Cours 4 : Compléments sur la régression linéaire

1 Retour sur les hypothèses de la régression linéaire : identifiabilité et analyse des résidus

On reprend ici des exercices de la fiche précédente.

1.1 Exercice : Identifiabilité

On considère le jeu de données "ocde.csv" à télécharger, et charger sous le nom `ocde`. On y trouve, en 1960, pour les pays de l'OCDE, le revenu par habitant `PCINC`, la part (en pourcents) de la population active travaillant dans l'agriculture `AGR`, dans l'industrie `IND` et dans les services `SER`.

1. Comparer les coefficients de détermination des ajustements linéaires suivants : `PCINC~AGR+IND+SER` et `PCINC~IND+SER`. Cela suggère-t-il quelque chose sur l'influence de la part d'emplois agricoles sur le revenu moyen ? Regarder les coefficients.
2. Ajouter au data.frame `ocde` une variable égale à la somme des 3 proportions. L'observer, et l'utiliser pour définir une variable `SER2` (toujours dans le data.frame `ocde`) qui modifie légèrement `SER` pour avoir toujours une somme de 100 %. Que donne l'ajustement selon `PCINC~AGR+IND+SER2` ?

Corrigé

```
ocde=read.table("ocde.csv",sep=",",dec=".",header=T,row.names="COUNTRY")
head(ocde)
```

```
##             PCINC AGR IND SER
## CANADA       1536  13  43  45
## SWEEDEN      1644  14  53  33
## SWITZERLAND  1361  11  56  33
## LUXEMBOURG   1242  15  51  34
## U.KINGDOM    1105   4  56  40
## DENMARK      1049  18  45  37
```

```
reg=lm(formula=PCINC~AGR+IND+SER,data=ocde)
summary(reg)
```

```
##
## Call:
## lm(formula = PCINC ~ AGR + IND + SER, data = ocde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -354.56 -201.37  -34.25  120.04  602.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4660.13   16298.65  -0.286    0.779
## AGR            40.93     163.03   0.251    0.805
## IND            59.89     162.89   0.368    0.718
## SER            59.24     162.49   0.365    0.720
```

```
##
## Residual standard error: 288.5 on 16 degrees of freedom
## Multiple R-squared:  0.6346, Adjusted R-squared:  0.5661
## F-statistic: 9.262 on 3 and 16 DF,  p-value: 0.0008716
```

```
reg2=lm(formula=PCINC~IND+SER,data=ocde)
summary(reg2)
```

```
##
## Call:
## lm(formula = PCINC ~ IND + SER, data = ocde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -346.08 -208.91  -12.56  110.76  592.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -568.685    267.431  -2.126   0.0484 *
## IND           19.032      7.013   2.714   0.0147 *
## SER           18.521      9.744   1.901   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280.4 on 17 degrees of freedom
## Multiple R-squared:  0.6332, Adjusted R-squared:   0.59
## F-statistic: 14.67 on 2 and 17 DF,  p-value: 0.0001987
```

On observe que les coefficients de `IND` et `SER` changent nettement quand on ne tient plus compte de `AGR`, ce qui est peu compatible avec le fait que le coefficient de `AGR` n'est pas significativement non-nul. Cela peut conduire à considérer ce coefficient non nul, ou à s'interroger sur la validité des hypothèses sur le modèle... Ici, `AGR`, `SER` et `IND` sont presque liés à la variable constante 1 (cf. la première colonne de la matrice X de la régression). En effet, leur somme vaut presque 100 (%). Par suite, l'hypothèse de non-colinéarité des colonnes de X n'est pas vérifiée (ou presque), ce qui donne une non-unicité des solutions (en cas de colinéarité) et invalide le calcul. Le résultat dépend fortement de l'échantillon observé (la matrice reste inversible, donc donne une estimation, mais les coefficients obtenus dépendent des singularités qui sont liés à des erreurs d'arrondi dans les données), ce qui ne devrait pas être le cas et fausse les interprétations. Le calcul peut aussi être soumis à des erreurs numérique (X^TX est presque singulière), même si R ne signale pas d'erreur.

Les coefficients à retenir ici sont ceux de la régression à deux variables (choisies de façon quelconque).

```
ocde$SER2=100-ocde$IND-ocde$AGR
reg3=lm(formula=PCINC~IND+SER2+AGR,data=ocde)
summary(reg3)
```

```
##
## Call:
## lm(formula = PCINC ~ IND + SER2 + AGR, data = ocde)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.49 -213.18   -8.95  107.34  588.10
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -566.877    268.088  -2.115   0.0495 *
```

```
## IND           19.168      7.011   2.734   0.0141 *
## SER2          18.390      9.797   1.877   0.0778 .
## AGR              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281 on 17 degrees of freedom
## Multiple R-squared:  0.6316, Adjusted R-squared:  0.5882
## F-statistic: 14.57 on 2 and 17 DF,  p-value: 0.0002061
```

Ici, les coefficients de la variable `AGR` sont notés `NA` par R qui a détecté cette fois-ci la colinéarité. La variable n'a pas été prise en compte, comme on le constate en comparant aux résultats précédents.

Il est essentiel, pour interpréter les résultats de la régression, de s'assurer de l'absence de colinéarité entre les variables (y compris avec la variable 1). Cela peut être vérifié a priori, si on anticipe des colinéarités (on enlève alors des variables) ; une alternative serait de considérer des modèles de régression "ridge" qui minimisent non pas $\|Y - X\Theta\|_2$ mais $\|Y - X\Theta\| + \lambda\|\Theta\|$ (pour un certain réel λ à choisir), ce qui stabilise le minimum et permet d'éviter des valeurs trop erratiques.

1.2 Exercice : Analyse de résidus

Pour que l'interprétation des intervalles de confiances ou des tests soit pertinente, il faut vérifier les hypothèses statistiques : les résidus sont indépendants, centrés, gaussiens, de même variance. Regardons un exemple.

On considère le jeu de données `LifeCycleSavings` qui, pour 50 pays, relève le niveau d'épargne `sr` en fonction de la proportion de population âgée de moins de 15 ans `pop15`, de plus de 75 ans `pop75`, du revenu réel disponible par personne `dpi` et du taux de croissance du revenu disponible `ddpi` (tout ceci, en moyenne sur les années 1960–1970).

