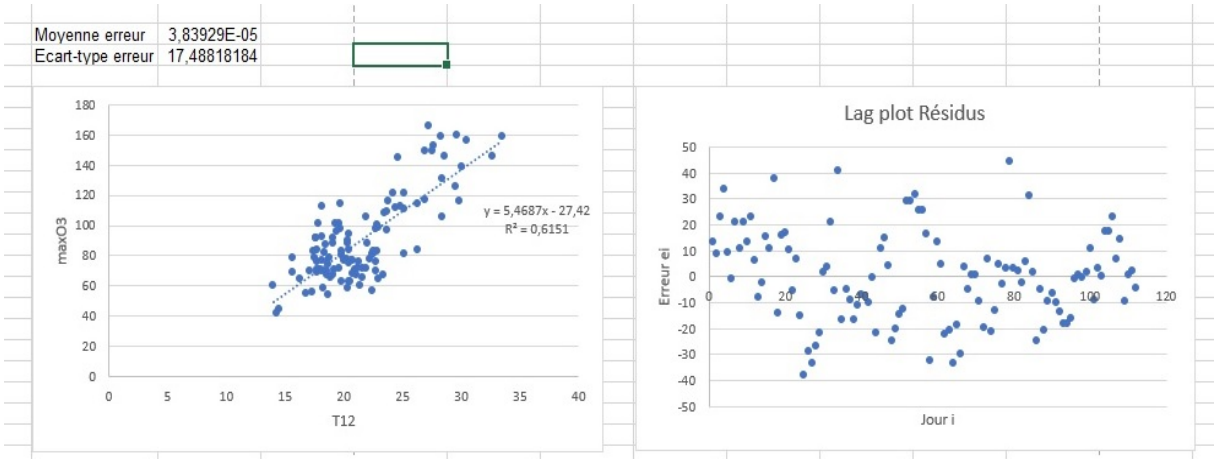
4 Exercice de synthèse

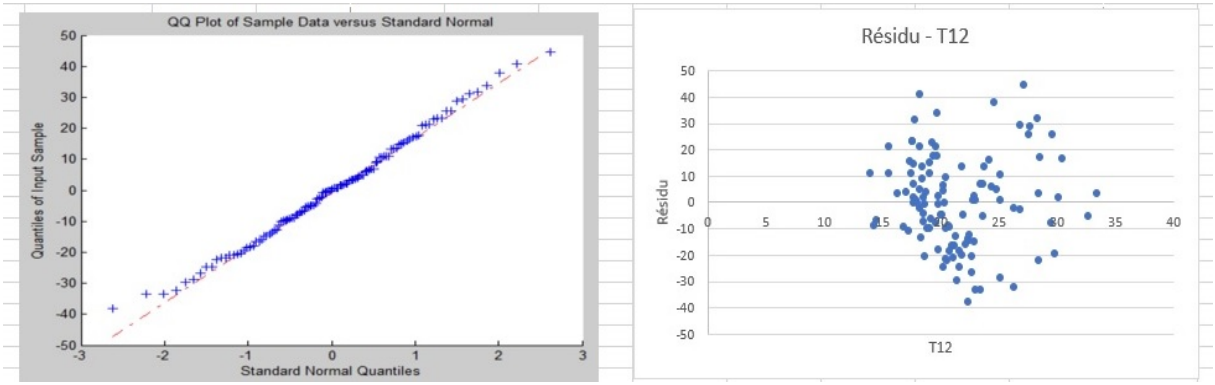
On dispose de données enregistrées dans une ville européenne pour une série de journées d'été (qui sont ici nos individus) :
– l'identifiant de la journée,
– le maximum d'ozone de l'air (variable maxO3),
– les températures à 9h, 12h et 15h (resp. T9, T12 et T15),
– la nébulosité (couverture nuageuse) à 9h, 12h et 15h (resp. Ne9, Ne12 et Ne15),
– la projection du vent sur l'axe est-ouest à 9h, 12h et 15h (resp. Vx9, Vx12 et Vx15),
– le maximum d'ozone de la veille (maxO3v).
Le but est de modéliser la valeur des pics d'ozone en fonction de grandeurs physiques facilement mesurables (température, nébulosité, vent) afin d'avoir des approximations de la qualité de l'air faciles et rapides à obtenir.

