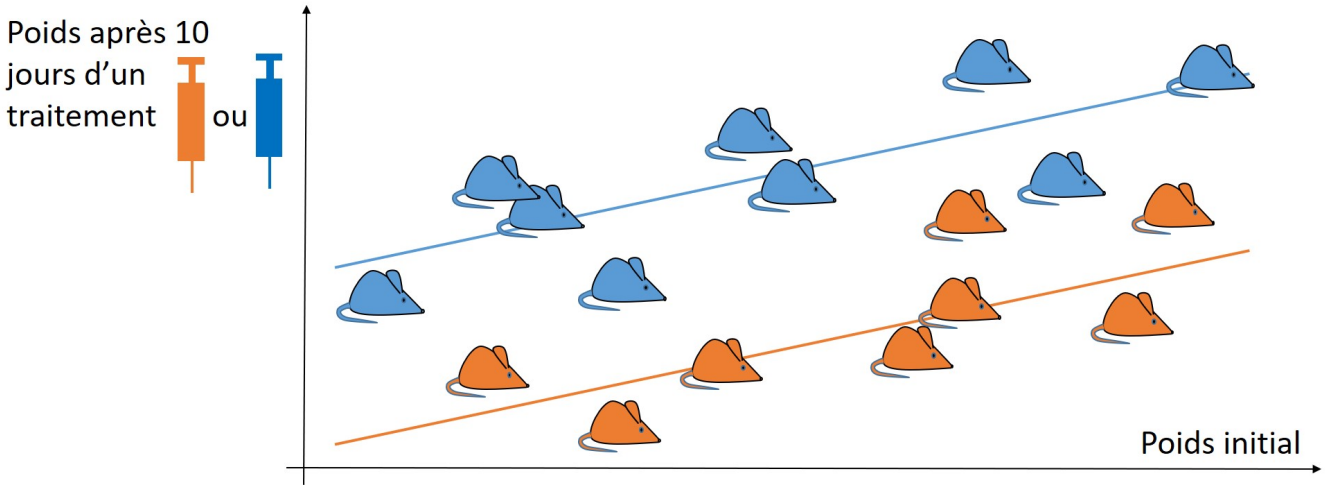
Le modèle linéaire : modélisation d'une variable quantitative en fonction de variables quantitatives ou qualitatives

Marie Laure Delignette-Muller - VetAgro Sup



Introduction

Les modèles linéaires nous fournissent les briques de base pour construire de très nombreux modèles utilisés en statistique.

Mes objectifs :

- vous faire bien **comprendre les principaux concepts** associés aux modèles linéaires,
- vous initier à la **manipulation des modèles linéaires sous R**,
- développer votre **lecture critique des travaux publiés** utilisant les modèles linéaires.

Notre exemple fil rouge

Nous présenterons différents modèles sur un unique jeu de données extrait d'une ancienne publication décrivant le **temps de survie** (en jours) des **tiques adultes** en fonction de la **température** et de l'**humidité relative** :

Milne, A. (1950). The ecology of the sheep tick, Ixodes ricinus L.: microhabitat economy of the adult tick. Parasitology, 40(1-2), 14-34.

Importation du jeu de données

```r
dtot <- read.table("DATA/Milne1950.txt", header = TRUE)
str(dtot)
```

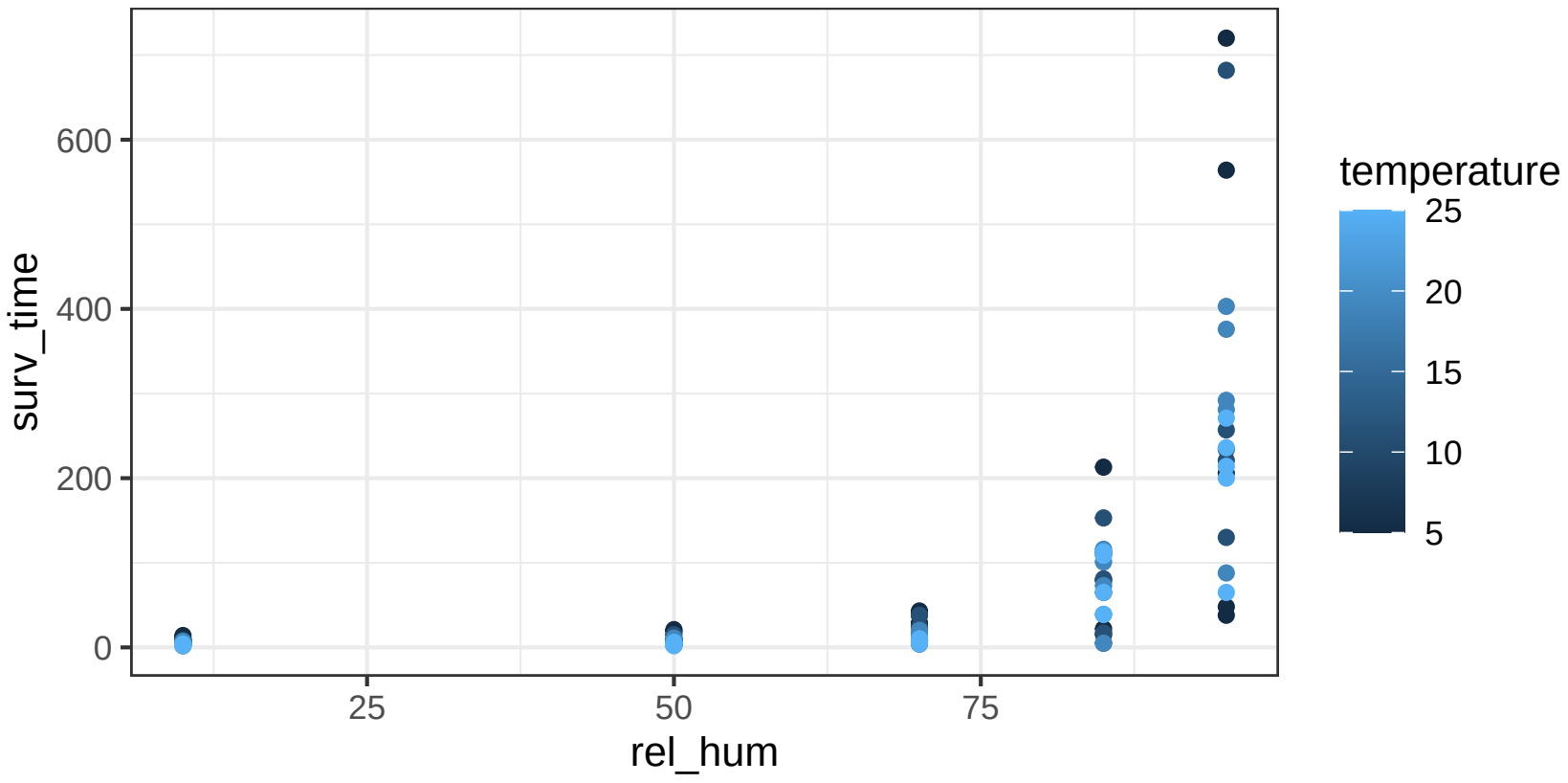
```
## 'data.frame':    100 obs. of  3 variables:
##  $ rel_hum    : int  0 50 70 85 95 0 50 70 85 95 ...
##  $ surv_time  : int  7 7 22 15 38 9 9 23 22 48 ...
##  $ temperature: int  5 5 5 5 5 5 5 5 5 5 ...
```

```r
# replacement of 0% humidity by 10%
# as in the paper Wongnak et al. 2022
dtot$rel_hum[dtot$rel_hum == 0] <- 10

# add of the log10 tranformed survival time
dtot$log10_surv_time <- log10(dtot$surv_time)
```

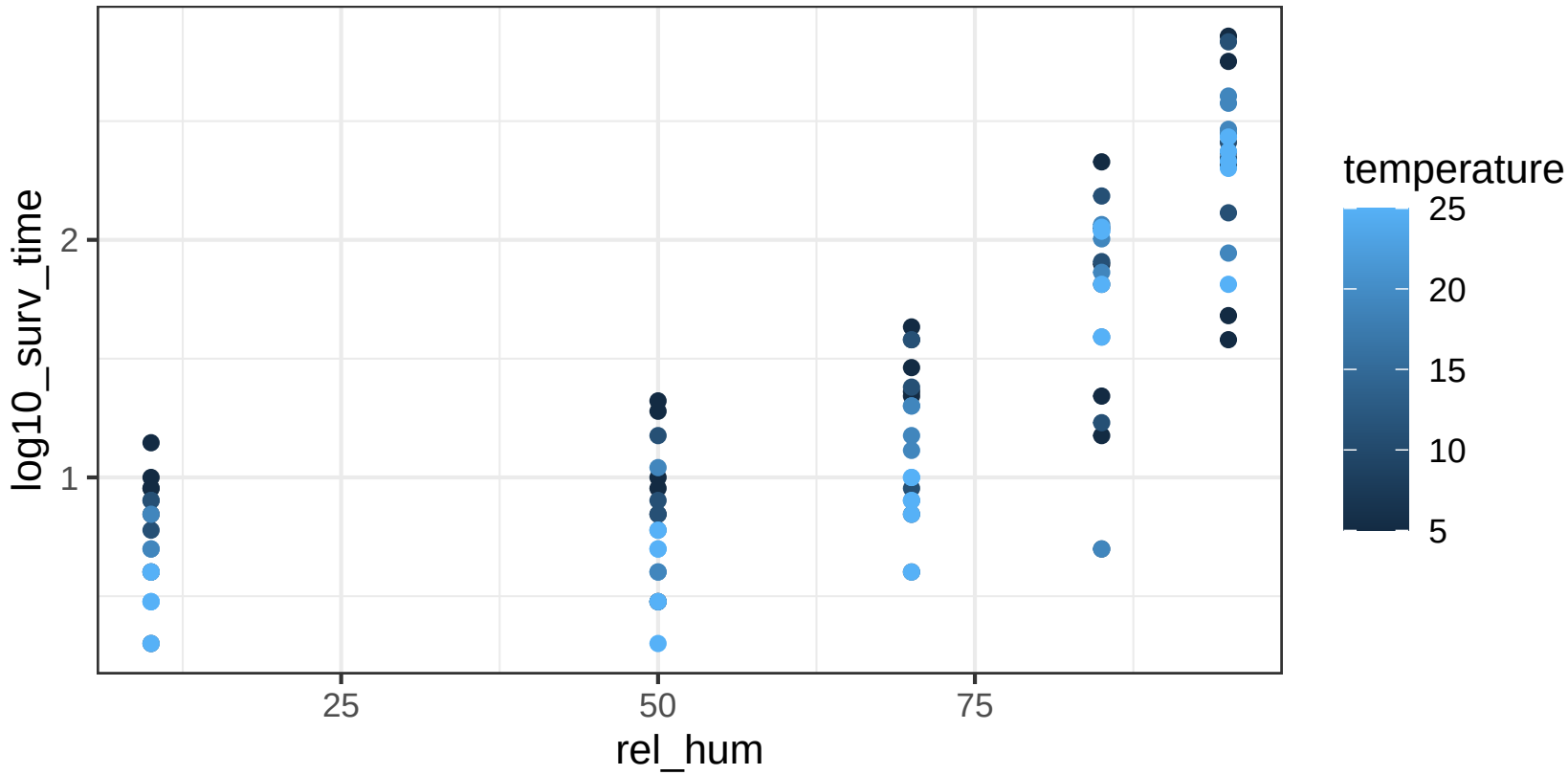
Représentation graphique des données

```
ggplot(data = dtot, aes(x = rel_hum, y = surv_time,
       col = temperature)) + geom_point()
```



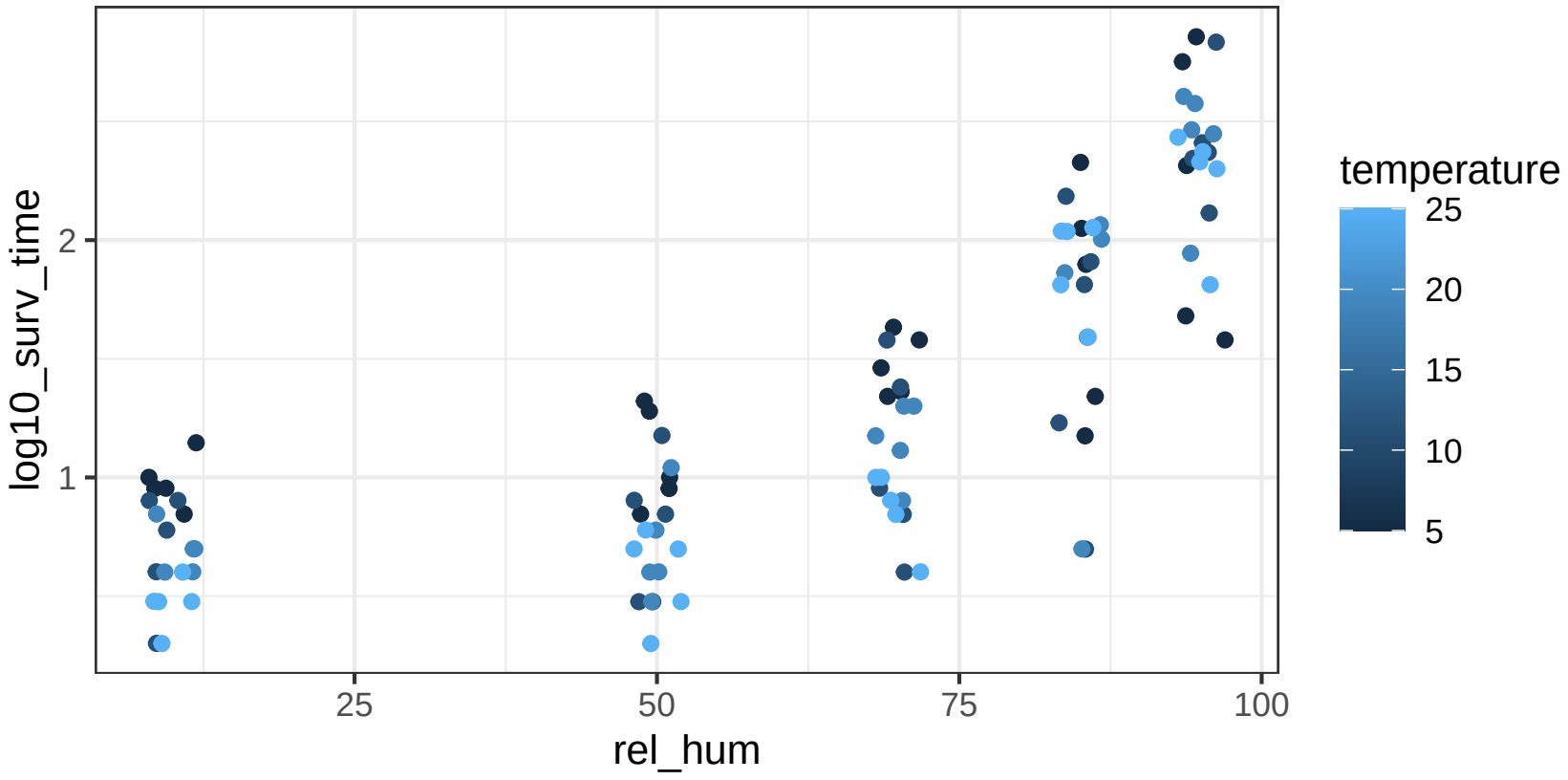
Représentation des données après transformation logarithmique de la variable à expliquer

```
ggplot(data = dtot, aes(x = rel_hum,
  y = log10_surv_time, col = temperature)) +
  geom_point()
```



Ajout de bruit sur l'axe des x pour voir tous les points

```
ggplot(data = dtot, aes(x = rel_hum,
  y = log10_surv_time, col = temperature)) +
  geom_jitter(width = 2)
```



Définition des termes de base

La **variable à expliquer** = la **variable dépendante** = le temps de survie (en jours) = une **variable continue**

Pas de censure ici car l'expérience a été poursuivie pour atteindre la mort pour chaque tique.

Les variables explicatives = les **variables indépendantes** :

- l'humidité relative (en %)
- la température (en degrés Celsius)

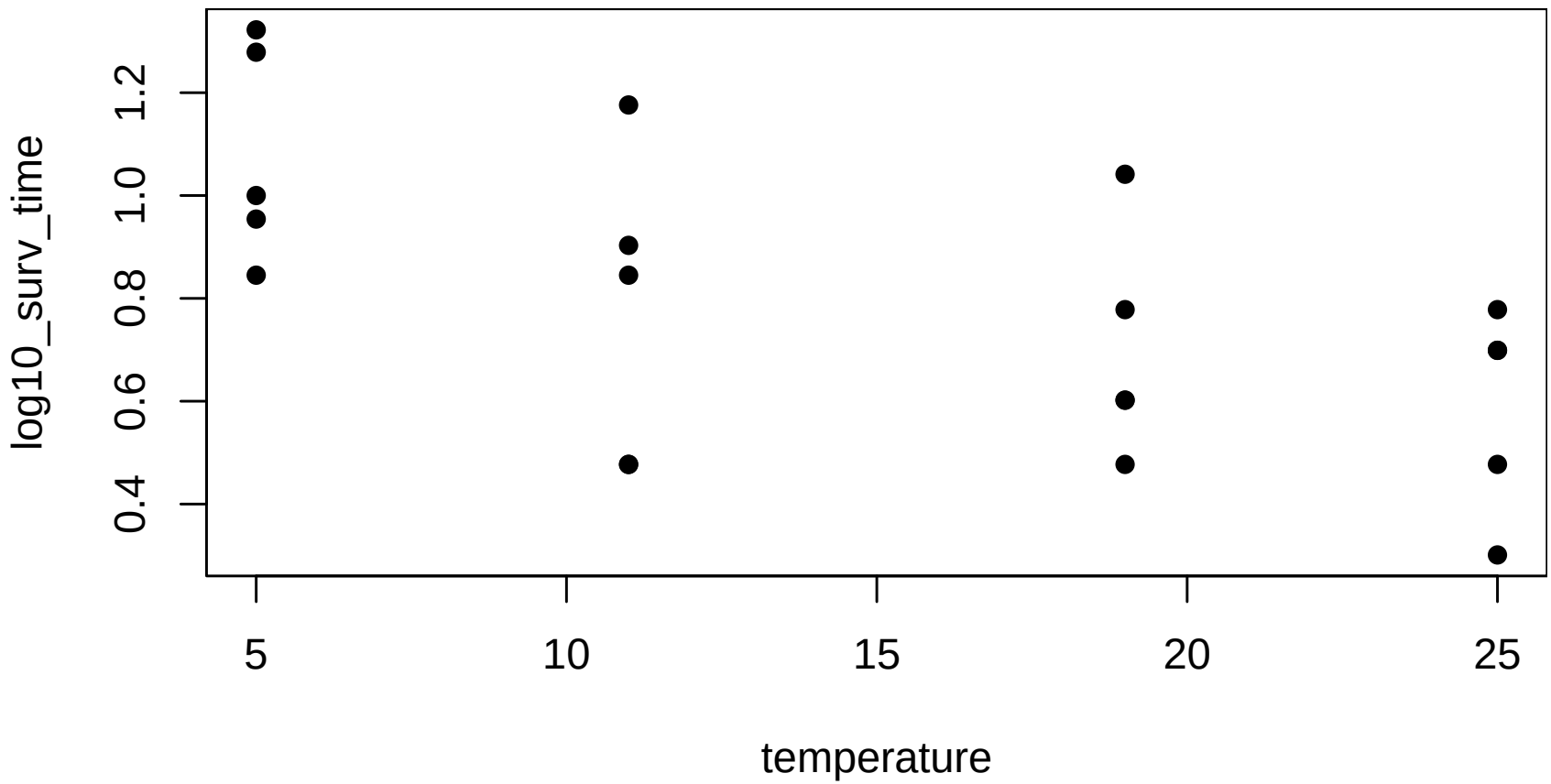
Dans cet exemple, les deux variables explicatives sont contrôlées (**étude expérimentale**).