1. Approcher linéairement le niveau d'épargne à l'aide des autres variables. Garder le résultat dans une variable `reg`.
2. Calculer avec `rstandard(reg)` et représenter graphiquement les résidus standardisés (sous hypothèse gaussienne, ils suivent une loi de student t_{n-p-1} où n est le nombre d'observations et p le nombre de variables de la régression). Donner un `boxplot` de la distribution des résidus, et (séparément) un graphe avec les numéros des observations en abscisse et les résidus en ordonnée. Ajouter sur ce 2è graphe deux lignes horizontales (`gg_hline(intercept= )`) aux quantiles 5% et 95% de la loi normale standard (ou, mieux, de la loi de student à `df`$=n-p-1$ degrés de liberté, calculés avec `qt(p=0.05 ,df= )`). À quels pays correspondent les valeurs en-dehors de ces lignes, et sont-elles vraiment aberrantes ?
3. Pour tester la normalité (ou "studentité") des résidus, on peut représenter leur QQ-plot qui compare quantiles théoriques (loi de student, ou loi normale approximativement) et observés. On l'obtient (pour la loi normale) via `geom_qq()` en ayant spécifié l'esthétique `sample` (égale ici aux résidus standardisés). Ajouter `geom_qq_line()` pour avoir la droite `y=x`. Expliquer, commenter. On peut aussi utiliser, de façon plus quantitative que visuelle, le test de Shapiro-Wilk par exemple : utiliser la fonction `shapiro.test` pour effectuer ce test de normalité.

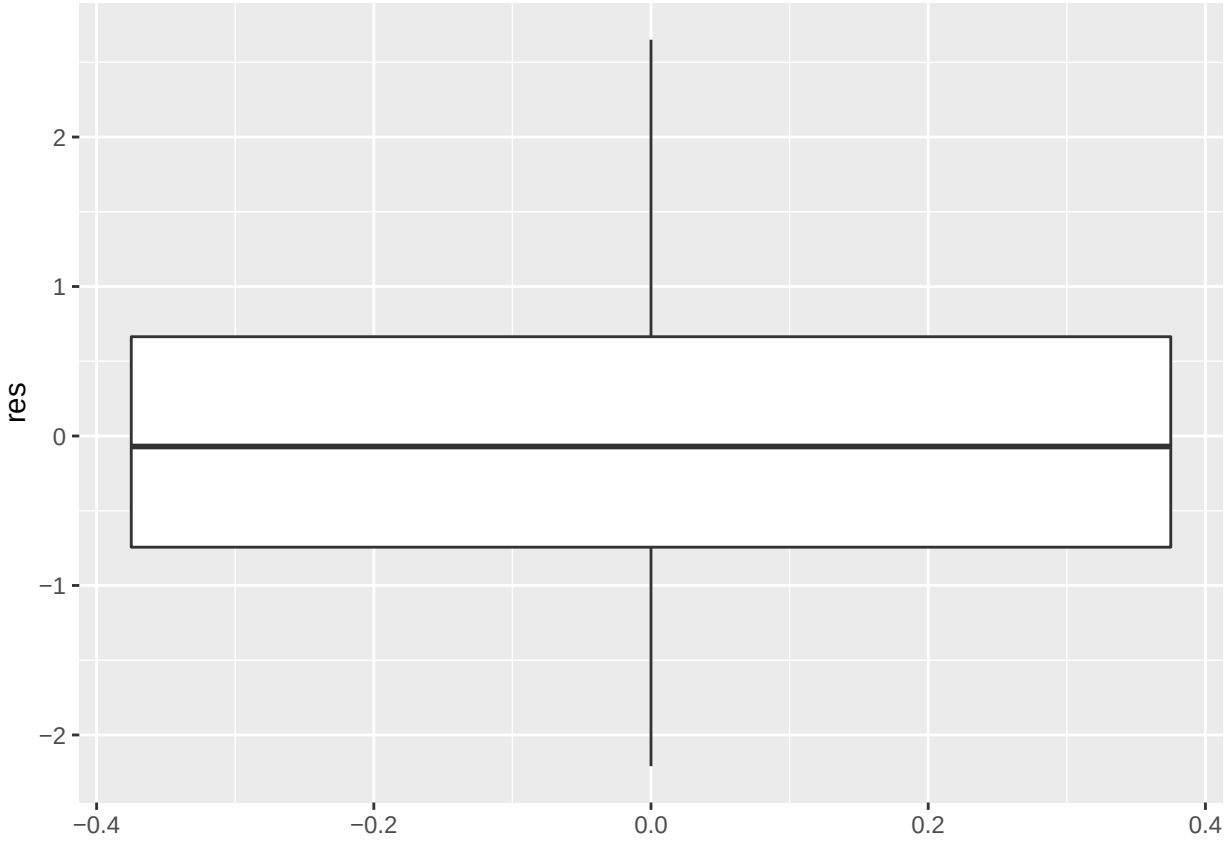
Corrigé

```
?LifeCycleSavings
lcs=LifeCycleSavings
reg=lm(formula = sr~pop15+pop75+dpi+ddpi,data=lcs)
summary(reg)
```