1. **Explication du pic d'ozone par la température à 12h**
   (a) Dessiner le nuage de points de maxO3 en fonction de T12. Interpréter.
   (b) Effectuer la régression simple de maxO3 en fonction de T12 en estimant les paramètres du modèle.
   (c) Calculer les erreurs.
   (d) Calculer la moyenne des erreurs, une estimation de son écart-type ainsi que le Q-Q plot associé au vecteur des résidus. Interpréter.
   (e) Tracer l'évolution des résidus en fonction du temps ainsi que les résidus en fonction de T12. Interpréter.
   (f) Calculer le coefficient de détermination. Interpréter.
2. **Explication du pic d'ozone par une régression linéaire multiple**
   (a) Dessiner maxO3 en fonction des différentes variables. Quelles sont celles qui sont a priori intéressantes ?
   (b) Effectuer la régression de maxO3 en fonction de toutes les variables.
   (c) Effectuer "à la main" une procédure "backward" pour sélectionner les variables : on estime le modèle, on retire la variable la moins significative (ayant un coefficient le plus faible) et on recommence (On calcule les différents indicateurs : $\overline{R}^2$, SCR, $\widehat{\sigma}$) pour décider du modèle.

Les résultats de la Régression Linéaire simple de maxO3 en fonction de T12 :

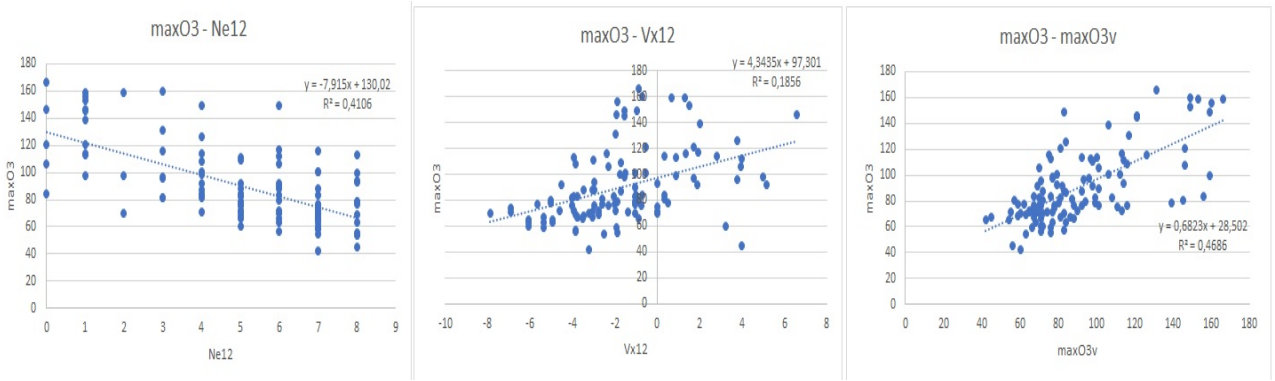


On remarque un nuage de point allongé. Il fait apparaître éventuellement une relation affine (pas assez forte) et positive entre la concentration maximale en Ozone en un jour donné et la température enregistrée à midi ce jour là.
Une estimation des paramètres du modèle est $\widehat{\beta}_0 = -27.42$ et $\widehat{\beta}_0 = 5.4687$ calculée directement avec un outil informatique et pouvant être vérifiée avec les formules du cours.
On calcule ensuite les erreurs $e_i = maxO3_i - (-27.42 + 5.4687 \cdot T12_i)$. On remarque que la moyenne est quasiment nulle $3.83929\ 10^{-5}$ et une estimation de l'écart-type des erreurs est donnée par $\widehat{\sigma} = 17.48$.



En observant le Q-Q plot où on distingue une grande correspondance entre les quartiles de la variable erreur et ceux d'une loi Normale centrée et d'écart-type celui estimé, on peut supposer que la distribution est bien une Gaussienne.
En observant l'évolution des résidus en fonction du temps ainsi que les résidus en fonction de T12, on remarque qu'il n'y a aucune forme apparente distinctive et donc on peut supposer que les erreurs sont indépendantes entre elles mais aussi indépendantes des observations de la température T12.
Toutes les hypothèses étant vérifiées, on peut dire que le modèle s'écrit $maxO3_i = -27.42 + 5.4687 \cdot T12_i + e_i$.

