TP1 Régression simple

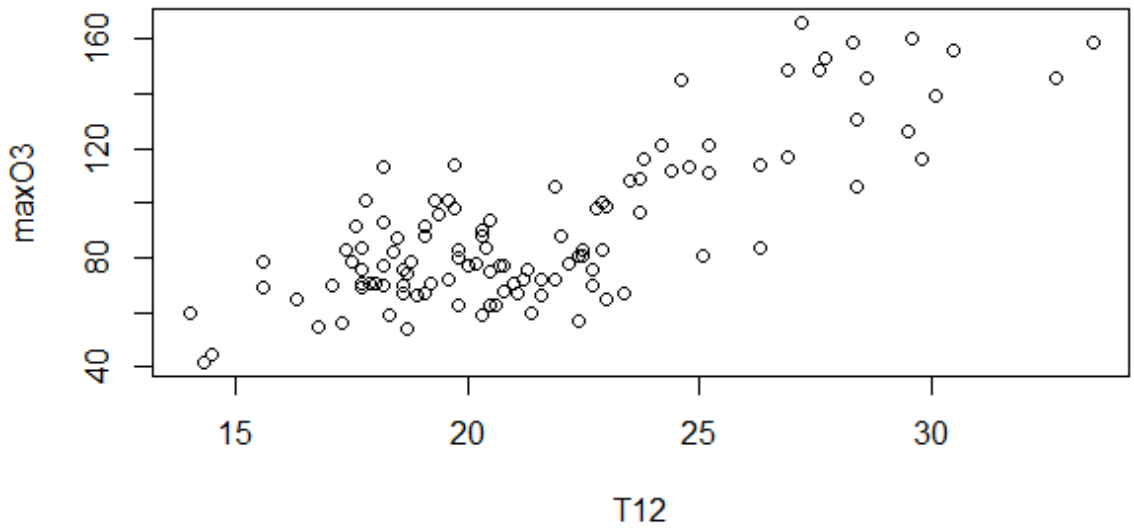
l'objectif de ce TP est d'expliquer la pollution d'ozone `maxO3` (concentration en $\mu g/m^3$) d'un jour donné à l'aide de la variable explicative `T12`, correspondante à la température mesurée à 12h en utilisant R. Il s'agit du modèle linéaire y en fonction de x :

$$\mathbf{maxO3} = \beta_0 + \beta_1 \mathbf{T12} + \varepsilon$$

```
> ozone <- read.table("ozone.txt",header=T,sep="")
> ozone
         maxO3   T9  T12  T15 Ne9 Ne12 Ne15     Vx9    Vx12    Vx15 maxO3v
vent
20010601    87 15.6 18.5 18.4   4    4    8  0.6946 -1.7101 -0.6946     84
Nord
20010602    82 17.0 18.4 17.7   5    5    7 -4.3301 -4.0000 -3.0000     87
Nord
20010603    92 15.3 17.6 19.5   2    5    4  2.9544  1.8794  0.5209     82
Est
20010604   114 16.2 19.7 22.5   1    1    0  0.9848  0.3473 -0.1736     92
Nord
20010605    94 17.4 20.5 20.4   8    8    7 -0.5000 -2.9544 -4.3301    114
Ouest
20010606    80 17.7 19.8 18.3   6    6    7 -5.6382 -5.0000 -6.0000     94
Ouest
20010607    79 16.8 15.6 14.9   7    8    8 -4.3301 -1.8794 -3.7588     80
Ouest
……..
```

Pour décrire le lien entre les deux variables, le nuage de points de `maxO3` con tre `T12`

```
plot(maxO3 ~ T12, data = ozone)
```



Le nuage de points met en avant une structure de dépendance linéaire croissante.