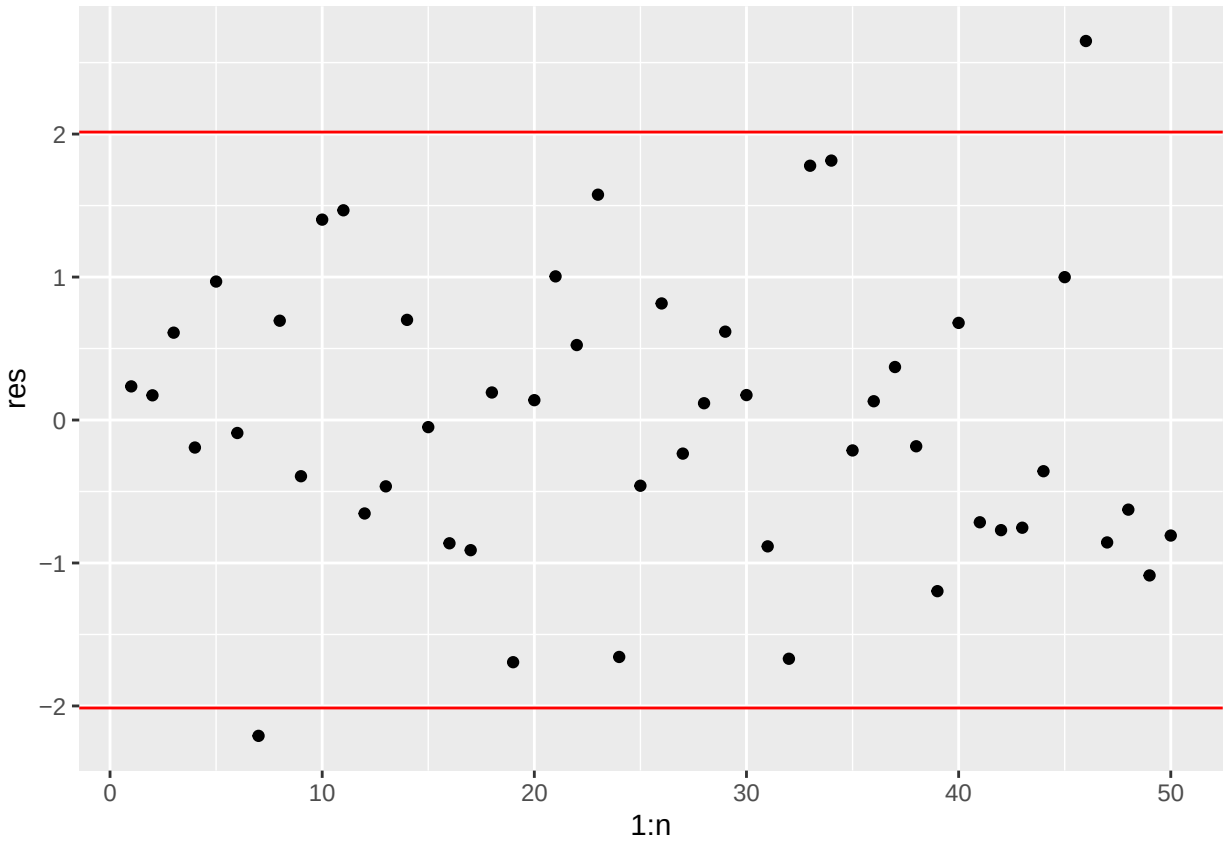
```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = lcs)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```r
res=rstandard(reg)
n=nrow(lcs)
p=4 # nb de régresseurs (variables dans la régression)
# Boxplot de la distribution de "res"
ggplot()+geom_boxplot(aes(y=res)) # Visiblement, pas de valeurs trop aberrante (entre -2 et 2.5)
```




```r
ggplot()+
  geom_point(aes(x=1:n,y=res))+
  geom_hline(yintercept=qt(p=0.025,df=n-p-1),colour="red")+
  geom_hline(yintercept=qt(p=0.975,df=n-p-1),colour="red")
```



On a ici représenté un intervalle de confiance de niveau 95% pour la loi de student à $n-p-1$ degrés de liberté, qui est la loi que devraient suivre les résidus standardisés, sous les hypothèses du modèle gaussien.

Le nombre de valeurs en-dehors de l'intervalle n'est pas aberrant, il est conforme à ce qui est attendu : 5% parmi 50 observation = 2.5 observations en-dehors de l'intervalle en moyenne. De plus l'amplitude de ces valeurs n'est pas signicativement différente de la borne de l'intervalle. Il peut rester intéressant d'identifier les observations concernées, pour le cas où une explication pourrait être apportée. Si des points avec des résidus particulièrement élevés avaient été observés, on aurait pu remettre en question l'homogénéité des variances, et éventuellement (si cela est justifié par l'expérience d'où sont tirées les données) considérer ces observations comme des cas particuliers que l'on pourrait retirer pour le calcul de la régression.

On note aussi une dispersion "raisonnable" des résidus : il n'y a pas de lot d'observations dont les résidus sont proches, par exemple, qui pourrait suggérer une dépendance entre les résidus.

```
# À quels pays correspondent ces 2 valeurs extrêmes ?
which(res>qt(p=0.975,df=n-p-1)) # Zambie
```

```
## Zambia
##     46
```

```
which(res<qt(p=0.025,df=n-p-1)) # Chili
```