Rappel sur la régression linéaire simple

Impact de la température pour des conditions d'humidité faible

en utilisant un **sous-ensemble** des données, restreint **à un niveau d'humidité de 50%**

```
dRH50 <- subset(dtot, rel_hum == 50); par(mar = c(4,4,0,0))
plot(log10_surv_time ~ temperature, data = dRH50, pch = 16)
```



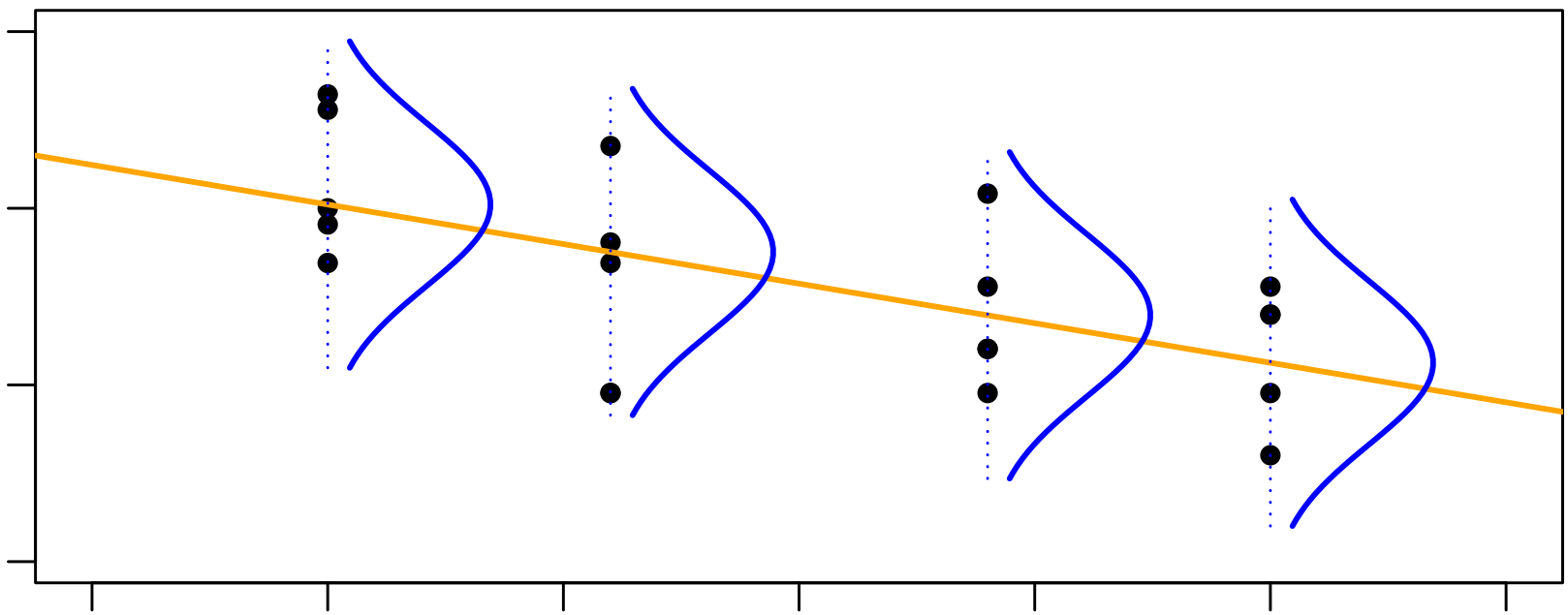
Modèle théorique

$Y_i = \alpha + \beta X_i + \epsilon_i$ avec $\epsilon_i \sim N(0, \sigma)$

Partie déterministe : lien linéaire

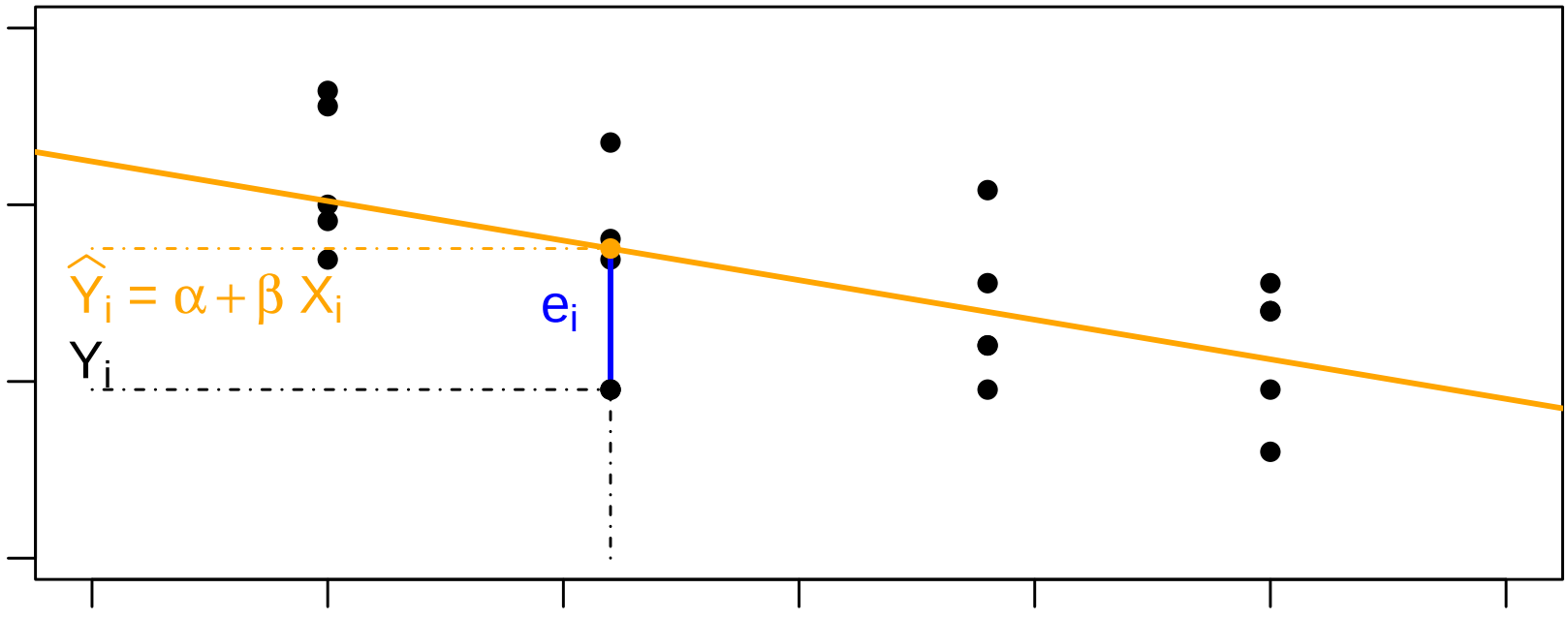
Partie stochastique : modèle gaussien

en supposant des résidus **aléatoires**, **indépendants** ϵ_i suivant une distribution **gaussienne** (normale) de variance constante σ^2.



Estimation des paramètres par la méthode des moindres carrés

Dans le cas de ce modèle, l'estimation du **maximum de vraisemblance** (maximisation de $Pr(Y|\alpha, \beta, \sigma)$) correspond à **l'estimation des moindres carrés** minimisant $SCE = \sum_{i=1}^{n} e_i^2$ avec $e_i = Y_i - \hat{Y}_i$



Estimation des paramètres à l'aide de la fonction `lm()` de R

```
(m <- lm(log10_surv_time ~ temperature, data = dRH50))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ temperature, data = dRH50)
##
## Coefficients:
## (Intercept)  temperature
##      1.1225      -0.0224
```

Résumé de l'ajustement du modèle linéaire

```
summary(m)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ temperature, data = dRH50)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3991 -0.1127 -0.0209  0.1560  0.3443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.12253    0.11411    9.84  1.1e-08 ***
## temperature -0.02239    0.00678   -3.30    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.231 on 18 degrees of freedom
## Multiple R-squared:  0.377,  Adjusted R-squared:  0.342
## F-statistic: 10.9 on 1 and 18 DF,  p-value: 0.00398
```

Comment interpréter la valeur p associée au coefficient de régression (pente) ?

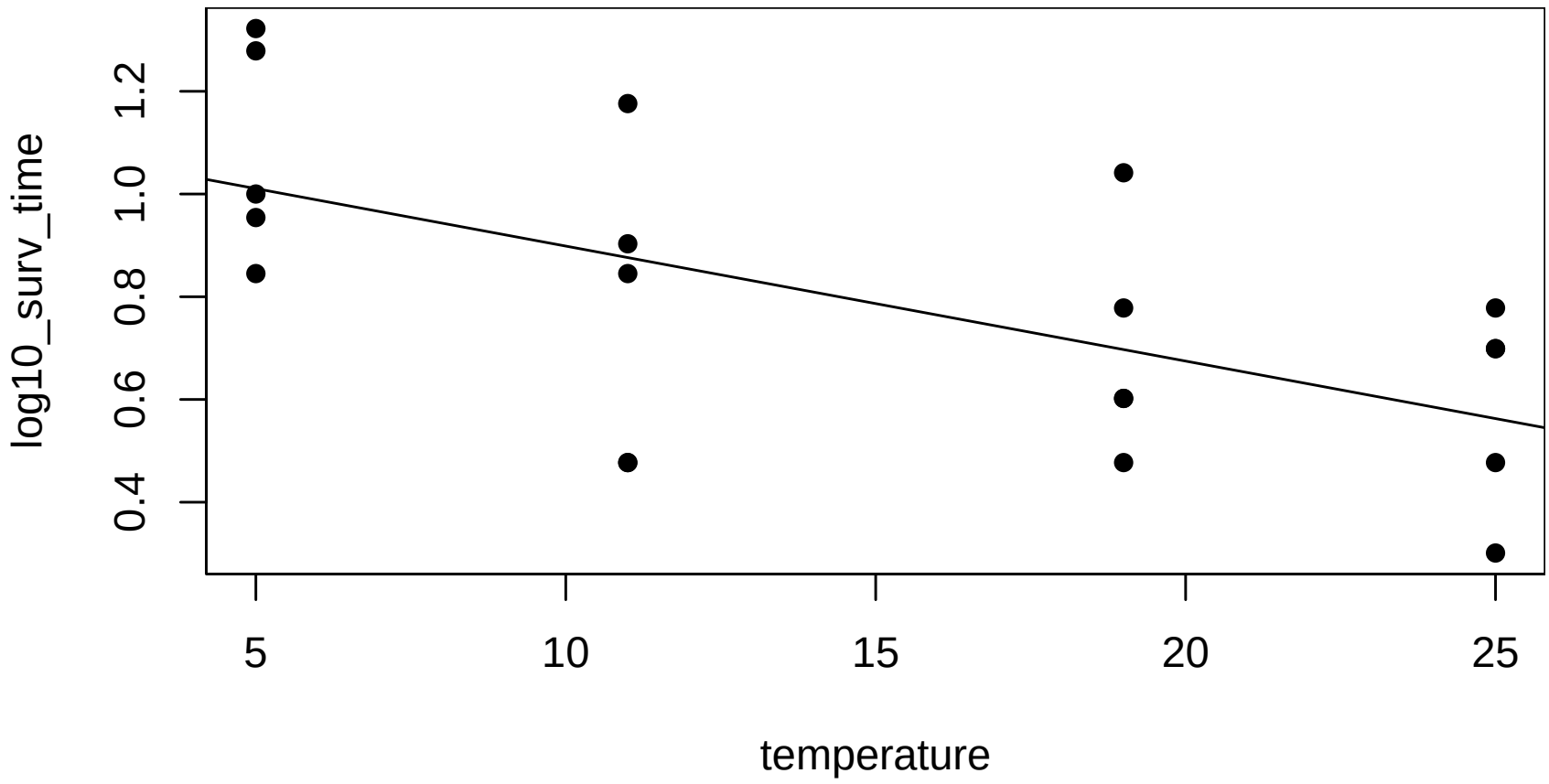
```
summary(m)$coefficients
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.1225    0.11411    9.84 1.15e-08
## temperature     -0.0224    0.00678   -3.30 3.98e-03
```

Il correspond au **test de significativité de la pente** avec H_0 l'hypothèse de pente nulle. Il permet donc en cas de rejet de H_0 de mettre en évidence une relation linéaire significative entre** la variable indépendante X **et** la variable dépendante Y.»

Représentation du modèle ajusté sur les données

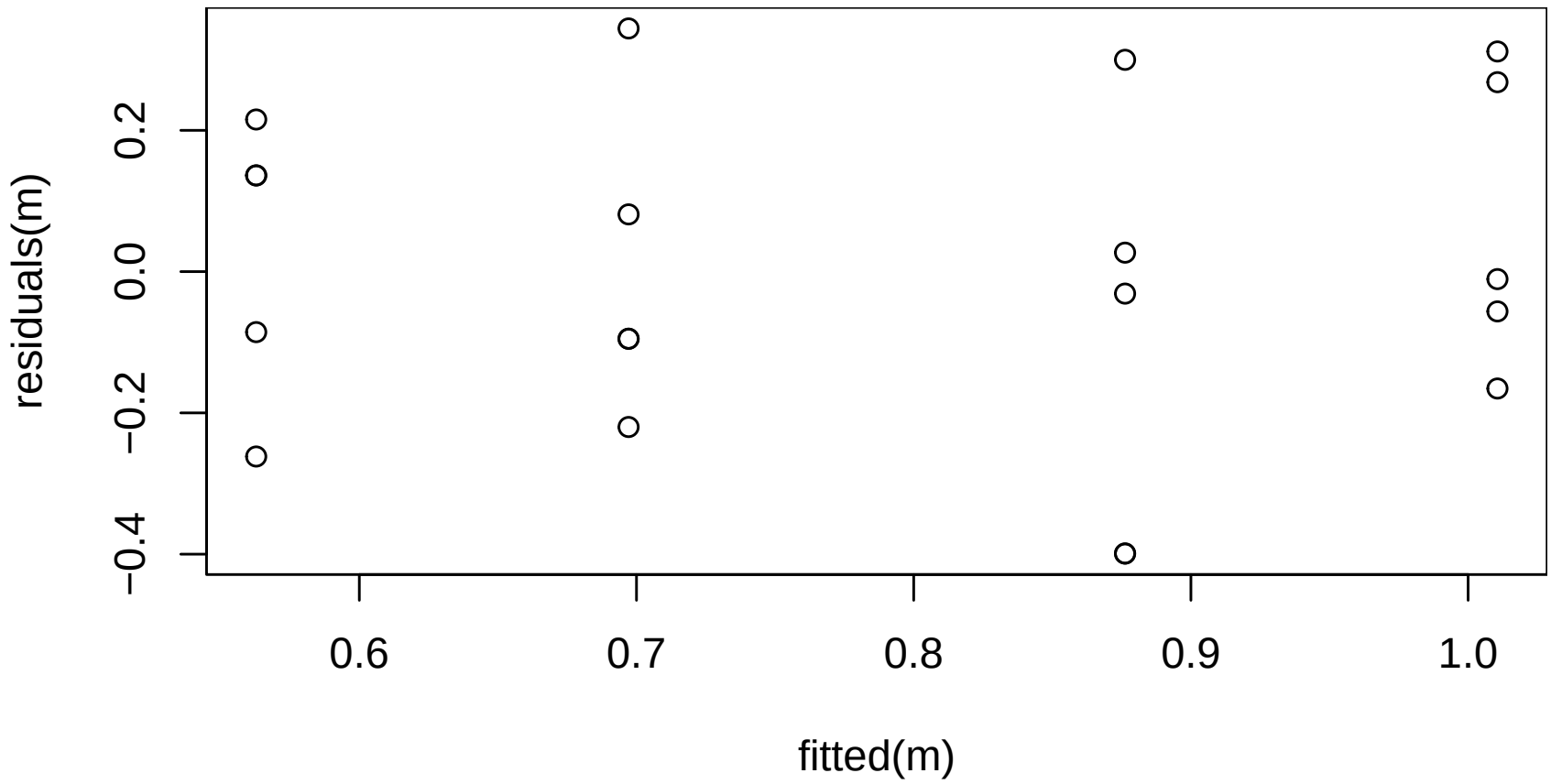
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ temperature, data = dRH50, pch = 16)
abline(m)
```



Vérification des conditions d'utilisation - graphe des résidus

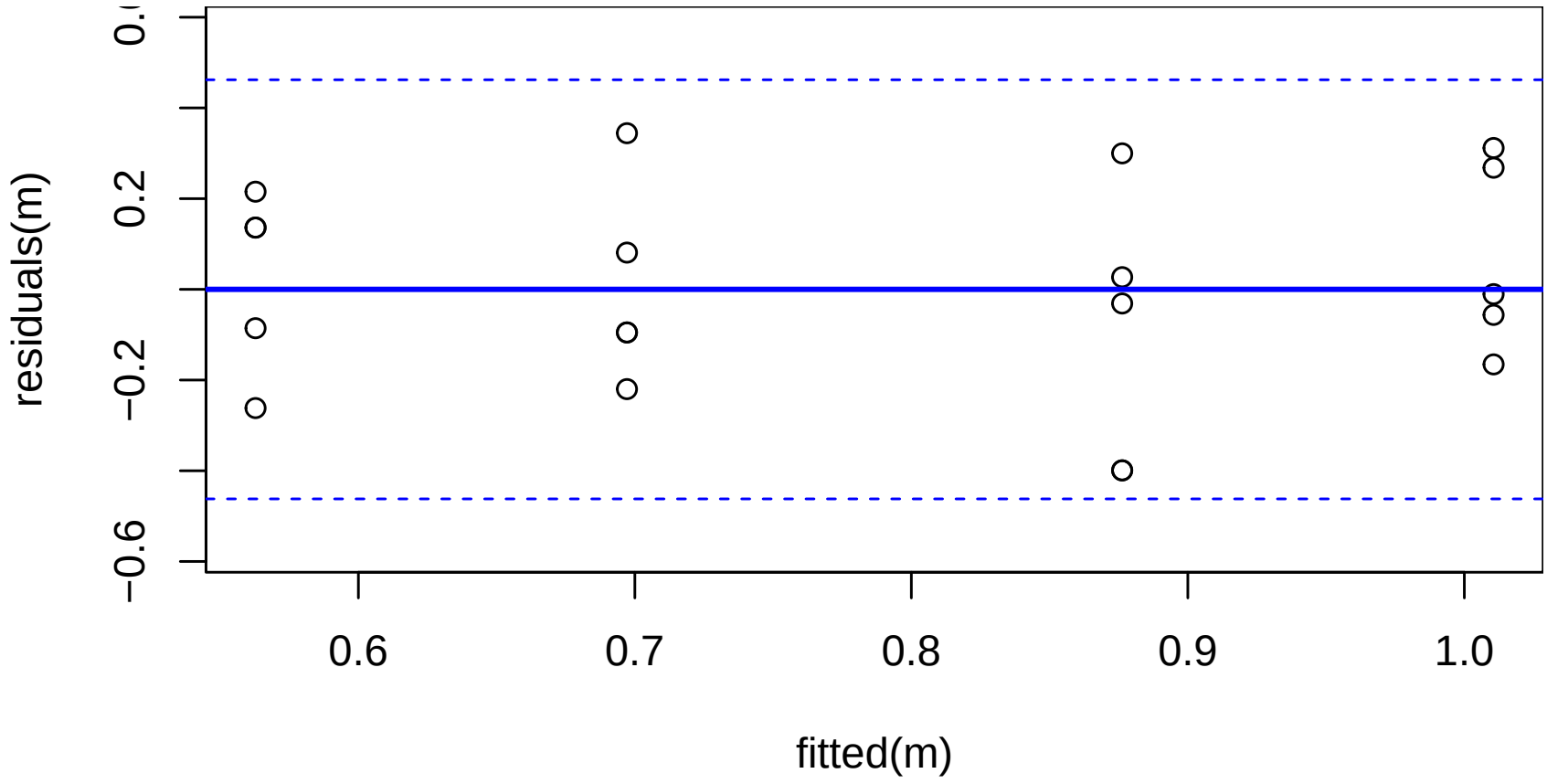
Résidus en fonction des valeurs prédites / ajustées