Pour compléter cette visualisation, on calcule le coefficient de corrélation linéaire entre ces deux variables :

```
> cor(ozone$maxO3, ozone$T12)
[1] 0.7842623
reg<-lm(maxO3 ~ T12, data = ozone)
> reg

Call:
lm(formula = maxO3 ~ T12, data = ozone)

Coefficients:
(Intercept)          T12
    -27.420        5.469

> summary(reg)

Call:
lm(formula = maxO3 ~ T12, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-38.079 -12.735   0.257  11.003  44.671

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035    0.003 **
T12           5.4687     0.4125  13.258   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

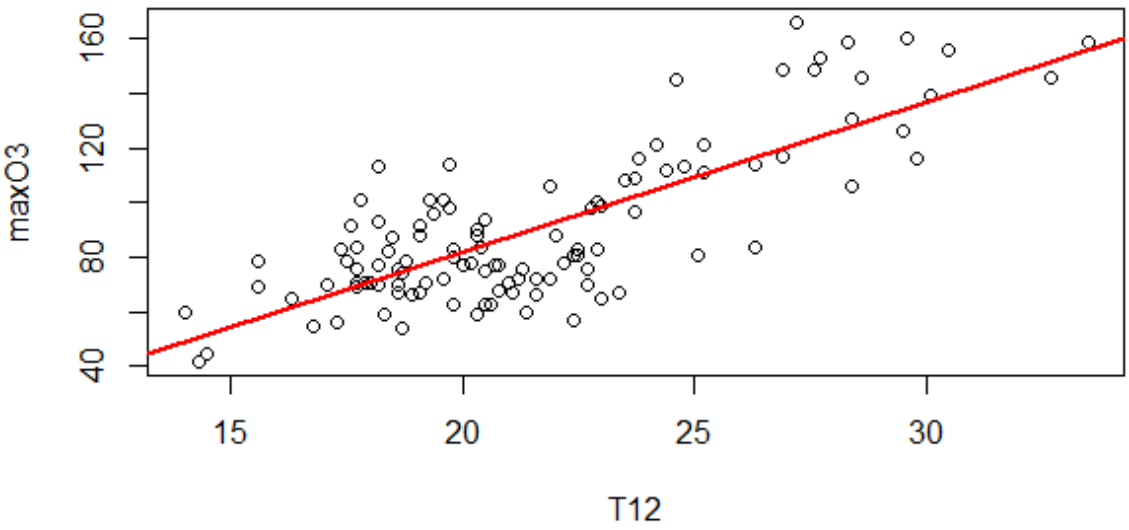
```
Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared:  0.6151,    Adjusted R-squared:  0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

```
> reg$fitted.values
```

```
> reg$residuals
> summary(reg)$sigma
```

```
[1] 17.56749
```

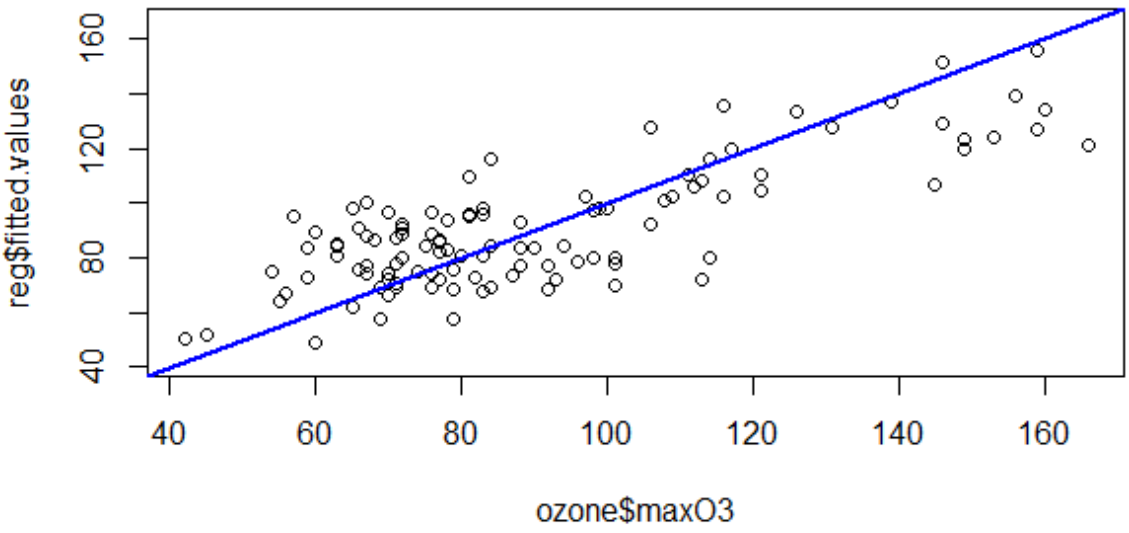
```
abline(reg, col) = "red", lwd=2
```



Graphique des valeurs ajustées en fonction des valeurs observes:

Le graphique des valeurs ajustées y^i en fonction des yi observés, ainsi que l'ajout de la droite d'ordonnée à l'origine 0 et de coefficient directeur 1 (première bissectrice), peuvent s'obtenir par les instructions suivantes. Pour commencer, un objet ajuobs concaténant les valeurs de maxO3 ainsi que les valeurs ajustées est créé, afin de régler correctement l'étendue des axes du graphique :

```
> ajuobs<- c(ozone$maxO3, reg$fitted.values)
> plot(ozone$maxO3, reg$fitted.values, xlim = range(ajuobs), y
lim = range(ajuobs))
> abline(a = 0, b = 1, col = "blue", lwd = 2)
```