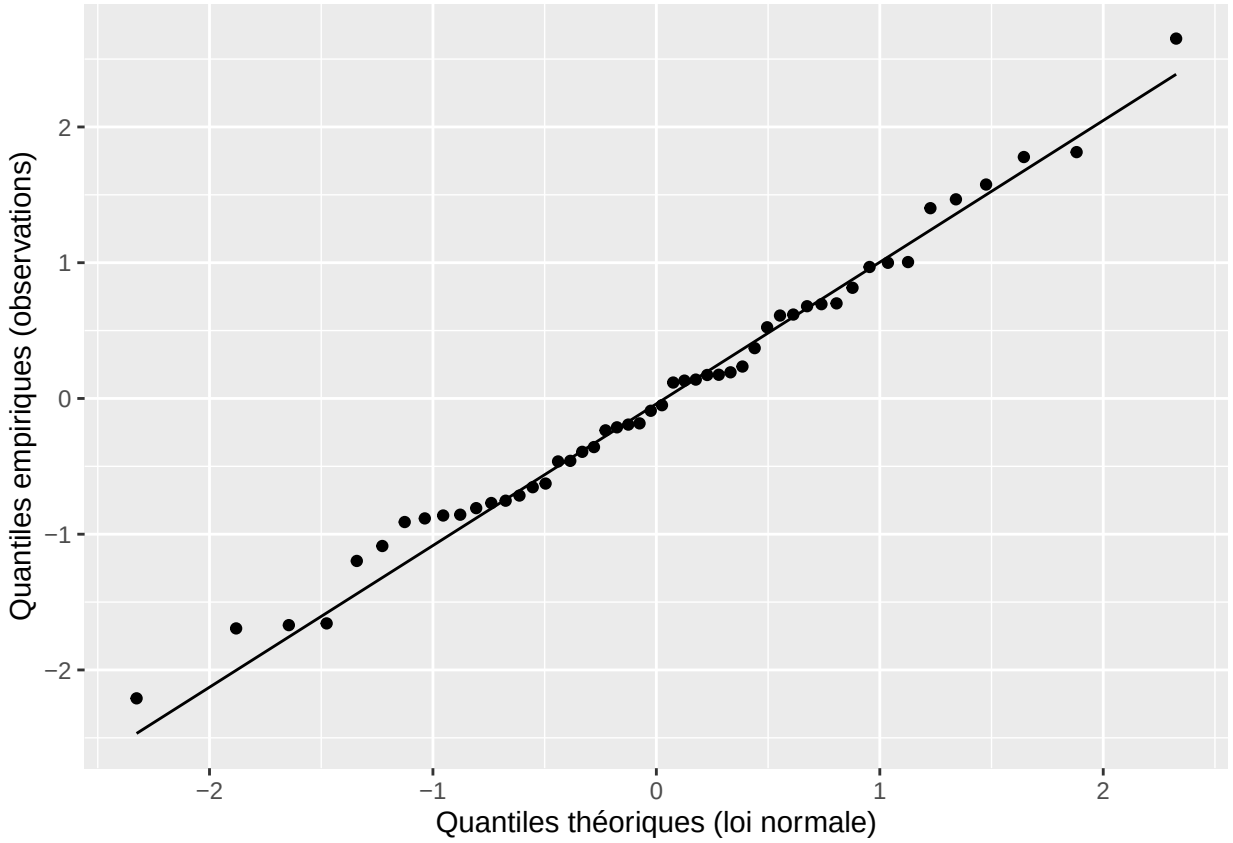
```
## Chile
##     7
```

Produisons avec `ggplot`un QQ-plot vis-à-vis de la loi normale $\mathcal{N}(0,1)$ (qui est une bonne approximation de la loi t_{n-p-1} dès que $n-p-1$ est relativement grand (ici 45)) ; cela a l'avantage d'être la distribution de probabilité par défaut pour les qq-plots.

Pour repréciser le contenu du graphe : les points représentés ont pour coordonnées $(F^{-1}(i/N), x_{(i)})$, où

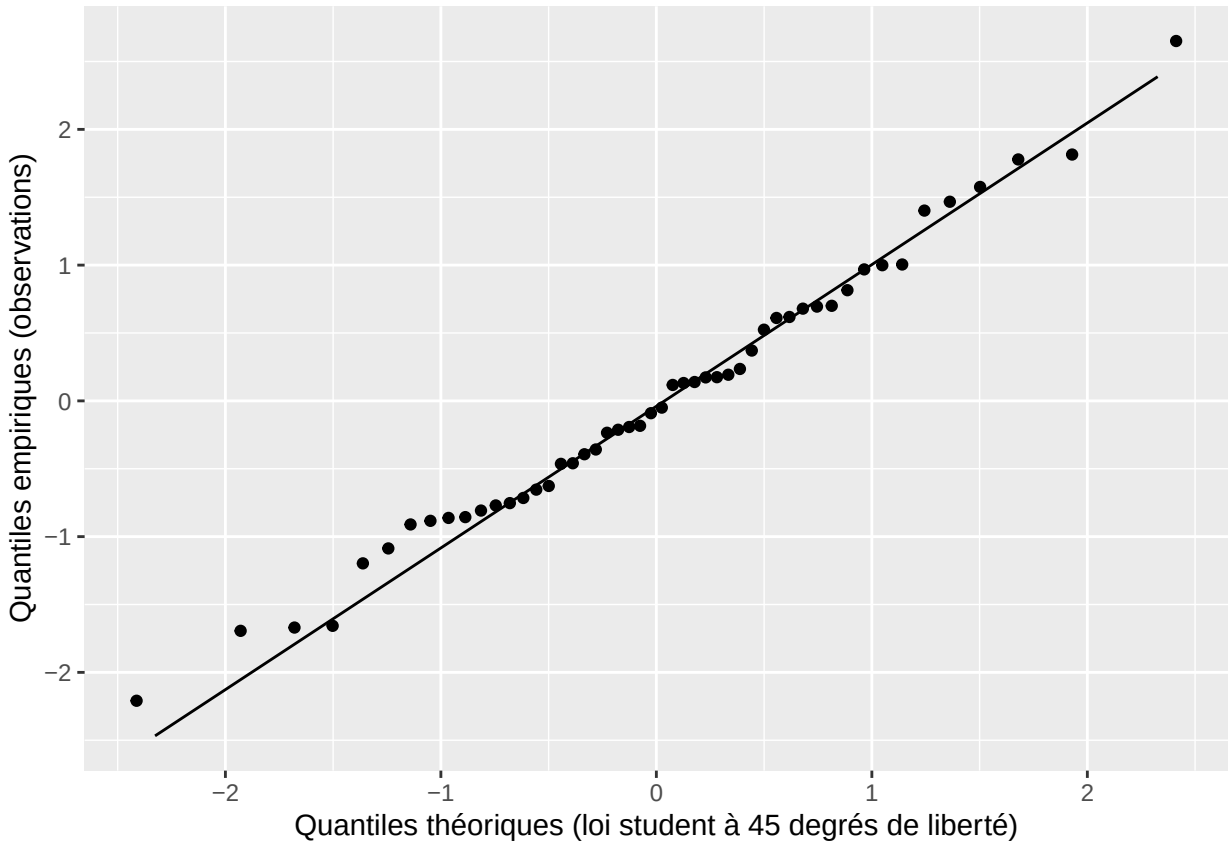
$x_{(1)}, \ldots, x_{(N)}$ sont les observations, classées par ordre croissant, et F est la fonction de répartition de la loi normale standard. En effet, $x_{(i)}$ est le quantile d'ordre $\frac{i}{N}$ des observations (une proportion $\frac{i}{N}$ des points est en-dessous), et $F^{-1}(\frac{i}{N})$ est le quantile d'ordre $\frac{i}{N}$ de la loi normale standard. Autrement dit, les points sont $(F^{-1}(u), \widehat{F}^{-1}(u))$ pour les u où $\widehat{F}$ présente un saut (où $\widehat{F}$ est la fonction de répartition empirique) ; en changeant de variable, c'est aussi le graphe de $(t, F(\widehat{F}^{-1}(t)))$, qui doit être proche de (t, t) pour n grand.

```
ggplot(mapping=aes(sample=res))+
  geom_qq()+
  geom_qq_line()+
  xlab("Quantiles théoriques (loi normale)")+
  ylab("Quantiles empiriques (observations)")
```