```
par(mar = c(4, 4, 0, 0))
plot(residuals(m) ~ fitted(m))
```



Graphe des résidus - attendus

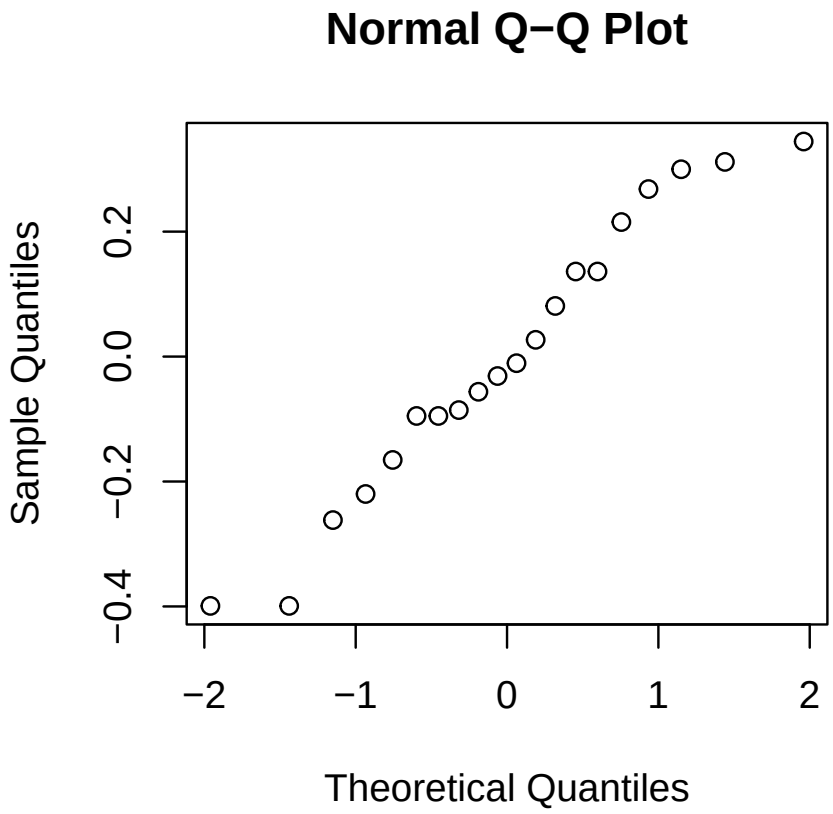
Attendus : centrés sur 0, 95% dans $[-2\sigma; 2\sigma]$, variance constante, sans tendance particulière.



Vérification des conditions d'utilisation - diagramme quantile-quantile des résidus

Distribution globale gaussienne (normale) des résidus attendue : points grossièrement alignés sur le Q-Q plot des résidus

```
qqnorm(residuals(m))
```



A vous de jouer avec trois autres exemples

Ajustez un modèle linéaire simple et examinez en particulier les résidus dans les trois cas suivants :

1. Ajuster un modèle linéaire simple sur le même exemple **sans transformation logarithmique** du temps de survie
2. Modéliser l'**impact de l'humidité relative sur le temps de survie** (après transformation logarithmique de log_{10}) à $25°C$.
3. Modéliser l'**impact de l'humidité relative sur le temps de survie** (après log_{10} de transformation) à $19°C$, en excluant les données dans les conditions les plus arrides.

Vous pouvez utiliser le script R "exos_intro2linmodel.R" pour vous aider si besoin.

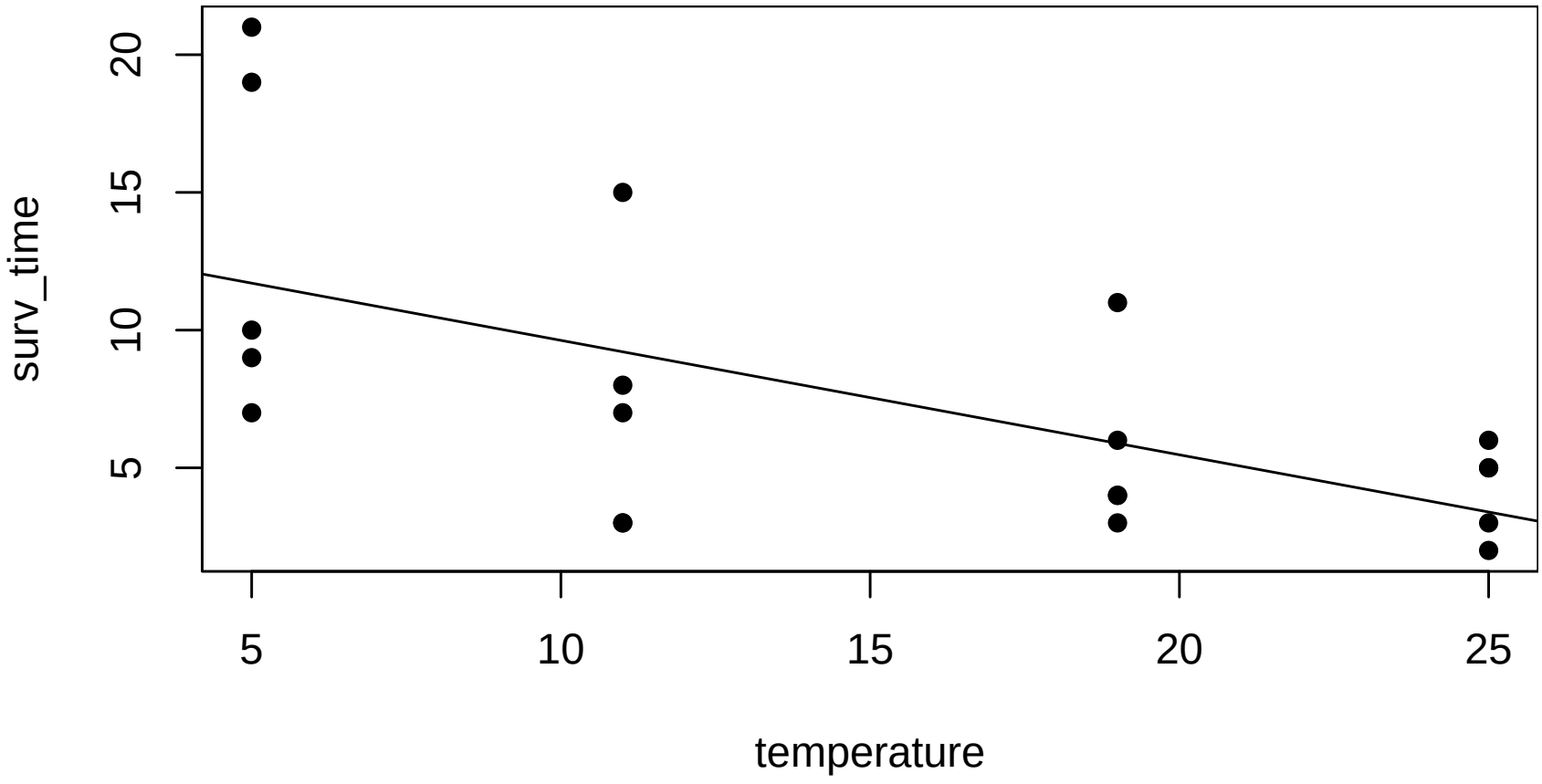
Exemple 1 - ajustement du modèle

```r
mnonlog <- lm(surv_time ~ temperature, data = dRH50)
summary(mnonlog)
```

```
##
## Call:
## lm(formula = surv_time ~ temperature, data = dRH50)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.21  -2.33  -1.30   1.85   9.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.783      2.161    6.38  5.3e-06 ***
## temperature   -0.416      0.128   -3.23   0.0046 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.38 on 18 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.332
## F-statistic: 10.5 on 1 and 18 DF,  p-value: 0.0046
```

Exemple 1 - tracé du modèle ajusté

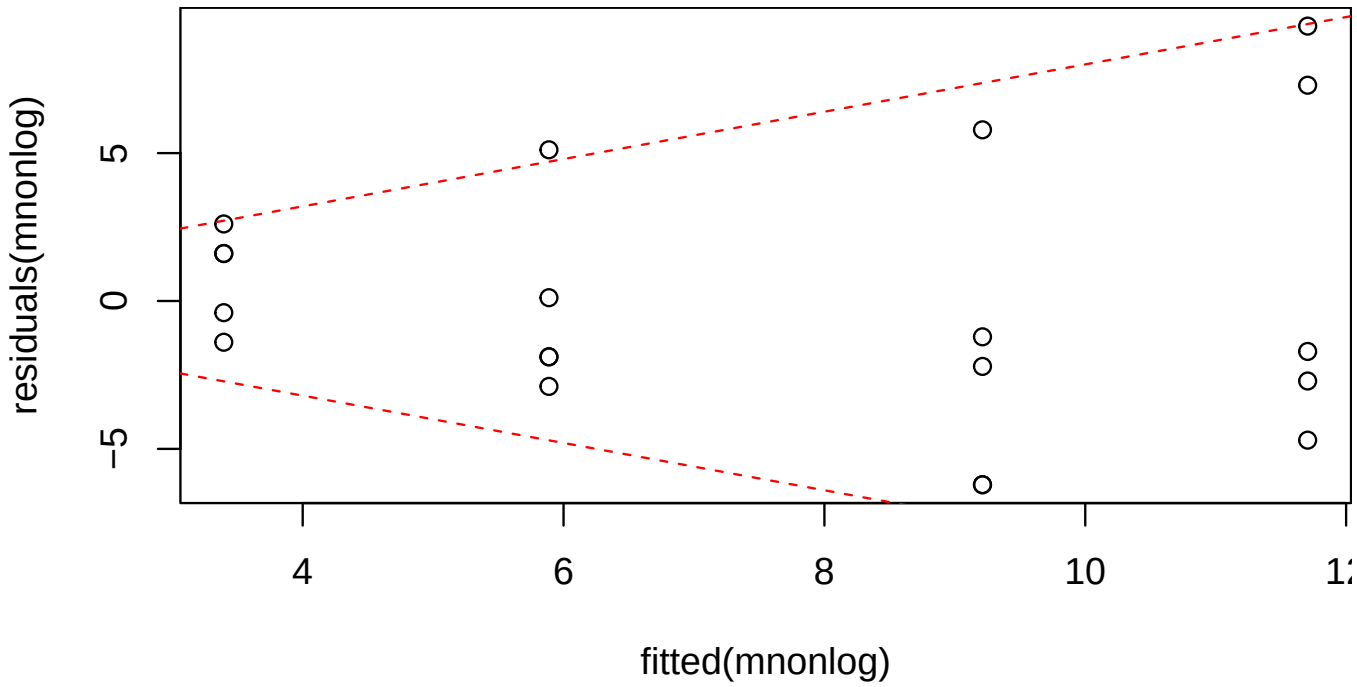
```
par(mar = c(4, 4, 0, 0))
plot(surv_time ~ temperature, data = dRH50, pch = 16)
abline(mnonlog)
```



Exemple 1 - graphe des résidus

Effet entonnoir (ou bottleneck pour les anglophones) = hétéroscédasticité = σ n'est pas constant

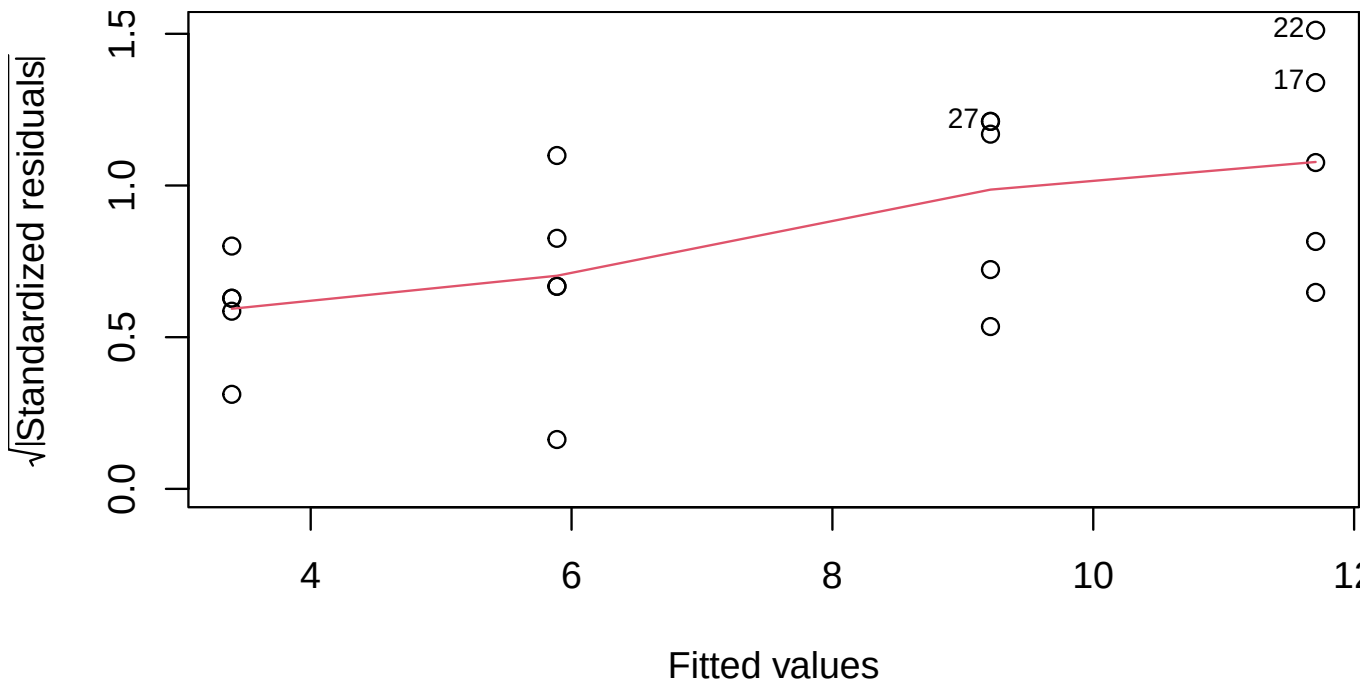
```
plot(residuals(mnonlog) ~ fitted(mnonlog))
```



Exemple 1 - Une variante fournie dans les graphes associés à la fonction `lm()` pour identifier facilement ce type de problème

"Scale-Location plot"

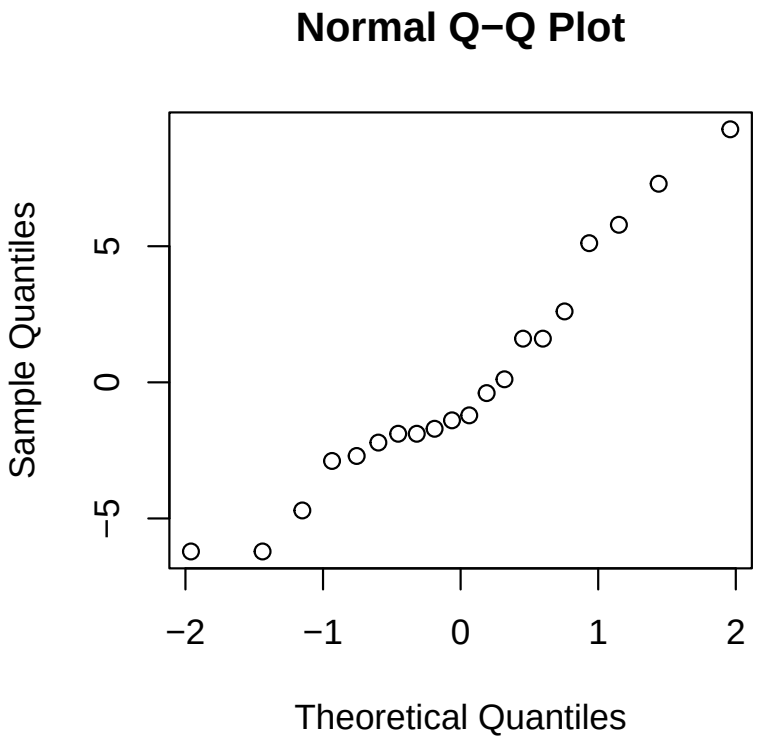
```
par(mar = c(4, 4, 0, 0))
plot(mnonlog, which = 3)
```



Exemple 1 - diagramme quantile-quantile des résidus

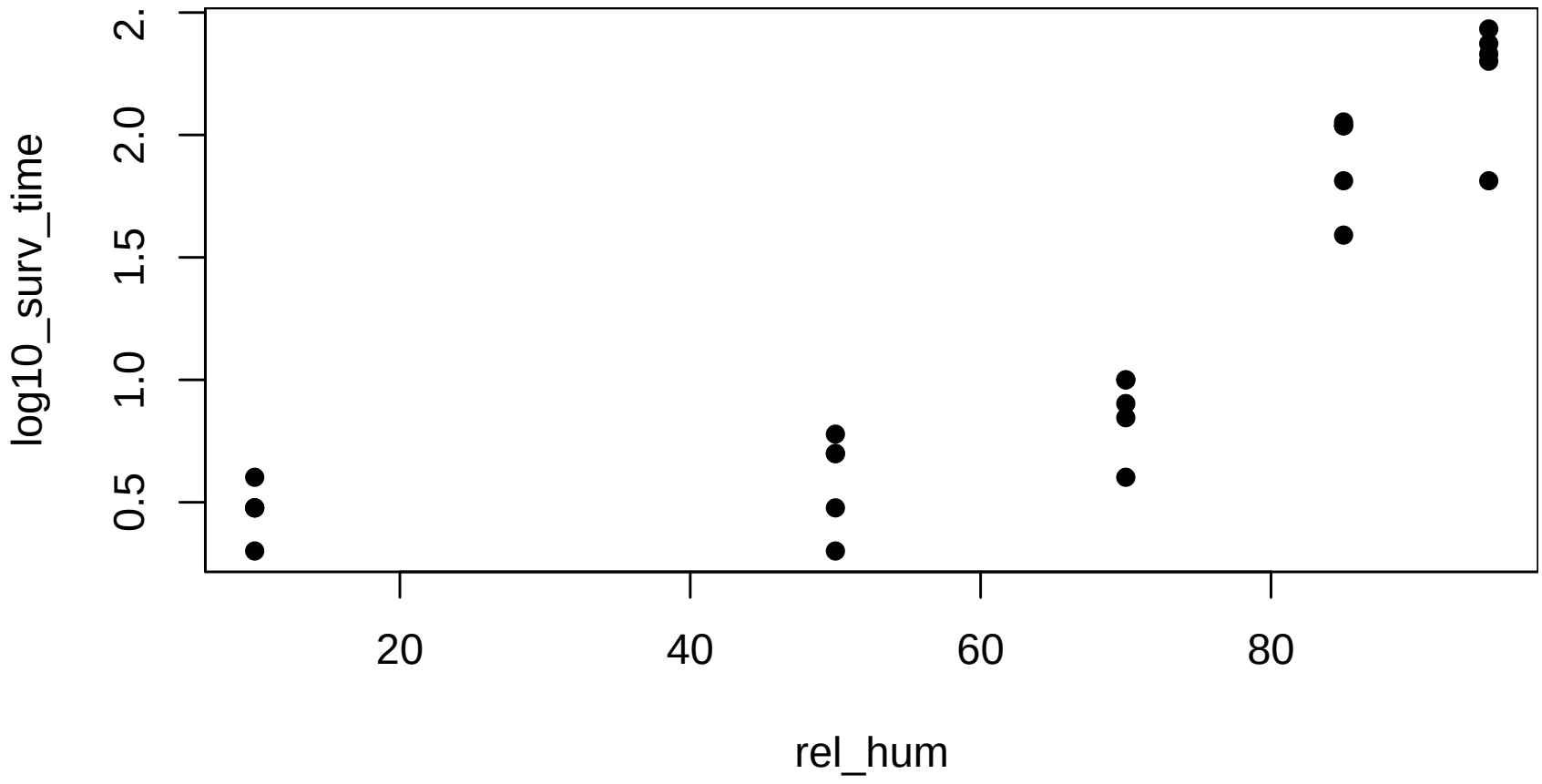
Dans cet exemple, le problème n'est pas détectable sur le Q-Q plot des résidus.