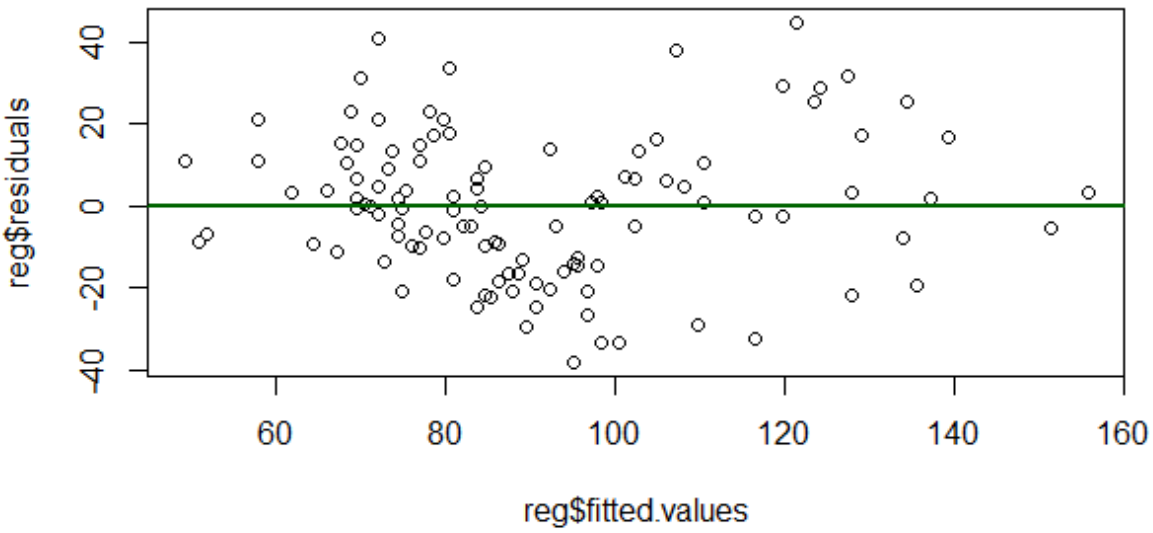
Ce graphique permet d'apprécier visuellement la bonne adéquation des données au modèle : les points tracés doivent être relativement proches de la première bissectrice et répartis de manière équilibrée autour de cette droite.

Graphique des résidus

L'analyse des résidus a pour objectif de tester la validité d'un modèle de régression. Elle permet de déceler les défaillances d'un modèle, c'est pourquoi il est nécessaire de l'effectuer avant toute analyse de régression.

Il est important de vérifier le bon comportement aléatoire des résidus. Une façon de faire est de tracer les résidus en fonction des valeurs ajustées. Les résidus étant centrés, la droite horizontale d'ordonnée 0 est ajoutée, pour pouvoir juger plus facilement de la répartition aléatoire des points. Si le graphique présente une quelconque structure, il convient de changer le modèle.

```
> plot(reg$fitted.values, reg$residuals)
> abline(h = 0, col = "darkgreen", lwd = 2)
```



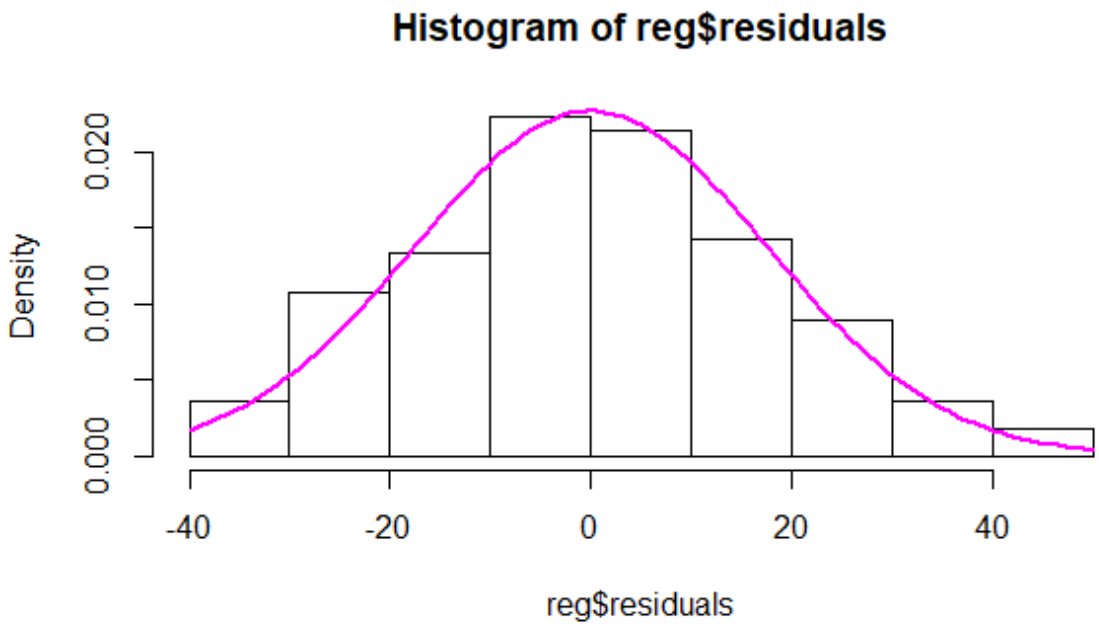
Histogramme des résidus

Les tests de normalité permettent de vérifier si des données réelles suivent une loi normale ou non.