Si on avait voulu vraiment comparer les quantiles des observations avec ceux de la loi t_{n-p-1}, on aurait pu faire ainsi :

```
ggplot(mapping=aes(sample=res))+
  geom_qq(distribution = stats::qt, dparams=list(df=n-p-1))+
  geom_qq_line()+
  xlab(paste("Quantiles théoriques (loi student à",as.character(n-p-1),"degrés de liberté)",sep=" "))+
  # paste pour concaténer, as.character pour convertir
  ylab("Quantiles empiriques (observations)")
```



Les points sont "proches" de la droite y=x (qui apparaît sur le graphe) : la distribution est compatible avec une loi gaussienne (ou student).

Pour un test plus quantitatif que graphique, on peut utiliser un test d'adéquation. Par exemple, le test de Shapiro-Wilk d'adéquation à une loi normale (sans préciser les paramètres) :

```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98869, p-value = 0.9109
```

La p-valeur étant supérieure à 0.05, les données sont compatibles avec une loi gaussienne.

On pourrait plus spécifiquement tester avec le test de Kolmogorov-Smirnov l'adéquation à la loi normale standard réduite $\mathcal{N}(0,1)$ ou, puisque c'est possible, à la loi de Student t_{n-p-1}. C'est un test général d'adéquation à une loi continue donnée.

```
ks.test(res,y="pnorm") # y est la fonction de répartition de la loi de référence
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  res
## D = 0.067026, p-value = 0.9671
## alternative hypothesis: two-sided
```

```r
ks.test(res,y="pt",n-p-1) # ici avec la loi de student à n-p-1 degrés de liberté
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  res
## D = 0.06756, p-value = 0.9647
## alternative hypothesis: two-sided
```

Là encore, les p-valeurs supérieures à 0.05 confirment la compatibilité avec une loi de Student à $n-p-1$ degrés de liberté (ou à une loi $\mathcal{N}(0,1)$, qui est très voisine).

Tout ceci valide les tests effectués sur la régression, qui étaient basés sur l'hypothèse du modèle gaussien.

2 Régression linéaire pondérée, et régression linéaire locale (LOESS)

2.1 Régression linéaire pondérée

On peut souhaiter associer un poids $w_i > 0$ à chaque donnée (x_i, y_i), au sens où on cherche la relation affine entre variables $y \simeq b + a_1 x_1 + \cdots + a_p x_p$ qui minimise plutôt la somme des erreurs pondérées :

$$\mathrm{RSS}_w := \|Y - \widehat{Y}\|_w^2 = \sum_{i=1}^n w_i (y_i - \widehat{y}_i)^2 = \sum_{i=1}^n w_i \varepsilon_i^2$$

où $\widehat{y}_i = b + a_1 x_{i,1} + \cdots + a_p x_{i,p}$ et $\varepsilon_i = y_i - \widehat{y}_i$.

On constate que cela revient à remplacer le produit scalaire canonique sur $\mathbb{R}^n$ par le produit scalaire $(x|y)_w = \sum_{i=1}^n w_i x_i y_i = {}^t x W y$ où W est la matrice diagonale ayant pour coefficients diagonaux $w_1, \ldots, w_n$. Ainsi, tout ce qui précède s'adapte à ce produit scalaire, et on obtient en particulier le minimiseur

$$\widehat{\Theta}_w = ({}^t X W X)^{-1}\, {}^t X W Y.$$

et la régression $\widehat{Y}_w = X\widehat{\Theta}_w$.

Une motivation pour ces poids peut être la non-homogénéité des variances de ε_i. S'il y a une raison *a priori* de supposer que la variance de l'erreur ε_i est λ_i fois plus grande que la variance de ε_1, choisir $w_i = \frac{1}{\sqrt{\lambda_i}}$ ramène au cas d'une variance homogène des résidus $\varepsilon_i^{(w)} = w_i \varepsilon_i$ et permet donc de faire les tests et intervalles de confiances déjà vus.

2.2 Régression LOESS

La régression non-paramétrique LOESS (LOcally Estimated Scatterplot Smoothing) est une régression locale : en chaque point $z = (z_1, \ldots, z_p) \in \mathbb{R}^p$, la régression est donnée par une régression linéaire $\widehat{z} = z\Theta_{w(z)}$ calculée avec des poids $w(z)_i$ dépendant de la proximité des points $x_i = (x_{i,1}, \ldots, x_{i,p})$ à z. Par exemple on peut considérer

$$w(z)_i = e^{-\|x_i - z\|^2} \qquad \text{ou} \qquad w(z)_i = \left(1 - \left(\frac{\|x_i - z\|}{D}\right)^3\right)^3$$

avec $D = \max_i \|x_i\|$. On obtient ainsi le graphe de régression $z \mapsto z\Theta_{w(z)} = z({}^t X W(z) X)^{-1}\, {}^t X W(z) Y$, qui hérite de la régularité de w.