```
qqnorm(residuals(mnonlog))
```



Exemple 2 - construction et examen du jeu de données

```
dT25 <- subset(dtot, temperature == 25)
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT25, pch = 16)
```



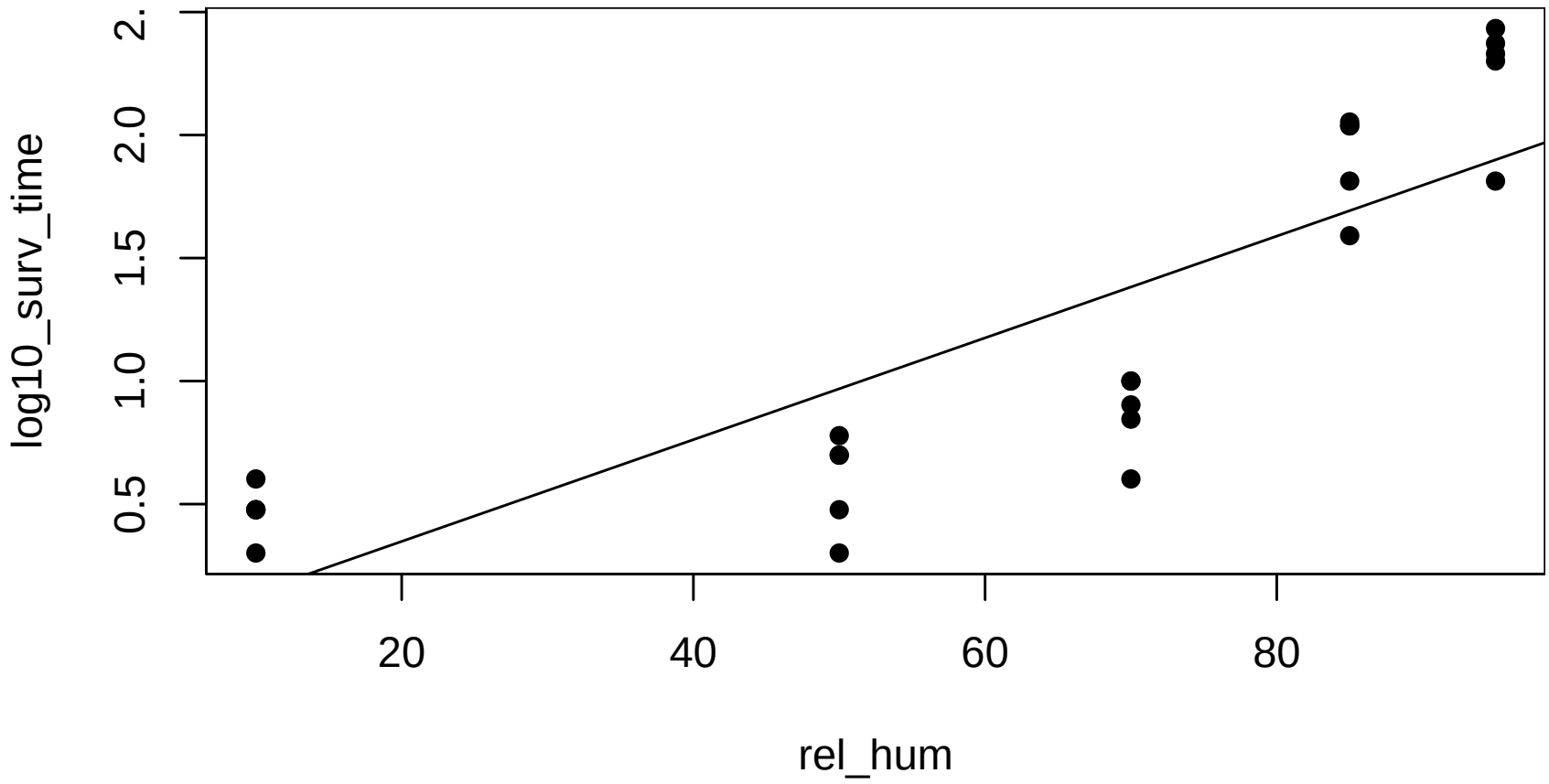
Exemple 2 - ajustement du modèle

```r
mnonlin <- lm(log10_surv_time ~ rel_hum, data = dT25)
summary(mnonlin)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum, data = dT25)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.780 -0.382  0.120  0.345  0.534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06545    0.19383   -0.34     0.74
## rel_hum      0.02068    0.00281    7.35  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.423 on 23 degrees of freedom
## Multiple R-squared:  0.702,  Adjusted R-squared:  0.689
## F-statistic: 54.1 on 1 and 23 DF,  p-value: 1.76e-07
```

Exemple 2 - tracé du modèle

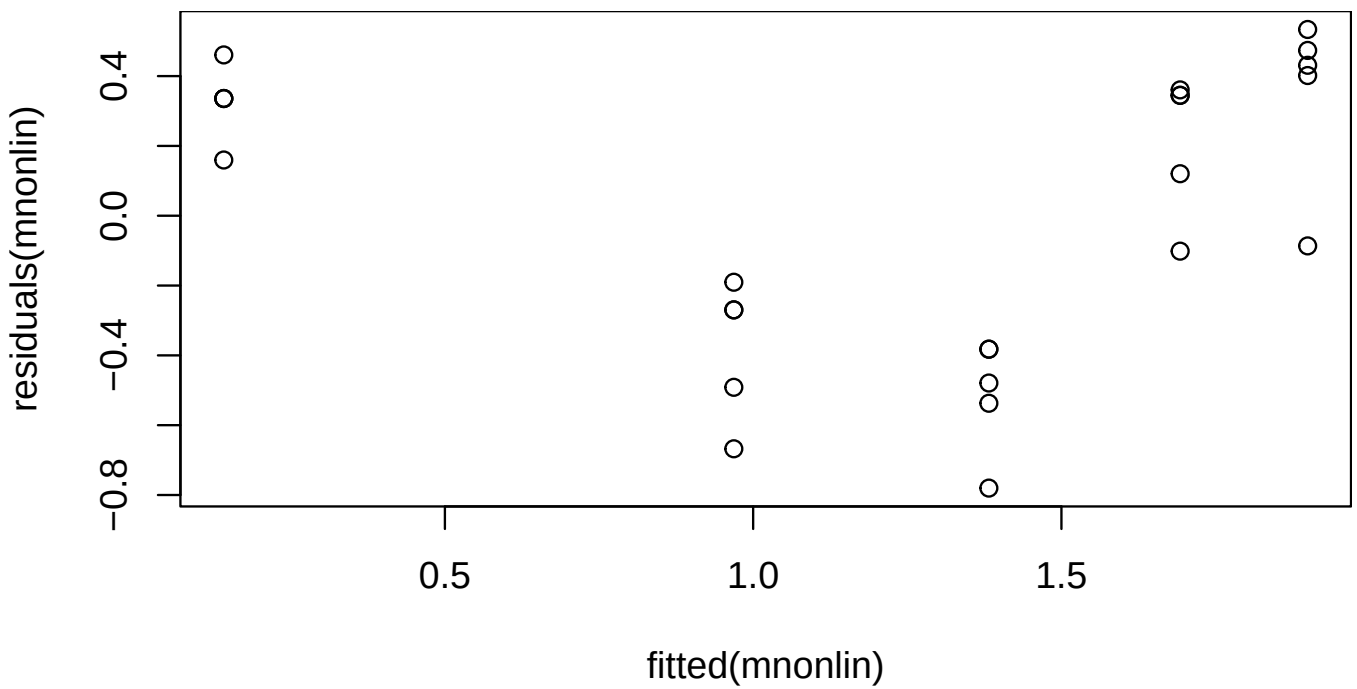
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT25, pch = 16)
abline(mnonlin)
```



Exemple 2 - Graphe des résidus

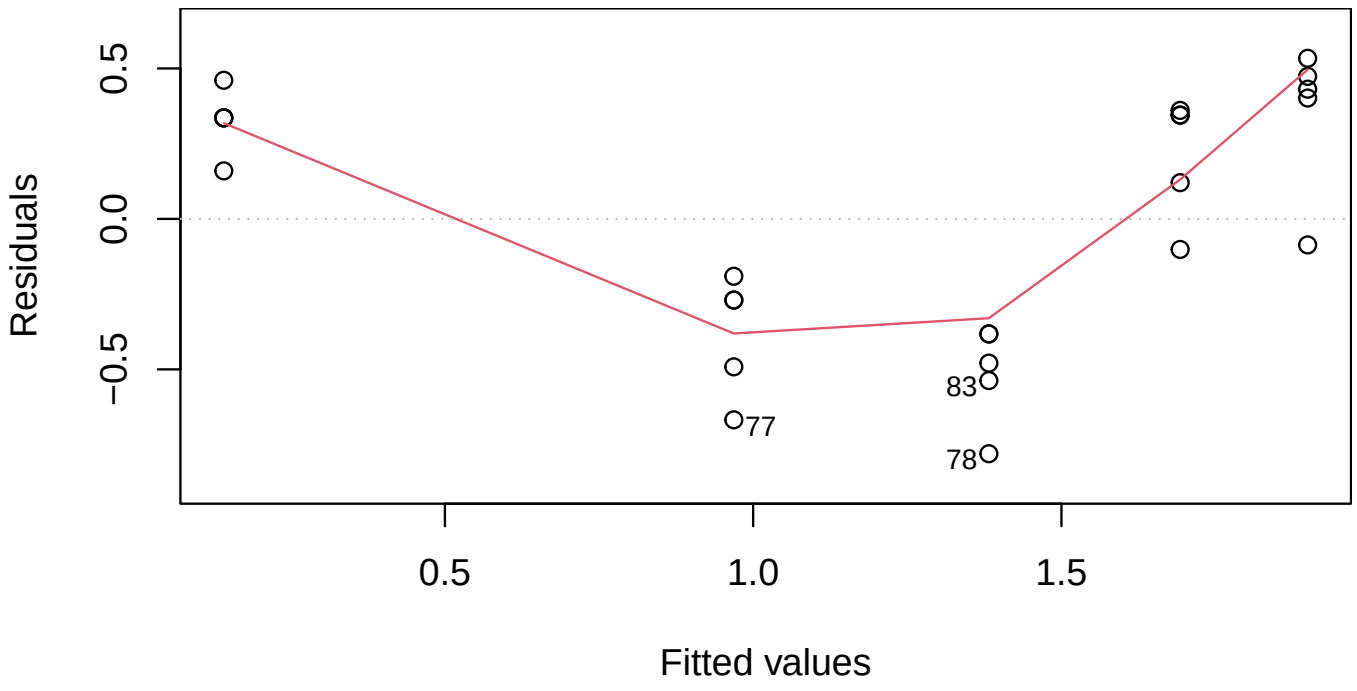
Tendance des résidus = les résidus ne sont pas indépendants, dans cet exemple en raison de la non linéarité de la relation

```
par(mar = c(4, 4, 0, 0))
plot(residuals(mnonlin) ~ fitted(mnonlin))
```



Exemple 2 - Une variante fournie dans les graphes associés à la fonction `lm()` pour identifier facilement ce type de problème

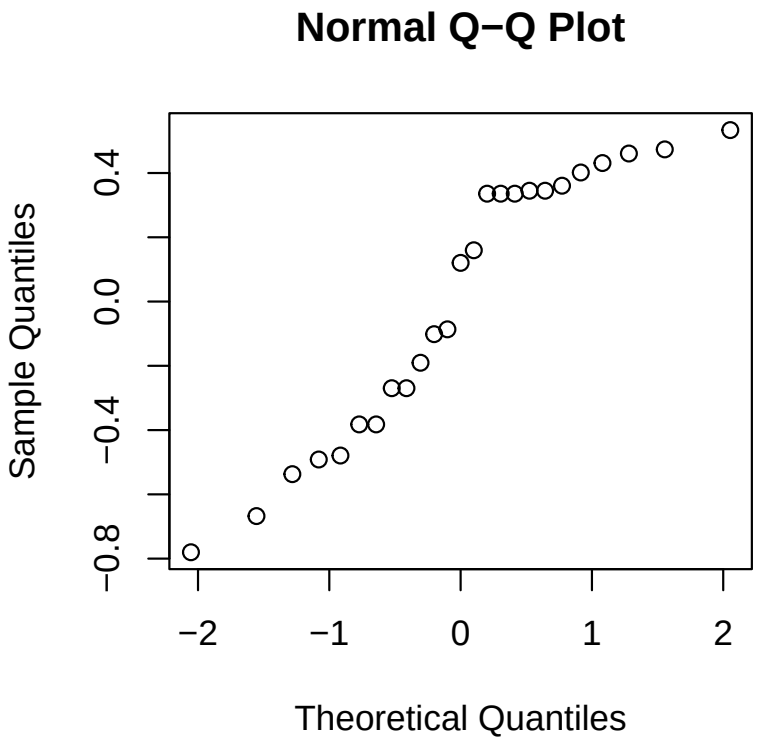
```
par(mar = c(4, 4, 0, 0))
plot(mnonlin, which = 1)
```



