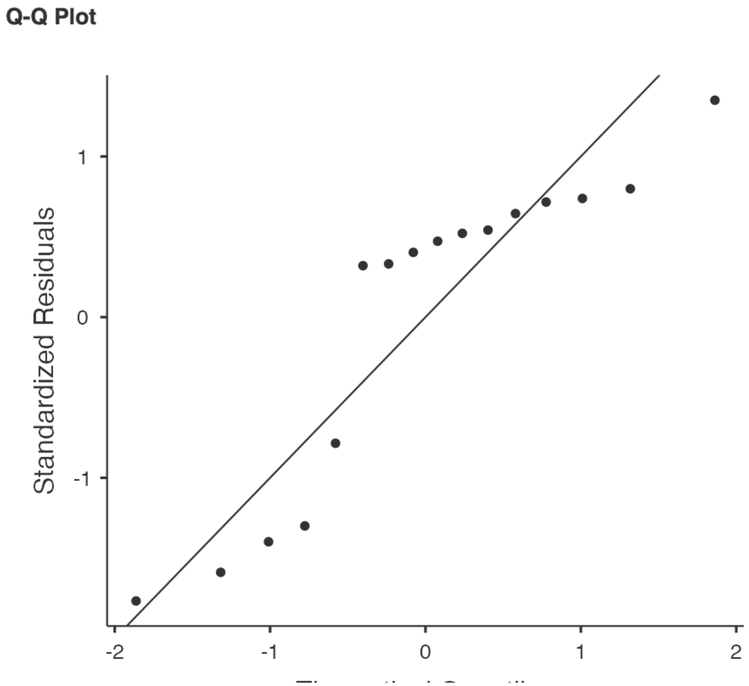
- Lorsque l'on a une seule variable prédictive on a $R^2 = r^2$
- Dans le cas d'une variable on peut également tester $H_0 : b_1 = 0$ de deuxfaçons différentes :
  - Avec le F de Fisher (table d'ANOVA)
  - Avec le t de Student (test sur un coefficient)

Vérifications post-estimation

- Les principales vérifications qui permettre de conforter le choix du modèle et la qualité de l'estimation concerne les résidus $(\varepsilon_i)_{i\in\{1,\dots n\}}$, il faut en particulier vérifier :
  - La normalité des résidus
  - L'homogénéité de la variance de ces résidus.
- L'analyse des résidus va aussi permettre d'identifier des points pour lesquels :
  - la valeur prédite est très éloignée de la valeur observée (i.e. le résidu associé est très grand)
  - Le point est très loin des autres points du nuage et a une forte incidence sur l'estimation du modèle.

Normalité des résidus

Normality Test (Shapiro-Wilk)	
Statistic	**p**
0.832	0.007



Points aberrants et leviers

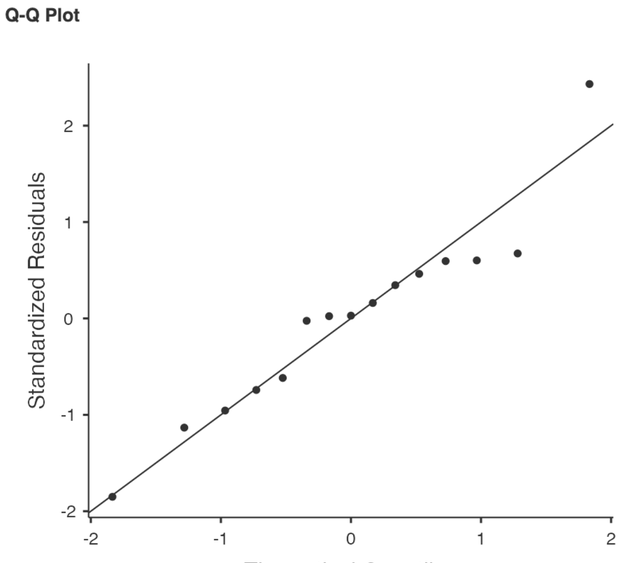
Définition

- Un point i est dit aberrant si le résidu standardisé (centré réduit) lui correspondant $\widetilde{\varepsilon}_i \notin [-3, 3]$
- Un point i est dit levier si la distance de Cook lui correspondant est telle que $d_i > 1$

Case Number	Std. Residual	Emissions.2003	Predicted Value	Residual	Cook's Distance
1	-1.783	10.200	16.270	-6.070	0.146
2	0.525	0.900	-0.891	1.791	0.012
3	0.544	1.500	-0.361	1.861	0.012
4	0.649	0.700	-1.514	2.214	0.018
5	-0.783	4.000	6.697	-2.697	0.020
6	0.721	0.900	-1.560	2.460	0.023
7	-1.588	5.600	11.057	-5.457	0.090
8	0.474	1.400	-0.223	1.623	0.009
9	0.744	0.700	-1.837	2.537	0.025
10	-1.397	5.700	10.503	-4.803	0.069
11	0.807	0.100	-2.644	2.744	0.030
12	0.321	2.100	1.000	1.100	0.004
13	0.333	0.800	-0.338	1.138	0.005
14	-1.299	6.500	10.964	-4.464	0.060
15	0.405	0.700	-0.684	1.384	0.007
16	3.704	69.000	64.362	4.638	48.481

Résultats sans les USA

Model Coefficients - Emissions.2003				
Predictor	**Estimate**	**SE**	**t**	**p**
Intercept	0.0340	0.18824	0.181	0.859
Population	0.1084	0.00519	20.899	< .001



Normality Test (Shapiro-Wilk)	
Statistic	**p**
0.940	0.380

Graphique Final