La régression LOESS est une façon d'obtenir une "tendance locale" très régulière d'un nuage de points qu'on ne pourrait pas résumer globalement à une relation affine. Le choix de w est bien sûr critique, de la même façon que la taille des corbeilles d'un histogramme : si w tend rapidement vers 0, la régression est très locale

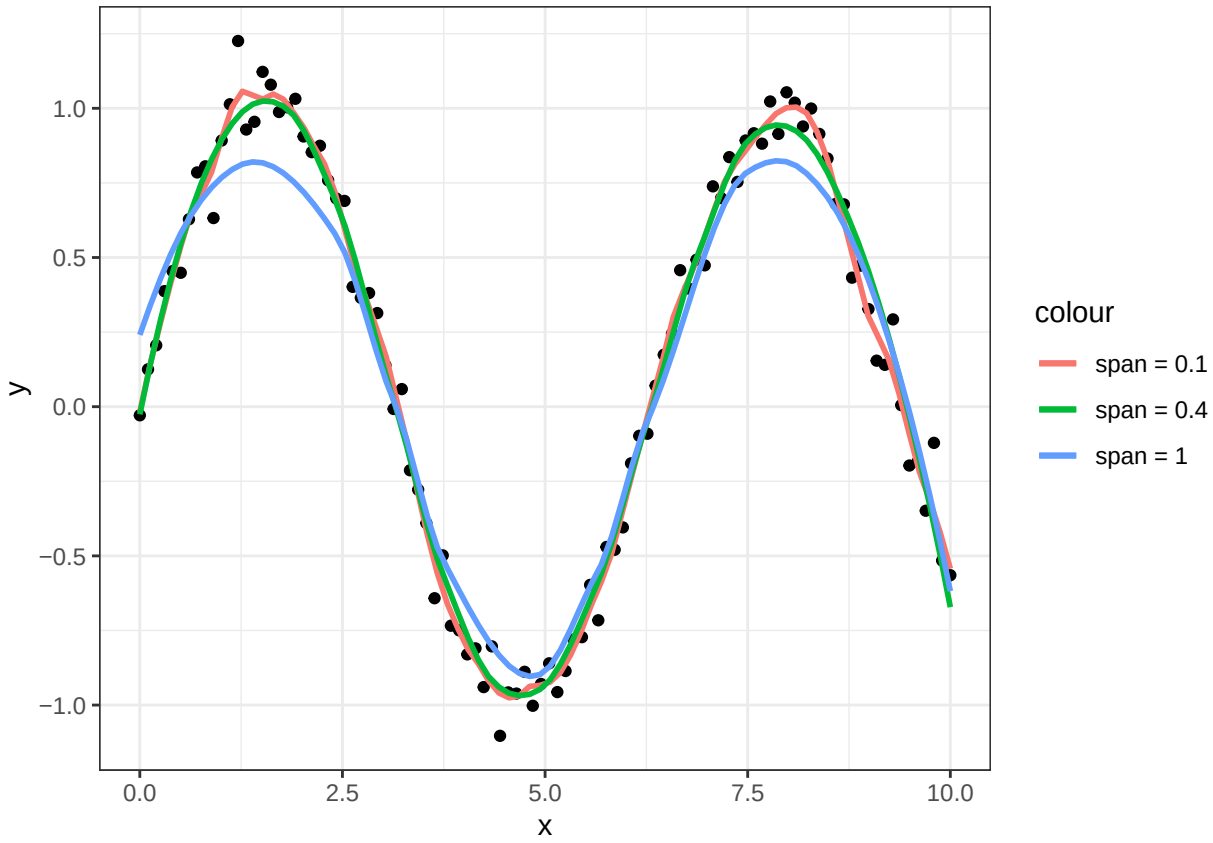
et ne décrit que les points individuellement, inversement si w tend lentement vers 0, la régression donne des poids voisins à tous les points et devient une régression linéaire globale.

Le choix classique pour la régression LOESS, qui dépend d'un paramètre `span`$\in [0,1]$, est la deuxième fonction $w(z)_i$ donnée plus haut si x_i fait partie des `span`$\cdot n$ points les plus proches de z, et $w(z)_i = 0$ pour les points plus éloignés, qui ne sont donc pas du tout pris en compte. (Il y a aussi d'autres paramètres, cf. `?geom_smooth` ou `?loess`). On pourrait écrire un programme pour chercher numériquement la valeur de `span` qui minimise par exemple la norme L^2 des erreurs aux points $x_1, \ldots, x_n$.