Exemple 2 - diagramme quantile-quantile des résidus

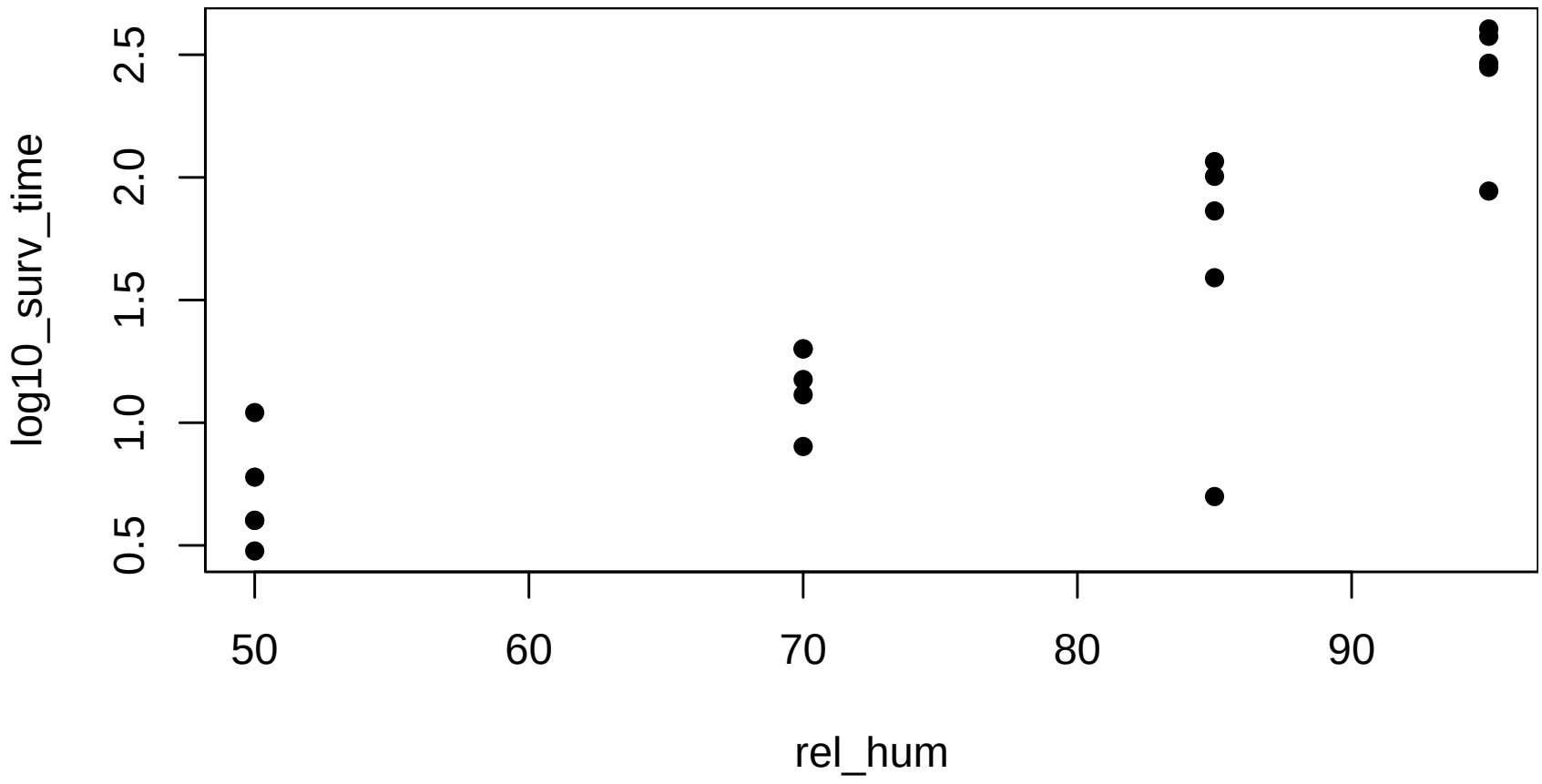
Dans cet exemple, le problème est également détectable sur le Q-Q plot des résidus.

```
qqnorm(residuals(mnonlin))
```



Exemple 3 - construction et examen du jeu de données

```
dT19 <- subset(dtot, temperature == 19 & rel_hum > 10)
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT19, pch = 16)
```



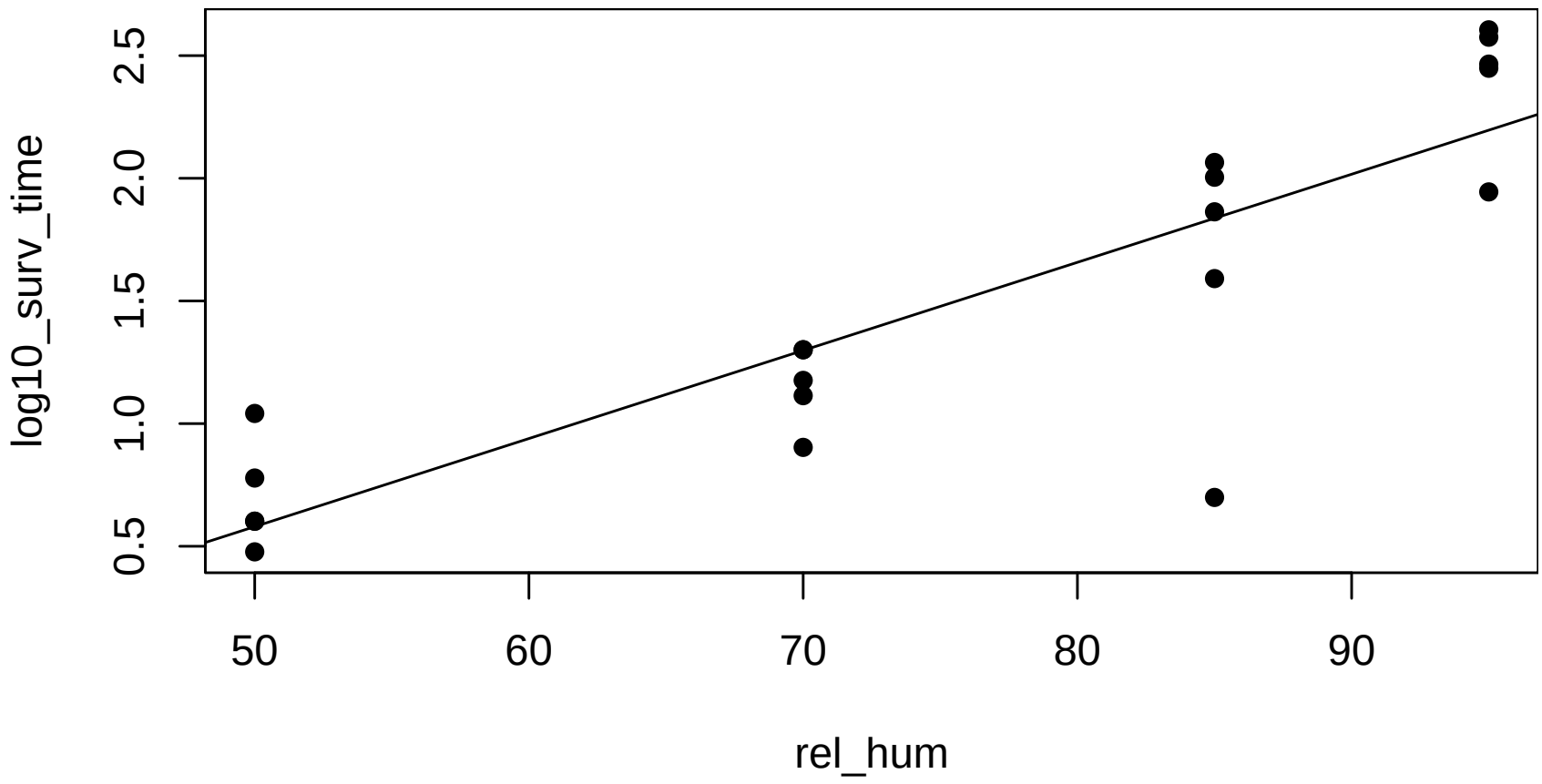
Exemple 3 - ajustement du modèle

```r
moutlier <- lm(log10_surv_time ~ rel_hum, data = dT19)
summary(moutlier)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum, data = dT19)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1380 -0.1377  0.0221  0.2337  0.4614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.21571    0.37147   -3.27   0.0042 **
## rel_hum      0.03591    0.00483    7.43  6.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.366 on 18 degrees of freedom
## Multiple R-squared:  0.754,  Adjusted R-squared:  0.741
## F-statistic: 55.3 on 1 and 18 DF,  p-value: 6.85e-07
```

Exemple 3 - tracé du modèle

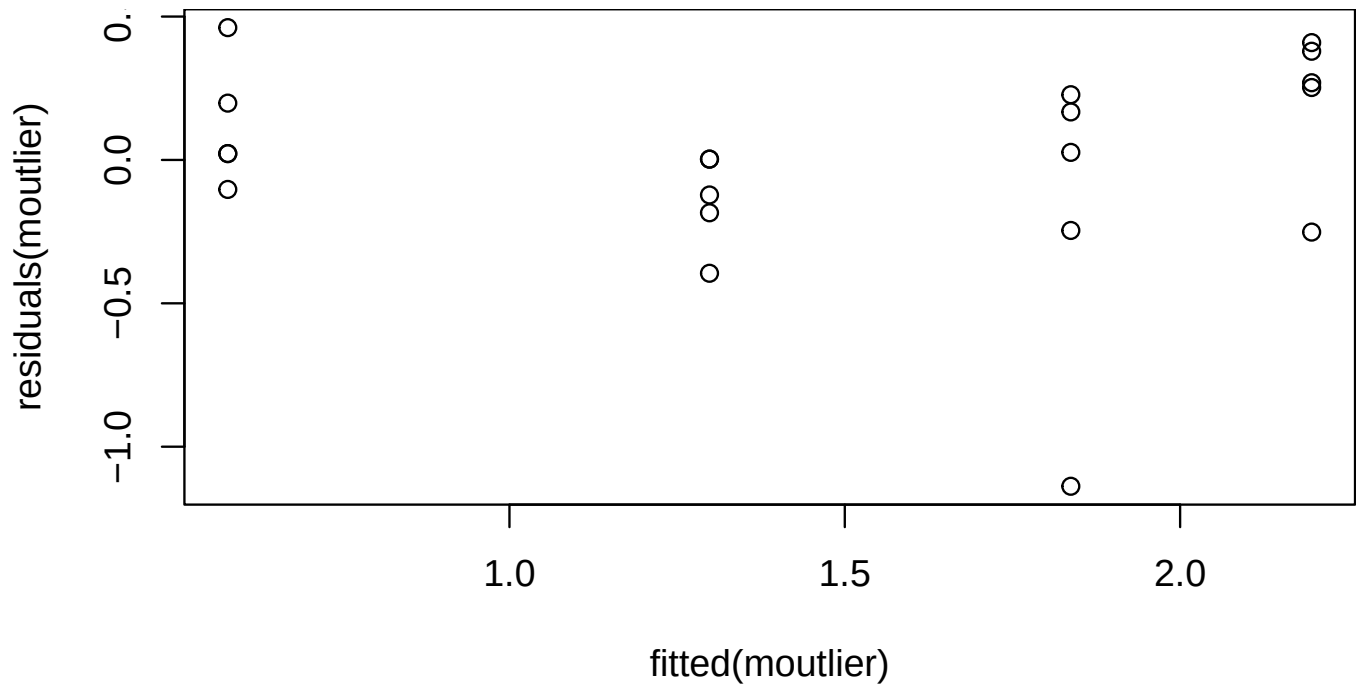
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT19, pch = 16)
abline(moutlier)
```



Exemple 3 - Graphe des résidus

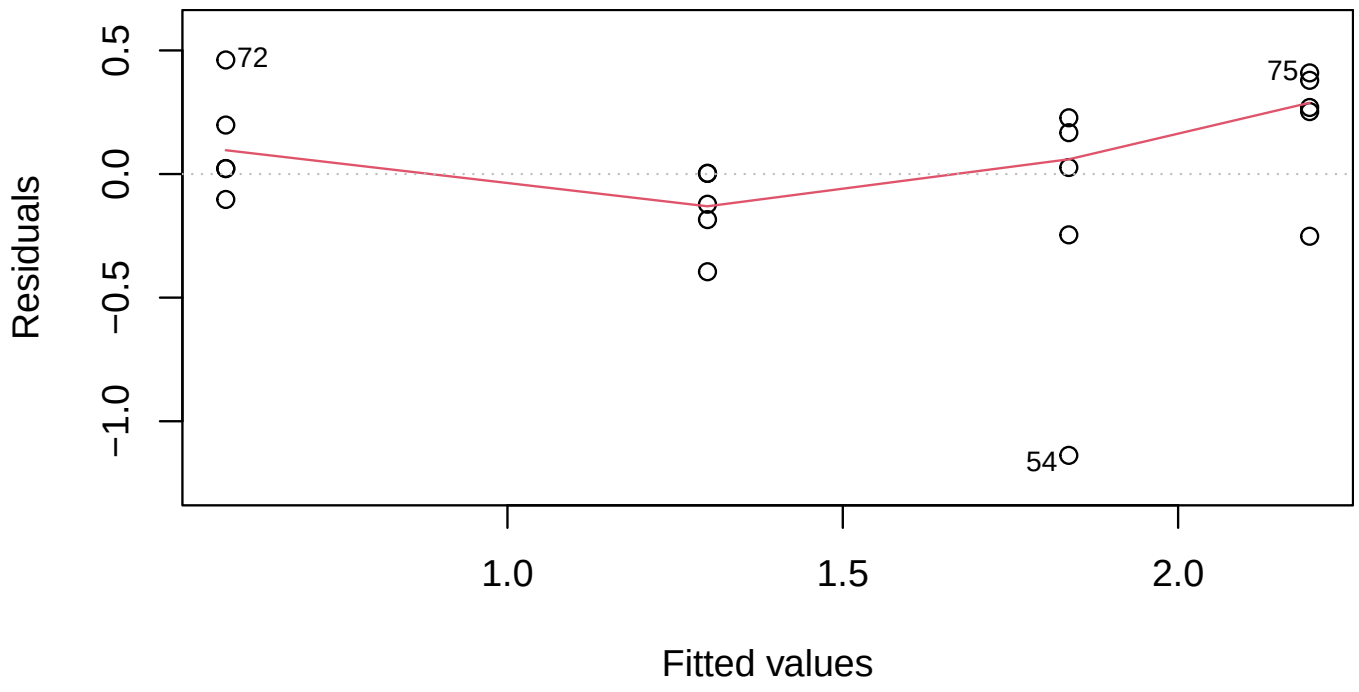
Une valeur extrême apparaît nettement sur le graphe des résidus.

```
par(mar = c(4, 4, 0, 0))
plot(residuals(moutlier) ~ fitted(moutlier))
```



Exemple 3 - Une variante fournie dans les graphes associés à la fonction `lm()` pour identifier les valeurs extrêmes

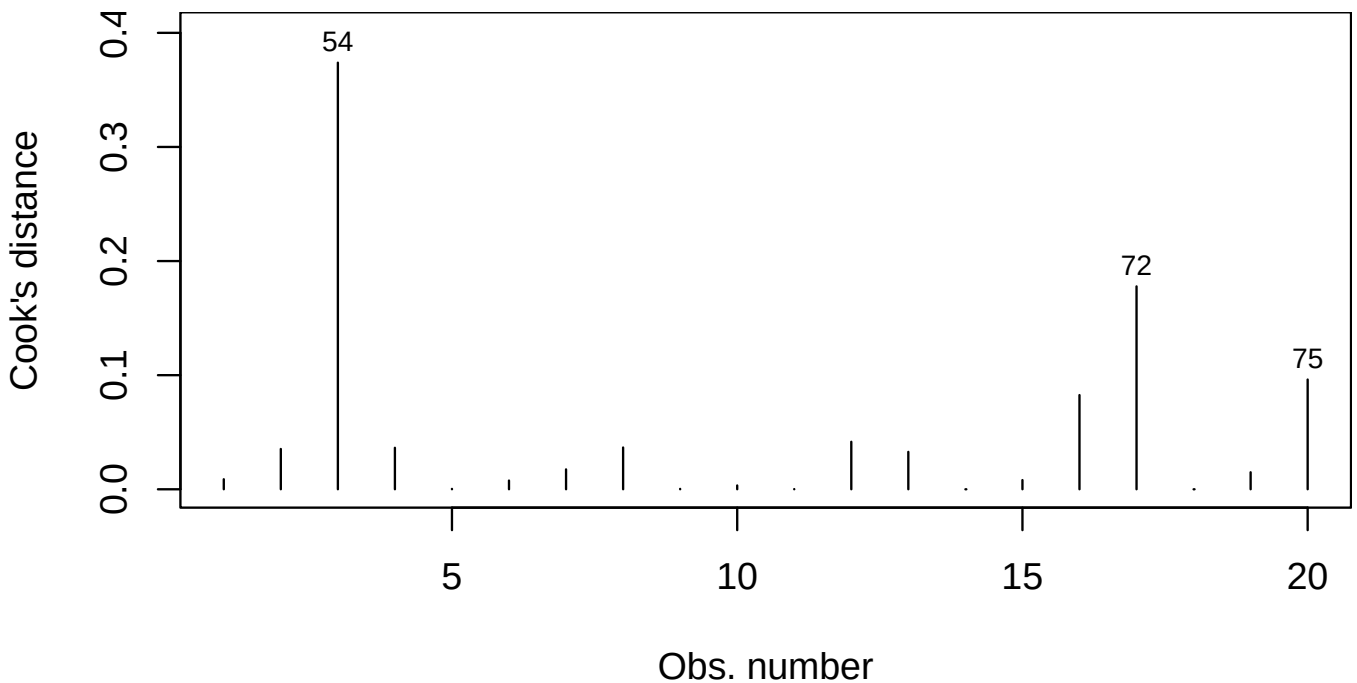
```
par(mar = c(4, 4, 0, 0))
plot(moutlier, which = 1)
```



Exemple 3 - Distances de Cook : impact / influence des valeurs extrêmes ?

Pour identifier les **observations influentes** : impact de l'élimination de chaque observation sur les estimations des paramètres.

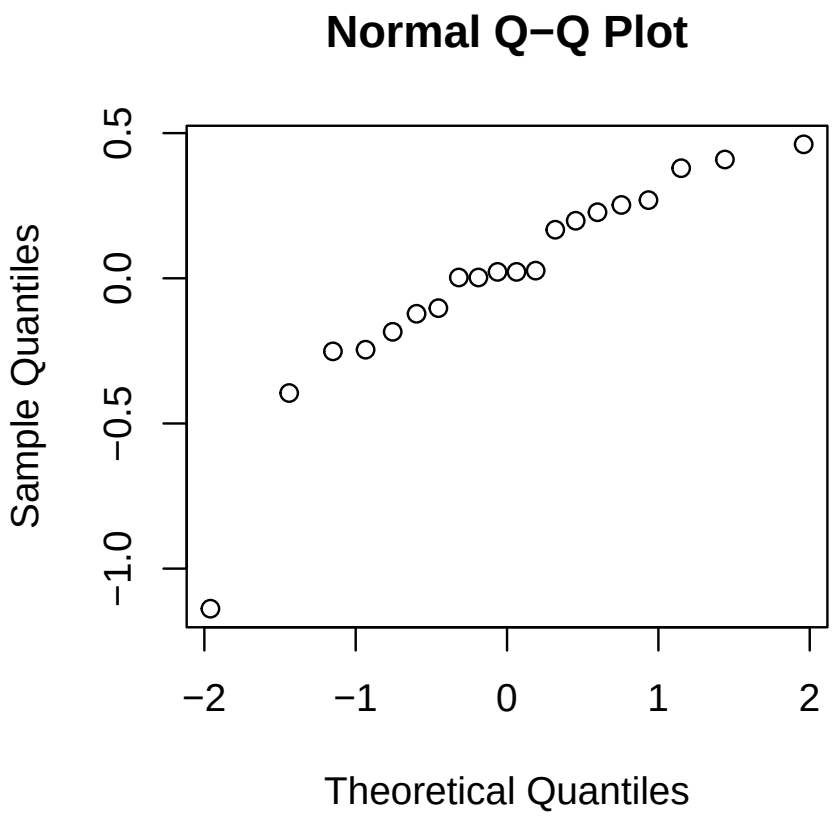
```
par(mar = c(4, 4, 0, 0))
plot(moutlier, which = 4)
```



Example 3 - Diagramme quantile-quantile des résidus

Dans cet exemple, le problème est également détectable sur le Q-Q plot des résidus.