Remarque : Le paramètre $\widehat{\beta}_1 = 5.4687$ représente la pente de la droite de régression mais aussi la sensibilité du modèle et on peut l'interpréter de la manière suivante : si la température de midi augmente en moyenne d'un degrés Celsius, le pic d'Ozone augmente d'un Dobson. *(L'unité Dobson est une unité de mesure de la masse surfacique de l'ozone atmosphérique).*
Le coefficient de détermination étant égale à 0.6151, on peut dire que 61.51% de la variation du pic d'Ozone en une journée est expliquée par la variation de la température enregistrée ce jour là à midi. Ce qui n'est pas suffisant. Car dans ce cas l'erreur prend une part importante dans l'explication de maxcO3. D'où la nécessité d'introduire de nouvelles variables qui peuvent expliquer la variation du pic d'Ozone.



En traçant les scatter plot entre maxO3 et les variables relatives à la nébulosité, notamment Ne12, on remarque que cette variable pourrait éventuellement être intéressante pour l'explication de maxO3 avec

• Contact : imen.kammoun@gmail.com

un coefficient qui va être négatif. Cependant la corrélation entre la vitesse du vent est-ouest du midi et le pic d'Ozone n'est pas très significative. Le pic d'Ozone enregistré la veille (maxO3v) peut aussi avoir une influence sur le maxO3 enregistré pendant une journée donnée vu que le nuage de points est assez allongé. On peut calculer le coefficient de détermination R^2 de chacun des modèles avec chacune des variables prises séparément. On a alors le tableau suivant :

	R2 ajusté
seulement T12	0.6151
seulement T9	0.4891
seulement T15	0.6000
seulement Ne12	0.4106
seulement Ne9	0.3865
seulement Ne15	0.2288
seulement Vx12	0.1856
seulement Vx9	0.2784
seulement Vx15	0.1536
selement maxO3v	0.4686

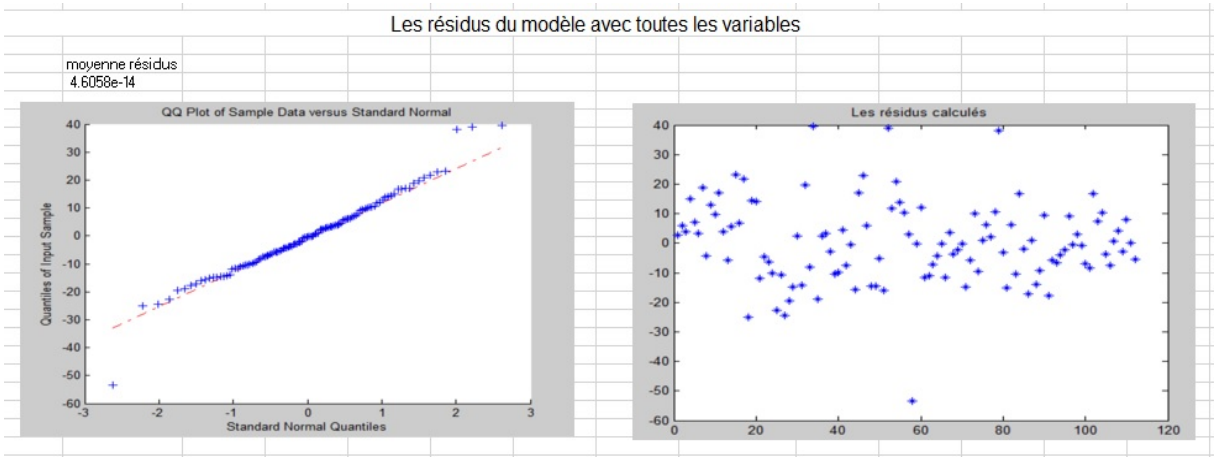
Il en découle que les variables $T12$, $T9$, $T15$, $Ne12$, $Ne9$ et $maxO3v$ peuvent être intéressantes pour l'explication de la variation de maxO3. Évidemment toutes ces variables ne peuvent pas à elles seules bien expliquer la variation du pic d'Ozone, mais ensemble ils pourront. C'est pour cette raison qu'on peut tenter une régression linéaire multiple.