Regardons un exemple sur des données simulées (100 abscisses équiréparties entre 0 et 10, et ordonnées données par le sinus de l'abscisse perturbé par une variable gaussienne $\mathcal{N}(0, 1/2)$), pour trois valeurs de `span` :

```
x=seq(from=0,to=10,length.out=100)
y=sin(x)+rnorm(n=100,0,.1)
ggplot(mapping=aes(x=x,y=y))+geom_point()+
  geom_smooth(span=.1,mapping=aes(colour="span = 0.1"),se=F)+
  geom_smooth(span=.4,mapping=aes(colour="span = 0.4"),se=F)+
  geom_smooth(span=.6,mapping=aes(colour="span = 1"),se=F)+
  theme_bw() # fond blanc
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



–>

3 Régression contre une variable catégorielle : ANOVA

Supposons que l'on dispose d'une variable continue y et d'une variable catégorielle x, dont on numérote les modalités $1, 2, \ldots, k$. On observe donc $(x_1, y_1), \ldots, (x_n, y_n)$ où $y_i \in \mathbb{R}$ et $x_i \in \{1, 2, \ldots, k\}$. Si l'on souhaite tester le fait que la variable y ne dépend pas de la valeur de x, on effectue une analyse de variance (ou "ANOVA", pour ANalysis Of Variance), dans le cadre d'un modèle linéaire gaussien.

Le modèle linéaire gaussien consiste à supposer que les observations $y_1, \ldots, y_n$ sont des réalisations de variables aléatoires indépendantes $Y_1, \ldots, Y_n$ ayant des lois gaussiennes de même variance σ^2 et de moyenne dépendant de la modalité de $x_1, \ldots, x_n$: si $x_i = j$ alors $Y_i \sim \mathcal{N}(a_j, \sigma^2)$, où $a_1, \ldots, a_j$ sont les moyennes "intra-groupes", c'est-à-dire pour chaque modalité. Le fait que y ne dépend pas de x revient à l'hypothèse $a_1 = \cdots = a_k$.

On peut constater que ce modèle peut s'écrire sous forme de régression linéaire : on a

$$Y_i = b + a_1 \mathbf{1}_{\{x_i=1\}} + \cdots + a_p \mathbf{1}_{\{x_i=k\}} + \varepsilon_i$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Les variables de la régression sont alors y et les indicatrices $\mathbf{1}_{\{x_i=1\}}, \ldots, \mathbf{1}_{\{x_i=k\}}$. Cependant, les colonnes de la matrice X associée sont linéairement liées car $\mathbf{1}_{\{x_i=1\}} + \cdots + \mathbf{1}_{\{x_i=k\}} = 1$. On fait le choix arbitraire d'enlever $\mathbf{1}_{\{x_i=1\}}$:

$$Y_i = b + a_2 \mathbf{1}_{\{x_i=2\}} + \cdots + a_p \mathbf{1}_{\{x_i=k\}} + \varepsilon_i = \begin{cases} b & \text{si } x_i = 1 \\ b + a_2 & \text{si } x_i = 2 \\ \cdots & \\ b + a_k & \text{si } x_i = k. \end{cases} + \varepsilon_i.$$

On souhaite dans ce cas tester si $a_2 = \cdots = a_k = 0$. Puisque l'on s'est ramené à une régression linéaire, il s'agit du test de Fisher déjà vu. Si on souhaite ensuite tester si $a_k = 0$, c'est le test de Student déjà vu.

Il s'avère que R voit automatiquement une régression linéaire comme un problème d'analyse de variance, dès lors qu'une variable est catégorielle. On effectue donc l'ANOVA de la même façon qu'une régression linéaire, sans avoir à introduire explicitement les indicatrices ci-dessus.

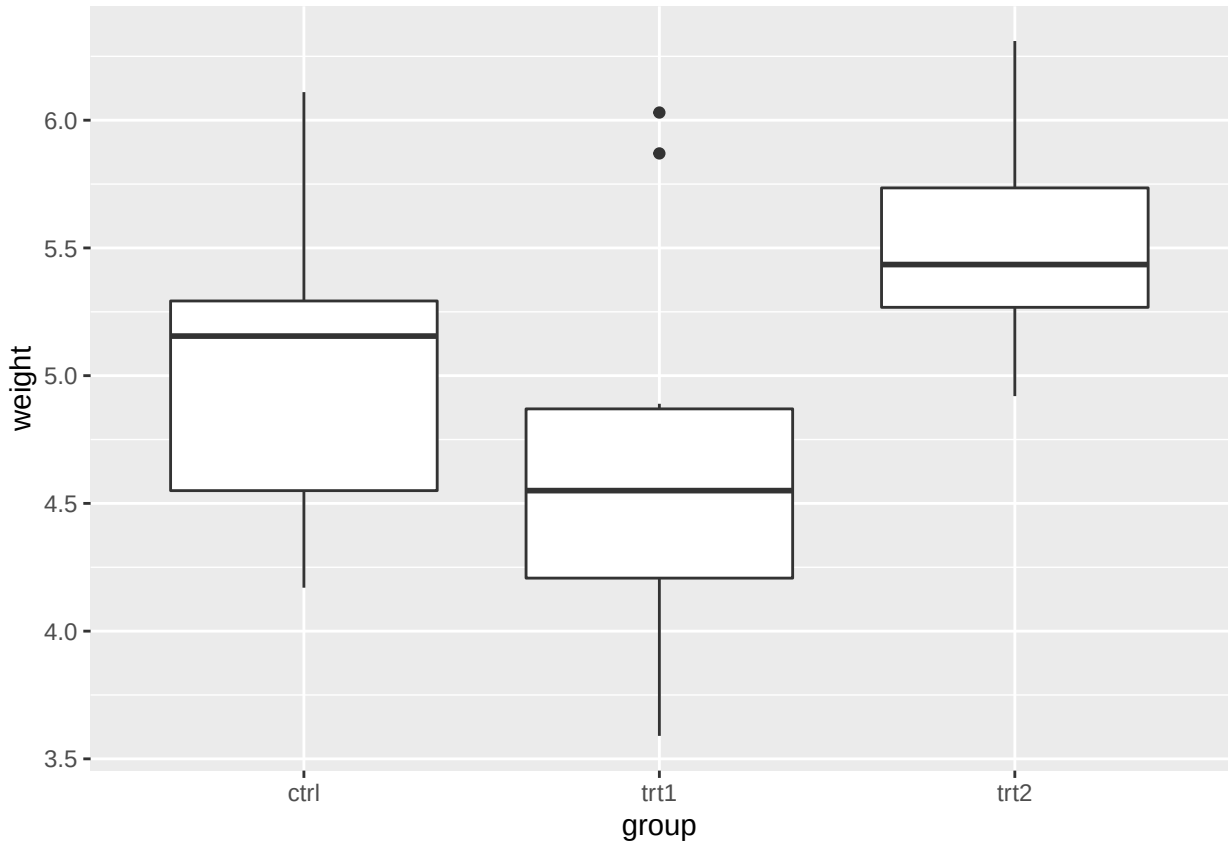
Exemple : on considère l'incidence du choix d'un traitement sur le poids de plantes.

```
?PlantGrowth
str(PlantGrowth)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  $ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# la variable "group" vaut
# - "ctrl", si aucun traitement n'a été appliqué
# - "trt1", si le traitement 1 a été appliqué
# - "trt2", si le traitement 2 a été appliqué

ggplot(data=PlantGrowth,aes(x=group,y=weight))+
  geom_boxplot()
```




```
# Visuellement, les moyennes sont très différentes selon les groupes, mais
# la variance rend la conclusion moins claire puisqu'une part importante des
# plages de valeurs du poids des groupes se chevauchent.