```
qqnorm(residuals(moutlier))
```



Inférence à l'aide d'une régression linéaire simple

Possible si votre modèle n'est pas invalidé par l'examen des résidus.

Reprenons notre exemple initial qui respecte les conditions de la régression linéaire

Estimations avec leur intervalle de confiance

```
coef(m)
```

```
## (Intercept) temperature
##      1.1225      -0.0224
```

```
confint(m)
```

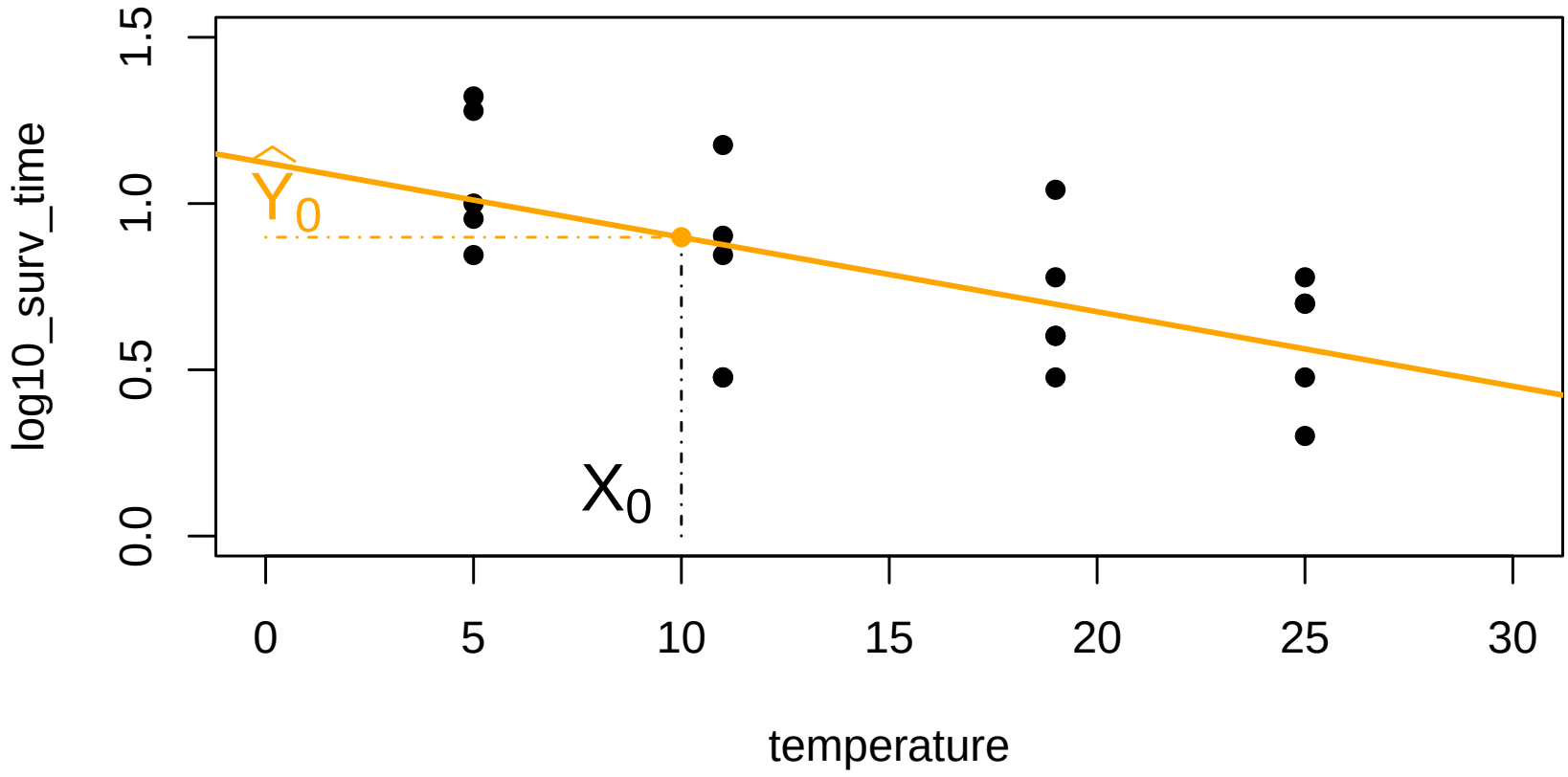
```
##                 2.5 %   97.5 %
## (Intercept)  0.8828  1.36228
## temperature -0.0366 -0.00814
```

- **ordonnée à l'origine (intercept)** : valeur estimée de Y pour $X = 0$ (ne peut avoir d'interprétation biologique que si 0 se situe dans l'intervalle des valeurs observées de X)
- **pente (coefficient de régression)** : **changement de la variable à expliquer** correspondant à **un changement d'une unité de la variable explicative**

Prédiction à partir du modèle

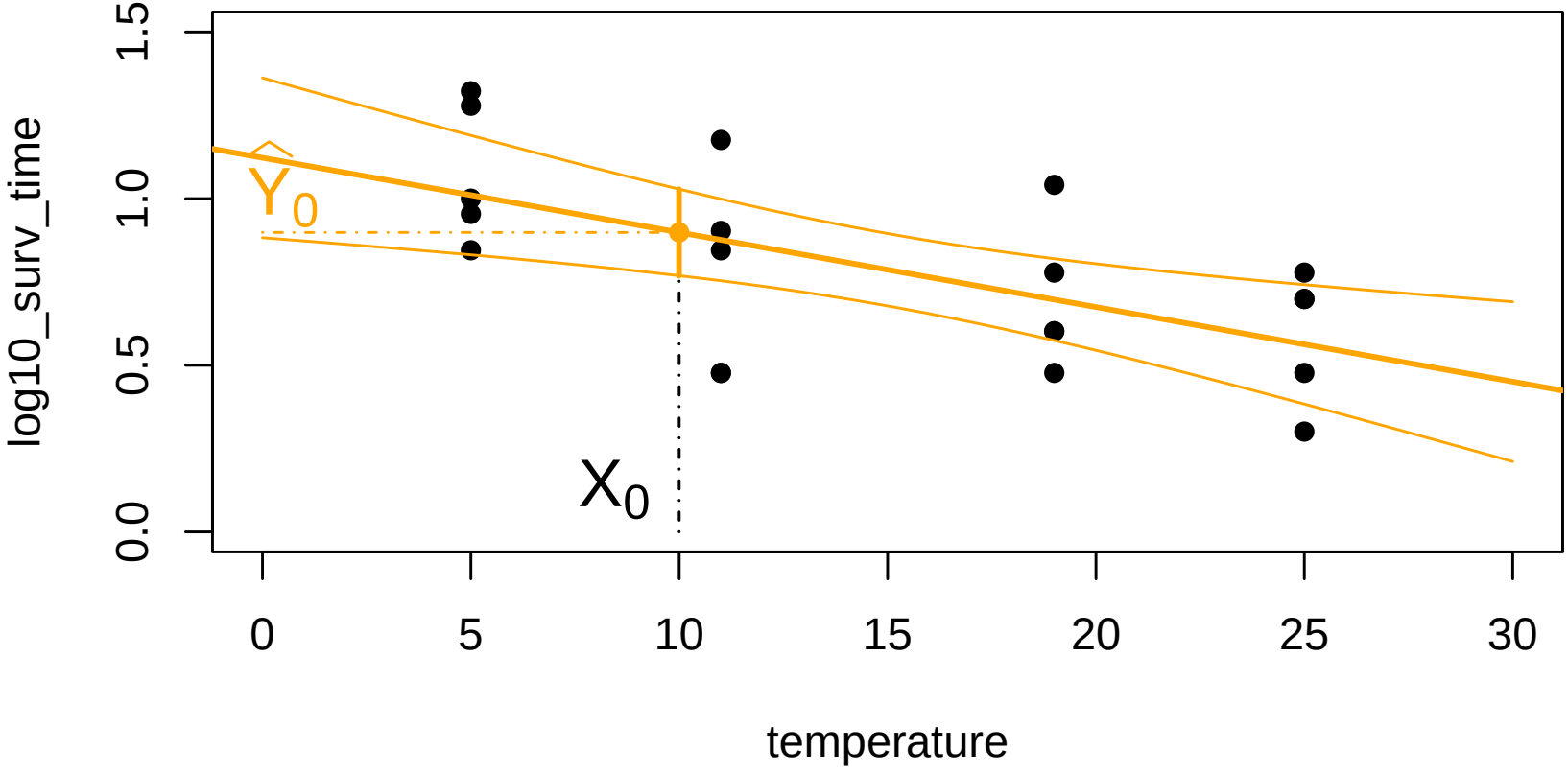
Prediction de Y_0 pour $X = X_0$ choisis dans la gamme observée.

ATTENTION: pas d'extrapolation !



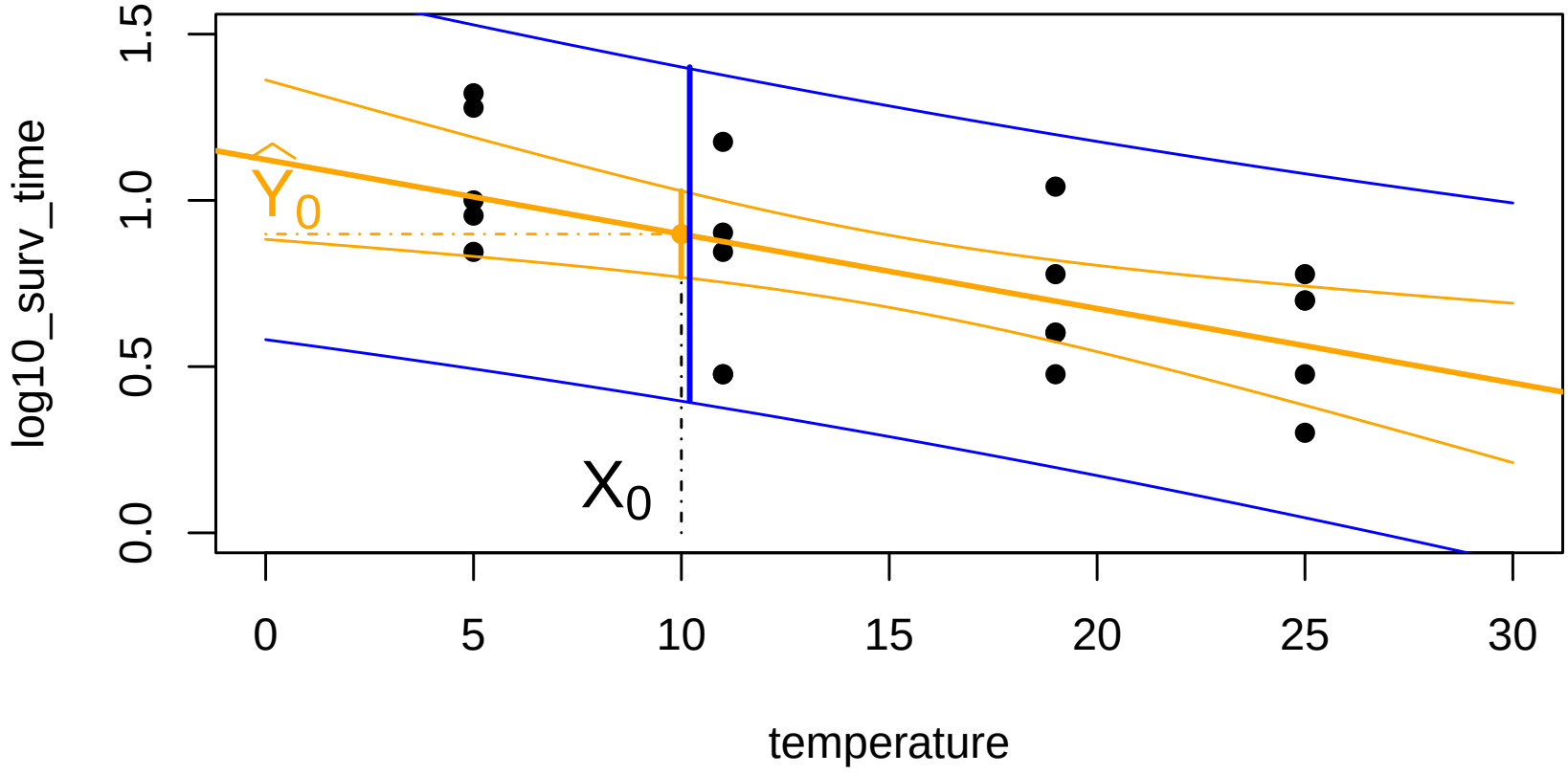
Intervalle de confiance sur la moyenne prédite

Incertitude sur la partie déterministe du modèle (la droite)



Intervalle de prédiction = intervalle de confiance sur une prédiction individuelle

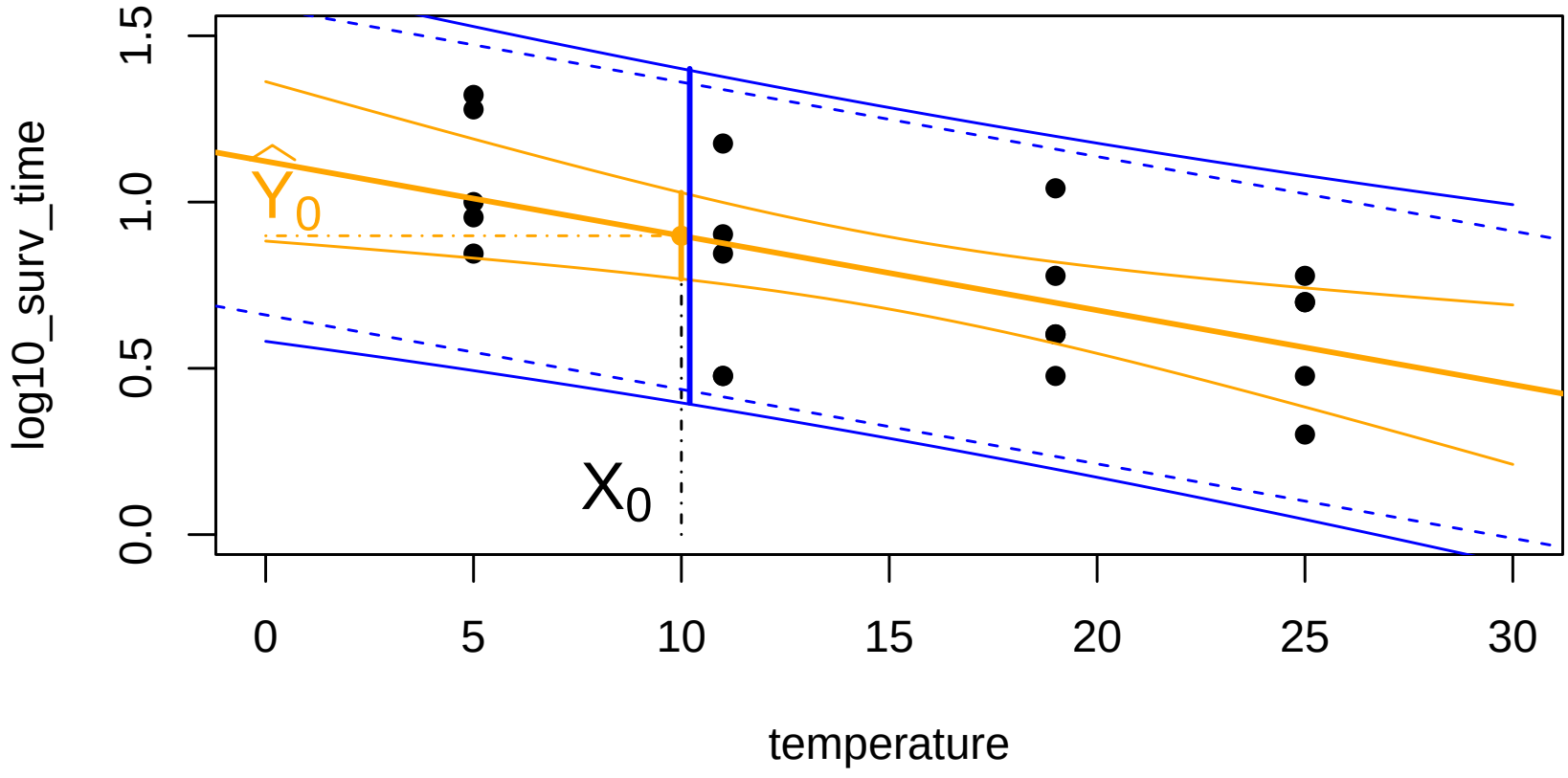
Incertitude sur la partie déterministe + la partie stochastique



Approximation de l'intervalle de prédiction

Incertitude sur la partie déterministe + la partie stochastique souvent approchée par défaut par la seule partie stochastique

$\hat{Y}_0 \pm 2 \times \sigma$



Calcul des intervalles de confiance et de prédiction avec R

Intervalle de confiance sur la moyenne

```
data4pred <- data.frame(temperature = 10)
predict(m, interval = "confidence", newdata = data4pred)
```

```
##      fit   lwr  upr
## 1 0.899 0.769 1.03
```

Intervalle de prédiction (individuelle)