Dans le modèle linéaire, la normalité du terme d'erreur aléatoire ε est supposée. Il faut vérifier cette hypothèse, qui est très importante.

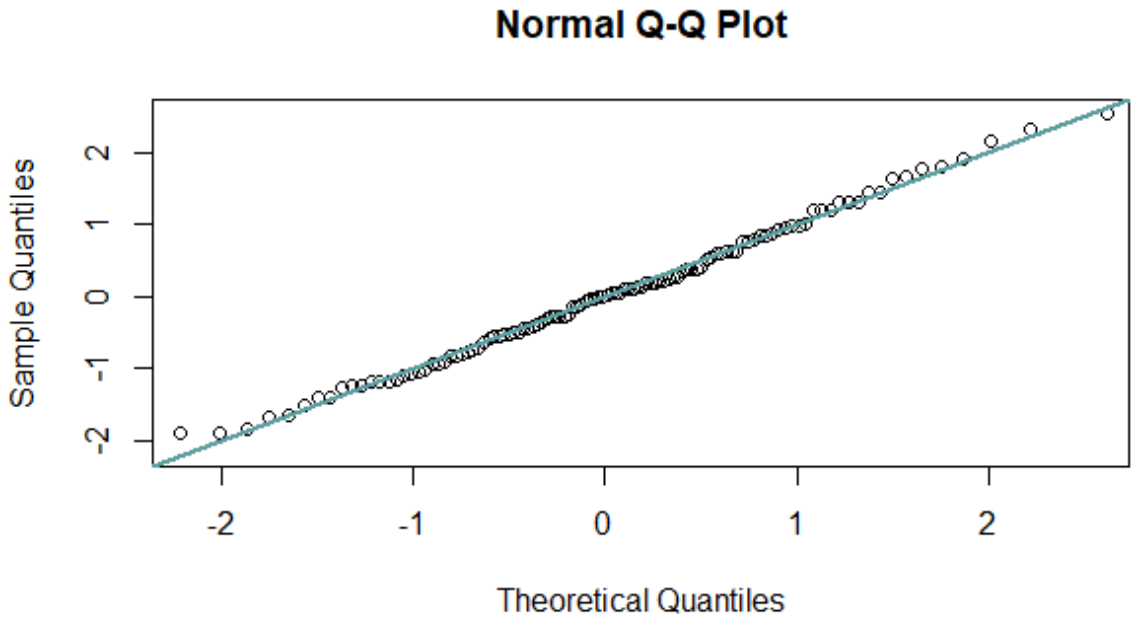
Dans un premier temps, une vérification graphique peut être faite, en traçant l'histogramme des résidus. De plus, la courbe représentative de la densité de la loi normale de paramètres 0 et l'écart-type estimé des résidus peut être superposée.

```
> histo <- hist(reg$residuals, probability = TRUE)
> ec_typ <- summary(reg)$sigma
> curve(dnorm(x, 0, ec_typ), from = min(histo$breaks), to = ma
x(histo$breaks),
+       add = TRUE, type = "l", col = "magenta", lwd = 2)
```



Graphique quantile par quantile

```
> ec_typ <- summary(reg)$sigma
> normed_res <- reg$residuals/ec_typ
> qqnorm(normed_res, xlim = range(normed_res), ylim = range(no
rmed_res))
> abline(0, 1, col = "cadetblue", lwd = 2)
```



Le qqplot des résidus doit être relativement proche de la première bissectrice pour valider graphiquement l'hypothèse de normalité.

Test de normalité des résidus

L'hypothèse de normalité des résidus est testée au niveau 5% :

```
> shapiro.test(reg$residuals)

        Shapiro-Wilk normality test

data:  reg$residuals
W = 0.99235, p-value = 0.792
```

La p-value, de 0.792, est largement supérieure à 5%, donc on ne rejette pas l'hypothèse de normalité des résidus.

Prédiction

La fonction predict est utilisée pour prédire une valeur de y connaissant la valeur de x. Par exemple dans notre cas T12=26

```
> x_new <- data.frame(T12 = 26)
> x_pred <- predict(reg, newdata = x_new)
> x_pred
       1
114.7662
```