Matlab :

```
[beta,bint,r,rint,stats]=regress(maxO3,[ones(length(maxO3),1) T12 Ne12 Vx12 maxO3v T9 T15 Ne9 Ne 15 Vx9 Vx15])
```

	Toutes les variables	Sans T9	Sans Vx12	Sans Ne15	Sans Ne12	Sans maxO3v
R2 ajusté	0.7638	0.7638	0.7638	0.7638	0.7636	0.6827
Beta						
cste	12.2444	12.2537	12.3091	12.6524	10.8670	
T12	2.2212	2.2094	2.2057	2.3220	2.3676	
Ne12	-0.4210	-0.4278	-0.4328	-0.2998		
Vx12	0.0312	0.0273				
maxO3v	0.3520	0.3517	0.3516	0.3514	0.3518	
T9	-0.0190					
T15	0.5585	0.5563	0.5583	0.4458	0.4475	
Ne9	-2.1891	-2.1854	-2.1860	-2.2028	-2.3547	
Ne15	0.1837	0.1825	0.1827			
Vx9	0.9479	0.9538	0.9627	0.9693	0.9850	
Vx15	0.4186	0.4197	0.4352	0.4198	0.4264	
SCR	2.0827e+04	2.0827e+04	2.0827e+04	2.0834e+04		
ecart type erreur	13.6979	13.6979	13.6980	13.7002		

		modele T12 T9
	R2 ajusté	0.6153
	Beta	
	Cste	-28.4603
	T12	5.2757
	T9	0.2830



En réalisant la régression linéaire avec toutes les variables, on remarque que les hypothèses sur l'erreur sont vérifiées et qu'on obtient les estimation des différents paramètres (ce qui constitue β dans la première colonne du tableau ci-dessus). Avec un coefficient de détermination de 0.7638, on peut dire que les 10 paramètres expliquent à hauteur de 76.38% la variation du pic d'Ozone. Le modèle s'écrit alors

$$maxO3_i = 12.2444 + 2.2212\ T12_i - 0.4210\ Ne12_i + 0.0312\ Vx12 + 0.3520\ maxO3v_i - 0.0190\ T9_i +$$

$$0.5585\ T15_i - 2.1891\ Ne9_i + 0.1837\ Ne15_i + 0.9479\ Vx9_i + 0.4186\ Vx15 + e_i$$

On remarque que le coefficient de T_9 pourrait ne pas être significatif (-0.019) ce qui nous amène à dire que cette variable, évidemment en présence des autres, ne rajoute pas grand chose au modèle. On va l'enlever et refaire la régression avec les 9 autres paramètres.

Remarque : Le fait que le T_9 était intéressante à elle seule et que maintenant elle est exclue n'est pas contradictoire car $T12$ qui a le paramètre le plus élevée (donc la plus influente) elle peut contenir une grande partie de l'information donnée par T_9. (D'ailleur si on réalise la régression où il y a seulement les deux paramètres $T12$ et $T9$, on remarque que cette dernière n'est pas significative par rapport à $T12$. Ce qui explique le fait qu'on retrouve que ce modèle possède un coefficient de détermination proche de $T12$ toute seule.

En enlevant T9, on se retrouve avec un nouveau modèle qui vérifie les hypothèses sur l'erreur et possède le même $\overline{R}^2$. On calcule alors la somme carré des résidus SCR et d'estimateur de l'écart-type des erreur $\widehat{\sigma}$ qui doivent être minimales.
Remarquant que ces indicateurs sont aussi les mêmes, on va enlever la variable $Vx12$ qui semble être non significative avec un coefficient estimé à 0.0273. On remarque de même que les trois indicateurs du modèle sont pratiquement les mêmes. On refait la même chose en enlevant celle qui semble être la moins significative Ne15. Là on remarque que nous avons toujours le même coefficient de détermination mais une SCR plus importante donc on pourrait décider de garder finalement la variable $Ne15$ et que sa contribution à l'explication de maxO3 est significativement non nulle.