```
predict(m, interval = "prediction", newdata = data4pred)
```

```
##      fit   lwr upr
## 1 0.899 0.396 1.4
```

Statistiques d'ajustement

Interprétation du r^2 (coefficient de détermination)

Il correspond à la **proportion de la variation (somme des carrés des écarts à la moyenne) de la variable à expliquer que la variable explicative explique réellement** (par la relation déterministe linéaire).

Le r^2 donné en % est parfois appelé **le pourcentage de variation expliquée**.

L'écart-type σ est également une mesure de la qualité de l'ajustement, mais il doit être interprété en fonction de l'ordre de grandeur de Y.

Résumé de l'ajustement d'un modèle linéaire dans R

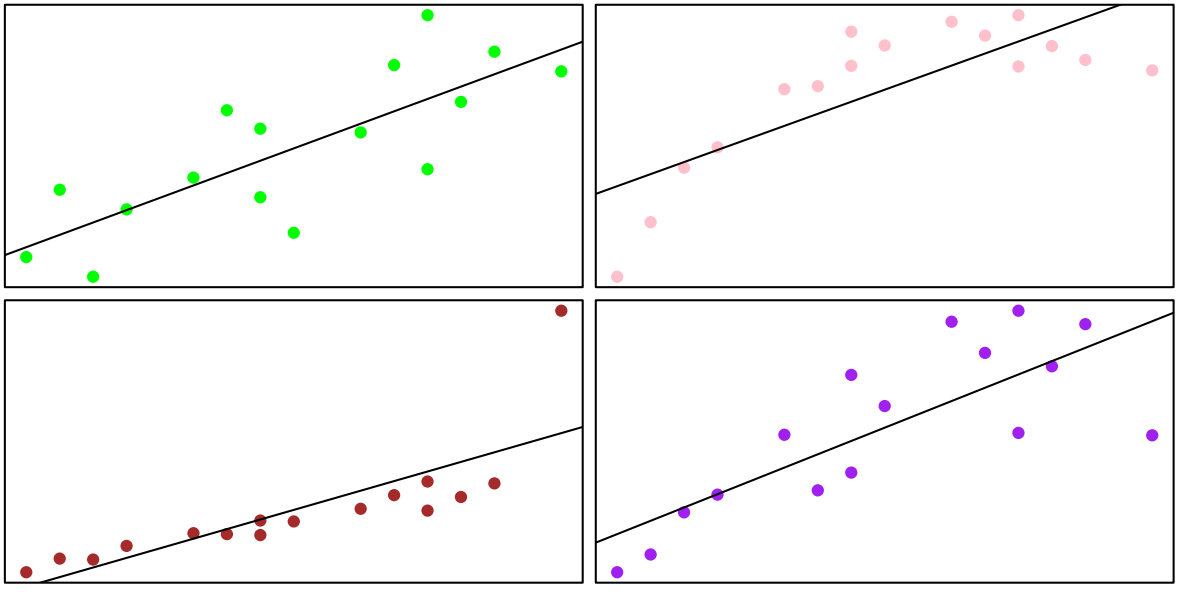
```
summary(m)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ temperature, data = dRH50)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3991 -0.1127 -0.0209  0.1560  0.3443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.12253    0.11411    9.84  1.1e-08 ***
## temperature -0.02239    0.00678   -3.30    0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.231 on 18 degrees of freedom
## Multiple R-squared:  0.377,  Adjusted R-squared:  0.342
## F-statistic: 10.9 on 1 and 18 DF,  p-value: 0.00398
```

Une valeur de r^2 proche de 1 ne vous informe pas sur le respect des conditions d'utilisation du modèle.

Pour vous en convaincre, regardez les quatre exemples suivants qui partagent exactement la même valeur de r^2, soit 62%.

extrait de R. Tomassone et al., 1992, La régression, nouveaux regards sur une ancienne méthode statistique.



A votre tour d'utiliser le modèle en inférence

Utilisez le modèle `m` pour répondre aux questions suivantes :

1. A une humidité relative de 50%, quelle est la **variation attendue du taux de survie (en log_{10})** attendue pour une **augmentation de la température de** $1°C$?
2. Traduire cette variation en un **coefficient mutiplicatif sur le temps de survie** et en sa **diminution relative**.
3. Donner une prédiction (avec son intervalle de confiance à 95%) **de la moyenne du temps de survie en échelle logarithmique**, des tiques à une humidité relative de 50% et à une **température de** $22°C$, et sa traduction en échelle brute (en jours).
4. Donner une prédiction (avec son IC à 95%) du **temps de survie d'une tique** exposée à une humidité relative de 50% et à une **température de** $\mathbf{22}°C$.
5. Donner une prédiction (avec son IC à 95%) de **la durée de survie d'une tique** exposée à une humidité relative de 50% et à une **température de** $40°C$.

Réponse à la question 1

A une humidité relative de 50%, quelle est la **variation attendue du taux de survie (en log_{10})** attendue pour une **augmentation de la température de** $1°C$?

```
coef(m)[2]
```

```
## temperature
##     -0.0224
```

Réponse à la question 2

Traduire cette variation en un **coefficient mutiplicatif sur le temps de survie** et en sa **diminution relative**.

si nous nommons st_0 le temps de survie initial, et st_c le temps de survie après le changement, on attend $log_{10}(st_c) - log_{10}(st_0) = b$ avec b le coefficient de régression, donc $log_{10}(\frac{st_c}{st_0}) = b$ donc $st_c = 10^b \times st_0$.

```
# multiplication by
10^coef(m)[2]
```

```
## temperature
##        0.95
```

```
# so a relative diminution of 5%
```

Réponse à la question 3

Donner une prédiction (avec son intervalle de confiance à 95%) **de la moyenne du temps de survie en échelle logarithmique**, des tiques à une humidité relative de 50% et à une **température de** $22°C$, et sa traduction en échelle brute (en jours).

```
data4pred <- data.frame(temperature = 22)
# prediction in log10(days)
(stinlog10 <- predict(m, interval = "confidence",
                      newdata = data4pred))
```

```
##    fit   lwr   upr
## 1 0.63 0.483 0.777
```

```
# prediction in days
10^stinlog10
```

```
##    fit  lwr  upr
## 1 4.27 3.04 5.99
```

Réponse à la question 4

Donner une prédiction (avec son IC à 95%) du **temps de survie d'une tique** exposée à une humidité relative de 50% et à une **température de 22**$^\circ C$.

```
data4pred <- data.frame(temperature = 22)
# prediction in log10(days)
(stinlog10 <- predict(m, interval = "prediction",
                      newdata = data4pred))
```

```
##     fit   lwr  upr
## 1 0.63 0.123 1.14
```

```
# prediction in days
10^stinlog10
```

```
##     fit  lwr  upr
## 1 4.27 1.33 13.7
```

Réponse à la question 5

Donner une prédiction (avec son IC à 95%) de **la durée de survie d'une tique** exposée à une humidité relative de 50% et à une **température de** $40°C$.

```
# Should we do that ?
data4pred <- data.frame(temperature = 40)
predict(m, interval = "prediction", newdata = data4pred)
```

```
##     fit    lwr   upr
## 1 0.227 -0.385 0.839
```

```
# NO !!!!!!!!!!!!!!!
```

ATTENTION EXTRAPOLATION INTERDITE !

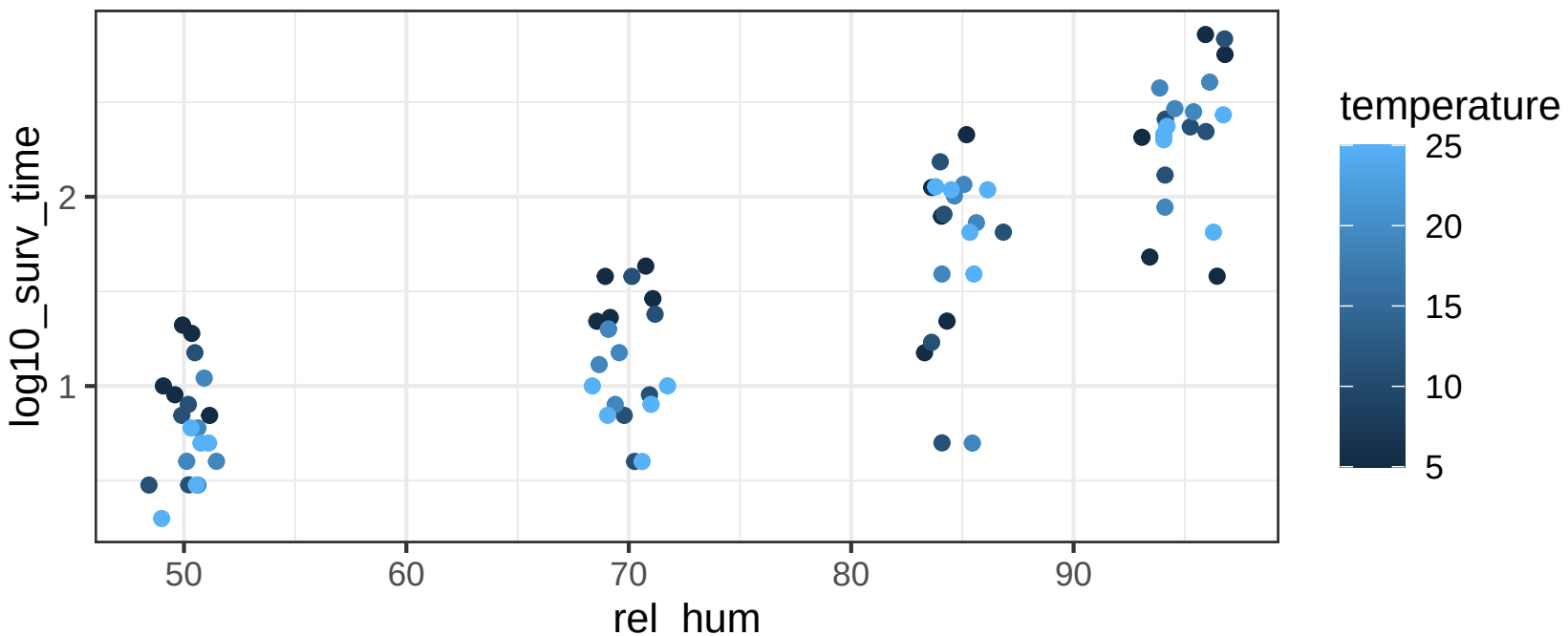
Ce jeu de données ne peut pas nous informer sur ce qui se passe au-dessus de $25°C$!

Régression linéaire multiple

Modélisation de l'impact de l'humidité relative et de la température hors conditions arrides

en utilisant un sous-ensemble des données dans toutes les conditions d'humidité **à l'exception de la condition la plus sèche**.

```
dhum <- subset(dtot, rel_hum > 10)
ggplot(data = dhum, aes(x = rel_hum, y = log10_surv_time,
col = temperature)) + geom_jitter(width = 2)
```



Le modèle théorique

Très similaire au modèle linéaire simple,

avec **plus d'une variable indépendante continue** (appelés aussi régresseurs),

donc **plus d'un coefficient de régression** (on ne parle plus de pente).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

avec $\epsilon_i \sim N(0, \sigma)$

Partie déterministe : lien linéaire
Partie stochastique : modèle gaussien

en supposant des résidus **aléatoires**, **indépendants** ϵ_i suivant une distribution **gaussienne** (normale) de variance constante σ^2.

L'estimation des paramètres par la méthode des moindres carrés

Comme dans le cas du modèle linéaire simple,

l'estimation du **maximum de vraisemblance**

(maximisant $Pr(Y|\beta_0, \beta_1, \ldots, \beta_p, \sigma)$) correspond toujours à **l'estimation des moindres carrés** minimisant $SCE = \sum_{i=1}^{n} e_i^2$ avec $e_i = Y_i - \hat{Y}_i$

Estimation des paramètres à l'aide de R

```r
(mm <- lm(log10_surv_time ~ rel_hum + temperature, data = dhum))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperature, data = dhum)
##
## Coefficients:
## (Intercept)      rel_hum  temperature
##    -0.86084      0.03334     -0.00971
```

Résumé de l'ajustement

```
summary(mm)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperature, data = dhum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1677 -0.2040  0.0649  0.2482  0.6337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.86084    0.21090   -4.08  0.00011 ***
## rel_hum      0.03334    0.00252   13.25  < 2e-16 ***
## temperature -0.00971    0.00560   -1.73  0.08691 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.382 on 77 degrees of freedom
## Multiple R-squared:  0.699,  Adjusted R-squared:  0.691
## F-statistic: 89.3 on 2 and 77 DF,  p-value: <2e-16
```

Affichage des coefficients avec leur intervalle de confiance à 95%

```
cbind(estimate = coef(mm), confint(mm))
```

```
##                estimate   2.5 %   97.5 %
## (Intercept) -0.86084 -1.2808 -0.44089
## rel_hum      0.03334  0.0283  0.03835
## temperature -0.00971 -0.0209  0.00144
```

Comment interpréter les valeurs p associées à chaque coefficient de régression ?

```
summary(mm)$coefficients
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.86084    0.21090   -4.08 1.08e-04
## rel_hum      0.03334    0.00252   13.25 1.46e-21
## temperature -0.00971    0.00560   -1.73 8.69e-02
```

Chaque valeur p correspond au **test de signification de chaque coefficient de régression** avec H_0 l'hypothèse nulle de chaque coefficient, les autres étant conservés dans le modèle. Il permet donc de mettre en évidence une relation linéaire significative entre** le régresseur X_i **et** le résultat Y **lorsque les autres régresseurs** X_j avec $j \neq i$ **ont déjà été pris en compte**.»

Dans cet exemple, nous voyons un impact significatif de l'humidité relative sur le taux de survie, mais l'ajout de la température comme second régresseur n'améliore pas significativement le modèle.

Une autre façon équivalente de comparer deux modèles emboîtés à l'aide d'un test F - impact de la température

```
mm <- lm(log10_surv_time ~ rel_hum + temperature, data = dhum)
mrel_hum <- lm(log10_surv_time ~ rel_hum, data = dhum)
anova(mm, mrel_hum)
```

```
## Analysis of Variance Table
##
## Model 1: log10_surv_time ~ rel_hum + temperature
## Model 2: log10_surv_time ~ rel_hum
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     77 11.2
## 2     78 11.6 -1    -0.438 3.01  0.087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

L'ajout de la température dans le modèle n'améliore pas significativement l'ajustement.

Une autre façon de comparer deux modèles emboîtés à l'aide d'un test F - impact de l'humidité relative

```
mm <- lm(log10_surv_time ~ rel_hum + temperature, data = dhum)
mtemperature <- lm(log10_surv_time ~ temperature, data = dhum)
anova(mm, mtemperature)
```

```
## Analysis of Variance Table
##
## Model 1: log10_surv_time ~ rel_hum + temperature
## Model 2: log10_surv_time ~ temperature
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1     77 11.2
## 2     78 36.8 -1     -25.6 176 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 
```

L'ajout de l'humidité relative dans le modèle améliore significativement l'ajustement.

Possibilité de faire tous les tests F en enlevant les coefficients de régression un à un automatiquement

```
drop1(mm, test = "F")
```

```
## Single term deletions
##
## Model:
## log10_surv_time ~ rel_hum + temperature
##             Df Sum of Sq  RSS    AIC F value Pr(>F)
## <none>                    11.2 -151.2
## rel_hum      1     25.57 36.8  -58.2  175.69 <2e-16 ***
## temperature  1      0.44 11.6 -150.2    3.01  0.087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Vérification des conditions d'utilisation

Il n'est plus possible de tracer le modèle ajusté sur les données en 2D.

C'est l'une des difficultés rencontrées dans le processus de vérification de l'adéquation du modèle aux données !

Tracé des résidus en fonction des valeurs ajustées, comme dans le cas d'une régression linéaire simple

```
par(mar = c(4, 4, 0, 0))
plot(residuals(mm) ~ fitted(mm))
```

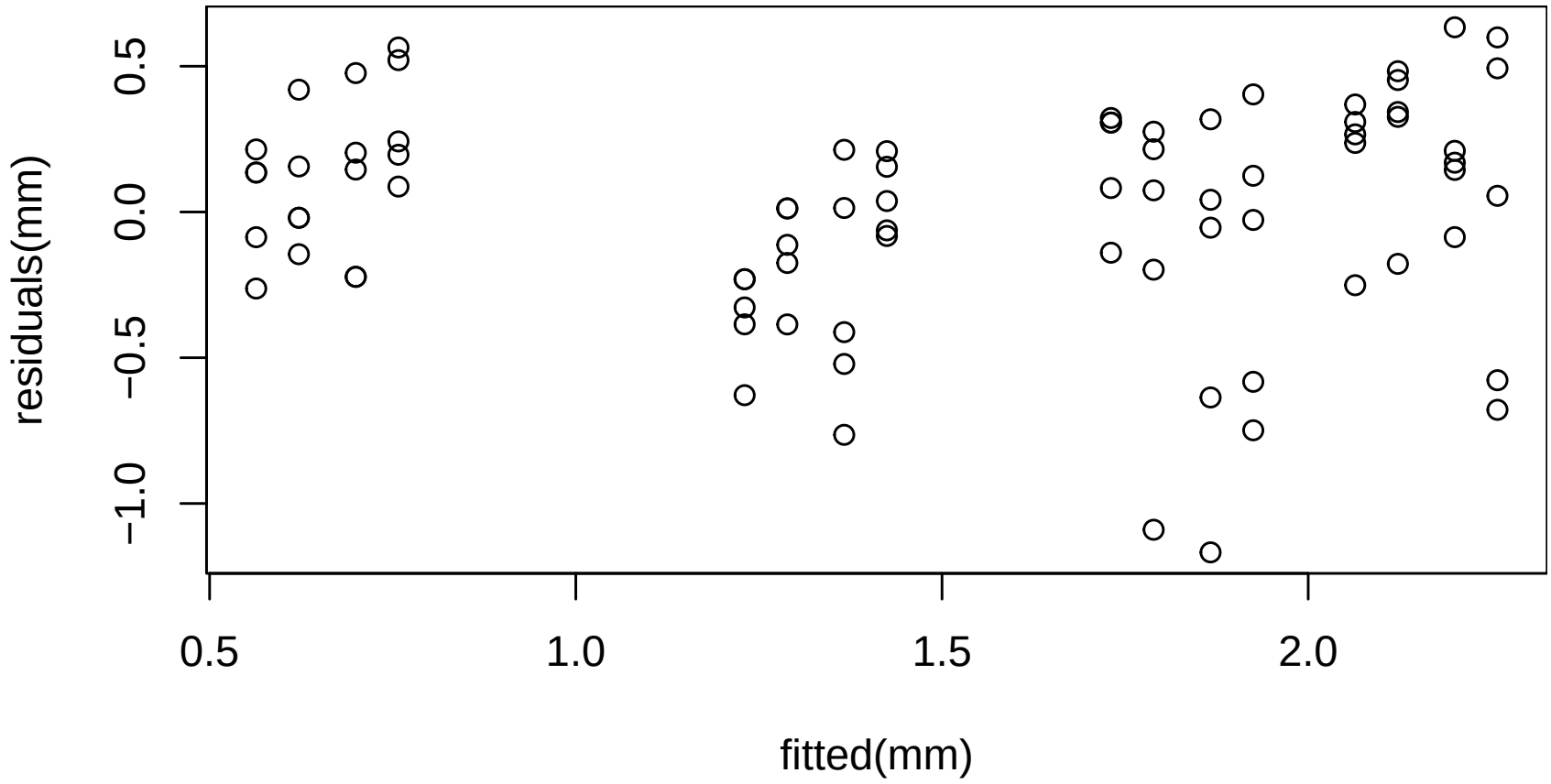
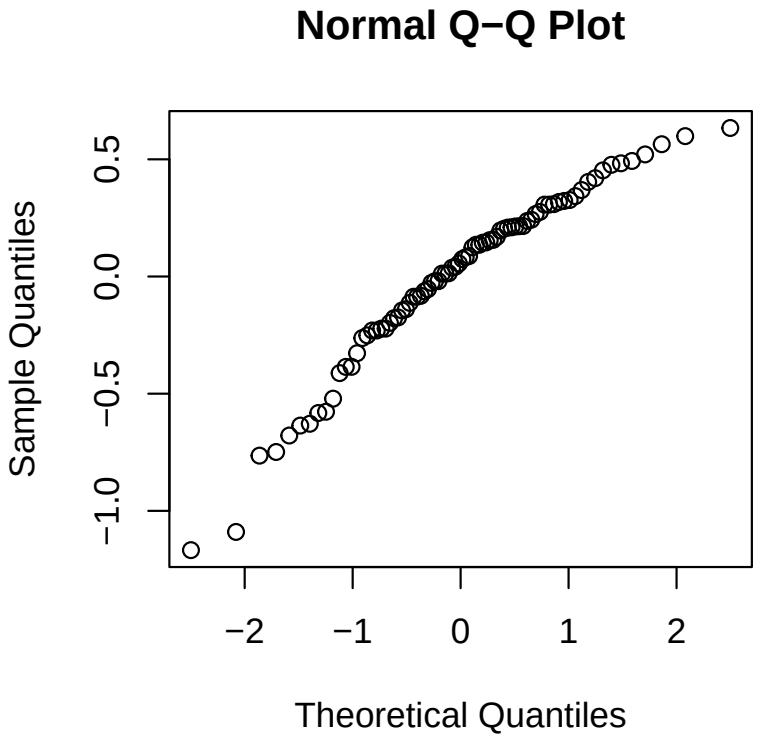


Diagramme quantile-quantile des résidus comme dans la régression linéaire simple