reg=lm(data=PlantGrowth,formula = weight~group)
summary(reg)
```

```
##
## Call:
## lm(formula = weight ~ group, data = PlantGrowth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0710 -0.4180 -0.0060  0.2627  1.3690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0320     0.1971  25.527   <2e-16 ***
## grouptrt1    -0.3710     0.2788  -1.331   0.1944
## grouptrt2     0.4940     0.2788   1.772   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6234 on 27 degrees of freedom
## Multiple R-squared:  0.2641, Adjusted R-squared:  0.2096
## F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

```
# À la fin : p-valeur de 0.01591 pour le test de nullité des coefficients
# (hors intercept), ce qui revient à rejeter l'hypothèse au niveau de confiance
# de 5%, ce qui revient à rejeter l'hypothèse d'homogénéité : les traitements
# ont bien un effet sur le poids des plantes.
# Par contre individuellement les p-valeurs 0.1944 et 0.0877 ne permettent pas
# de conclure au même niveau de confiance de 5% que l'un ou l'autre est
# significativement non nul. Il vaudrait mieux disposer d'un plus large
# échantillon pour pouvoir conclure sur l'utilité du traitement 1 pour diminuer
# le poids, ou sur l'utilité du traitement 2 pour l'augmenter.
```

Ensuite, on peut considérer plus généralement des modèles de régression avec plusieurs variables catégorielles (ANOVA à plusieurs facteurs), ou des modèles de régressions avec variables catégorielles et continues. Via la commande `lm`, R reformule le problème en termes d'indicatrices pour les modalités des variables catégorielles et le ramène au cas de régressions linéaires. (Où, à nouveau, une analyse des résidus doit être menée pour s'assurer des hypothèses du modèle linéaire.)

4 Régression logistique

Dans le cas de la section précédente, certaines des variables explicatives $x_1, \dots, x_p$ sont catégorielles, mais la réponse y reste continue.

Si la réponse y est catégorielle et $x_1, \dots, x_p$ sont continues, on ne peut pas s'attendre à une relation affine : on sort du cadre du modèle linéaire gaussien, où les variables ont des valeurs réelles quelconques.

On a vu que, si Y suit une loi gaussienne de moyenne donnée par une combinaison affine de $X_1, \dots, X_p$, alors l'estimateur du maximum de vraisemblance est donné par la régression linéaire. Pour Y discrète, on peut introduire un modèle de nature similaire où une loi discrète (loi de Bernoulli si y prend ses valeurs dans $\{0, 1\}$) remplace la loi gaussienne, avec toutefois un artifice pour ramener le paramètre dans $[0, 1]$:

$$Y \sim \mathcal{B}(g(b + a_1 X_1 + \cdots + a_p X_p)),$$

où $g : \mathbb{R} \to [0, 1]$ est une fonction continue strictement croissante telle que $\lim_{-\infty} g = 0$ et $\lim_{+\infty} g = 1$, et on peut recherche l'estimateur du maximum de vraisemblance pour une telle loi. C'est la **régression logistique**.

En pratique on utilise en effet souvent la fonction logistique $g : x \mapsto \frac{e^x}{1+e^x}$.

À la différence de la régression linéaire, le maximum de vraisemblance n'est pas explicite ici, et est obtenu de façon approchée par des méthodes numériques.

En R, on utilise la fonction `glm(y~x1+x2+x3,family=binomial(link="logit"),data=df)` (où `logit` désigne la fonction logistique, c'est en fait la valeur par défaut).

5 Exercice : Hauteur des eucalyptus

Lorsque l'on cherche à estimer la quantité de bois produite par une forêt, il est nécessaire de connaître la hauteur des arbres afin de calculer le volume par une formule (le volume d'un tronc de cône, approximativement). Cependant, mesurer la hauteur d'un arbre d'une vingtaine de mètres n'est pas simple : on mesure en général une méthode trigonométrique à partir de la mesure d'un angle de visée entre le sol et le sommet de l'arbre, ce qui suppose une vision claire de la cime et un recul suffisant pour avoir une mesure précise. Quand ce n'est pas possible, on peut effectuer une régression linéaire à partir de la mesure de la circonférence à 1,30 mètres du sol. Cela suppose de disposer d'un échantillon d'apprentissage : un ensemble d'arbres pour lesquels ont réellement été mesurées la circonférence et la hauteur. Le fichier "eucalyptus.txt" regroupe de telles données pour environ 1400 eucalyptus.

1. Charger les données dans un data frame. Les représenter graphiquement dans le plan (avec `ggplot`), ainsi que la droite de régression linéaire.

2. Afficher la valeur des coefficients de la régression, et du coefficient de détermination (R^2). Afficher les intervalles de confiance sur les coefficients (appliquer la fonction `confint( )` au résultat de la régression). L'approximation des données par une droite affine paraît-elle légitime ? Cela était-il attendu au vu du nuage de points ?
3. Au lieu du nuage de points, utiliser une représentation par histogramme 2D via `geom_bin_2d()` : en quoi cette représentation donne-t-elle une image plus fidèle du "nuage de points" ?
4. On dispose d'arbres de circonférence 50, 100, 200 et 500. Proposer des prédictions de leur hauteur. Pour cela, on utilisera la fonction `predict(reg,df2)` où `reg` est le résultat de la régression, et `df2` est un data frame contenant les mêmes variables explicatives que pour le calcul de `reg` (ici, juste la circonférence `circ`), avec les nouvelles données.
5. Pour améliorer la régression pour de petites valeurs, on propose le modèle $c = b + a_1 h + a_2\sqrt{h} + \varepsilon$ où c est la circonférence et h la hauteur. Effectuer le calcul de la régression. Afficher une représentation graphique (nb : dans `geom_smooth`, la formule relie `y` et `x` plutôt que les noms de variables). Que dire de l'amélioration ? Peut-on affirmer que le nouveau coefficient a_2 est significativement non nul ?
6. On s'intéresse aux résidus pour cette seconde régression.

a. Calculer les résidus standardisés. Combien sortent de l'intervalle de confiance de niveau 95% pour la loi gaussienne (ou de Student) ? Est-ce anormal ?
b. Représenter le graphe qqplot comparant la loi empirique des résidus normalisés avec la loi gaussienne standard. L'hypothèse du modèle gaussien paraît-elle justifiée ? Effectuer de plus un test de Shapiro-Wilk. Que donne le test de Kolmogorov-Smirnov ? (cf. corrigé de l'exercice 1.2) Comprenez-vous le problème, voyez-vous une fonction de le contourner ?