```
qqnorm(residuals(mm))
```

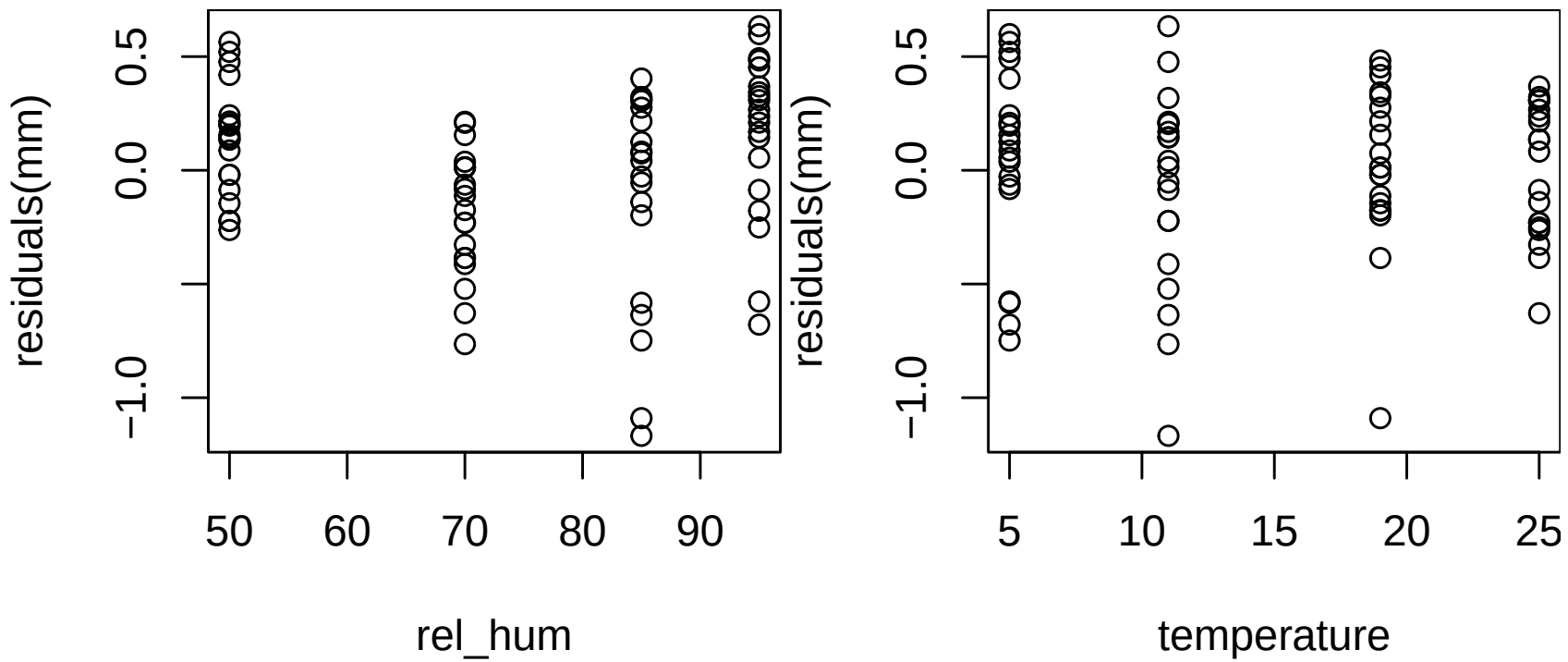


Ajout d'un graphe des résidus en fonction de chaque variable indépendante

Particulièrement utile pour détecter la violation de l'hypothèse de relation linéaire

(le cas de l'humidité relative dans cet exemple !)

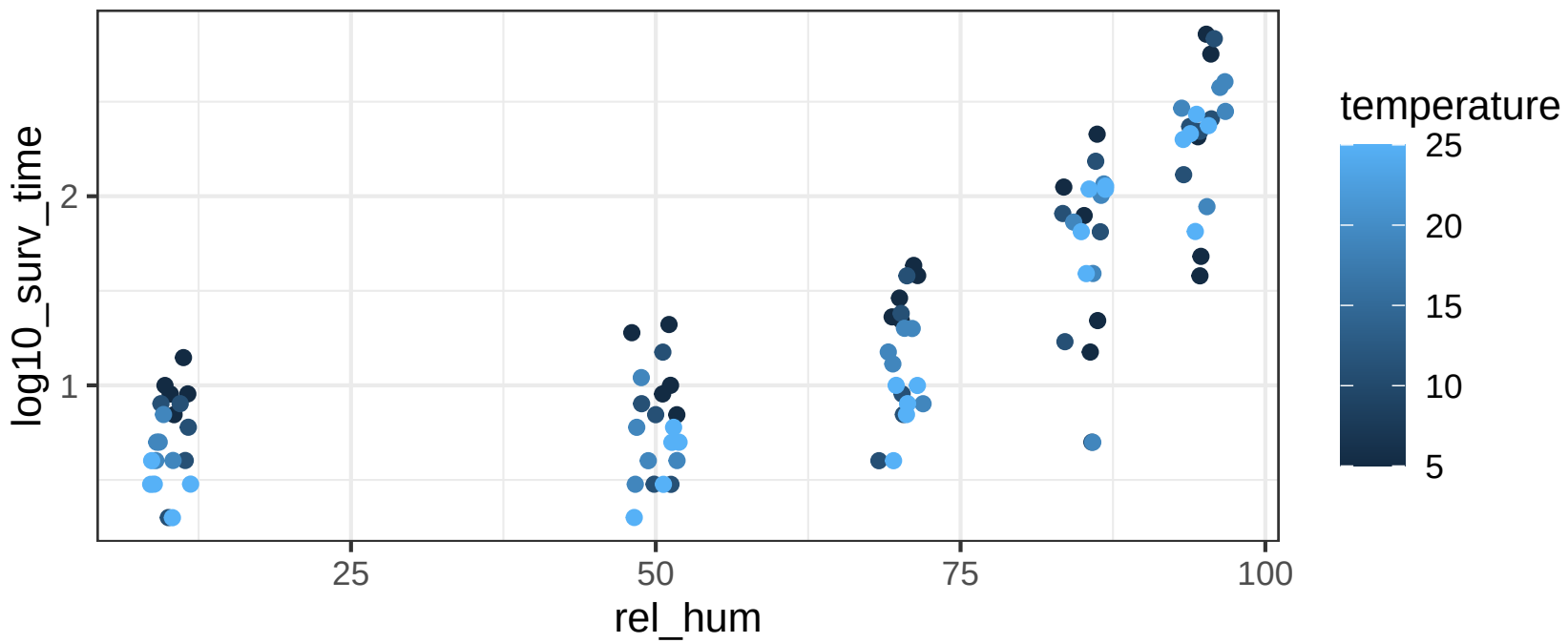
```
par(mar = c(4, 4, 0, 0), mfrow = c(1,2))
plot(residuals(mm) ~ rel_hum, data = dhum)
plot(residuals(mm) ~ temperature, data = dhum)
```



Retour sur les données

Nous aurions pu anticiper ce problème, et il aurait été pire si les données à 10% d'humidité avaient été conservées.

```
ggplot(data = dtot, aes(x = rel_hum,
  y = log10_surv_time, col = temperature)) +
  geom_jitter(width = 2)
```



Les modèles polynomiaux : une solution ?

Une façon simple de **prendre en compte la non-linéarité de la relation** entre la variable dépendante et un (ou plusieurs) régresseur(s) est d'utiliser des modèles polynomiaux.

```
(mm2 <- lm(log10_surv_time ~ rel_hum + I(rel_hum^2) +
                    temperature, data = dhum))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + I(rel_hum^2) + temperature,
##     data = dhum)
##
## Coefficients:
##  (Intercept)       rel_hum  I(rel_hum^2)   temperature
##     2.463772     -0.064501      0.000679     -0.009712
```

Ce modèle polynomial peut être ajusté par la méthode des moindres carrés : il s'agit d'un modèle linéaire avec une variable explicative supplémentaire (le carré de l'humidité relative).

A votre tour de manipuler la régression multiple

En utilisant le jeu de données `dhum` qui exclut la condition la plus arride :

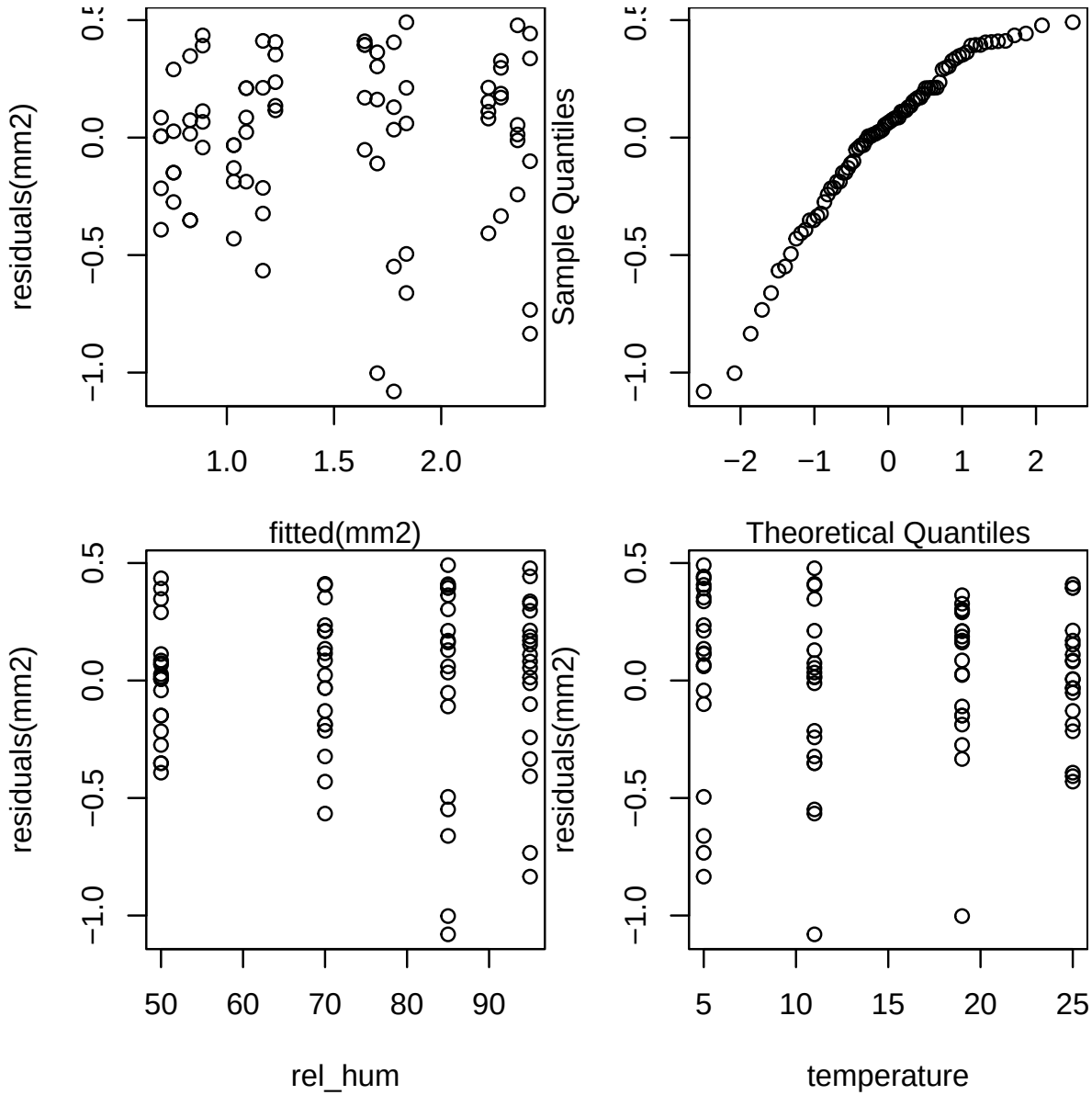
1. Examinez les **graphes des résidus** obtenus avec ce nouveau modèle. Le **problème de non-linéarité est-il résolu** ?
2. Regardez le résumé du nouveau modèle et essayez d'**interpréter les valeurs p**.

En utilisant l'ensemble des données `dtot` (avec toutes les conditions) :

3. Ajustez un modèle polynomial du second ordre **sans prendre en compte l'impact de la température**.
4. Regardez le résumé et les résidus.
5. Représentez le **modèle ajusté sur les données** et **interrogez-vous sur la pertinence biologique** de ce modèle. Pour cette question, vous devrez définir un nouveau jeu données avec des valeurs régulièrement espacées dans la gamme des conditions testées d'humidité relative et utiliser la fonction `predict()` directement sur ce nouveau jeu de données.

Réponse à la question 1

Il n'y a plus de tendance sur le graphe des résidus en fonction de l'humidité relative.



Réponse à la question 2

Amélioration significative de l'ajustement en ajoutant l'humidité relative au carré, mais pas d'amélioration significative en ajoutant la température.

```
summary(mm2)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + I(rel_hum^2) + temperature,
##     data = dhum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0802 -0.1873  0.0635  0.2185  0.4909
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.463772   0.904671    2.72  0.00801 **
## rel_hum      -0.064501   0.026104   -2.47  0.01572 *
## I(rel_hum^2)  0.000679   0.000180    3.76  0.00033 ***
## temperature  -0.009712   0.005176   -1.88  0.06444 .
## ---
```

Réponse à la question 3

```
(m2tot <- lm(log10_surv_time ~ rel_hum + I(rel_hum^2), data = dtot))
```

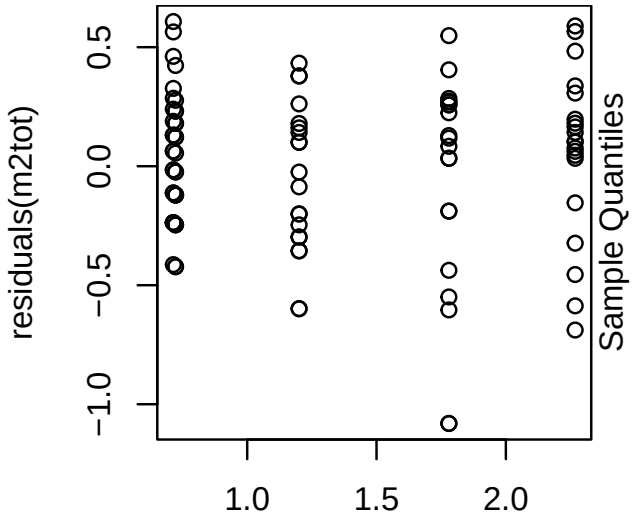
```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + I(rel_hum^2), data = dtot)
##
## Coefficients:
##  (Intercept)       rel_hum  I(rel_hum^2)
##     0.929364     -0.024727      0.000409
```

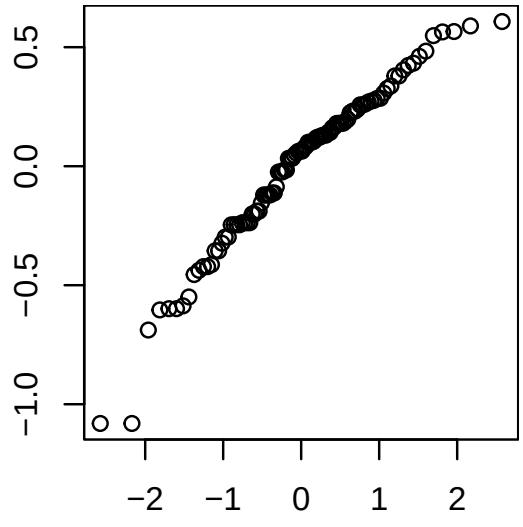
Réponse à la question 4 - résumé

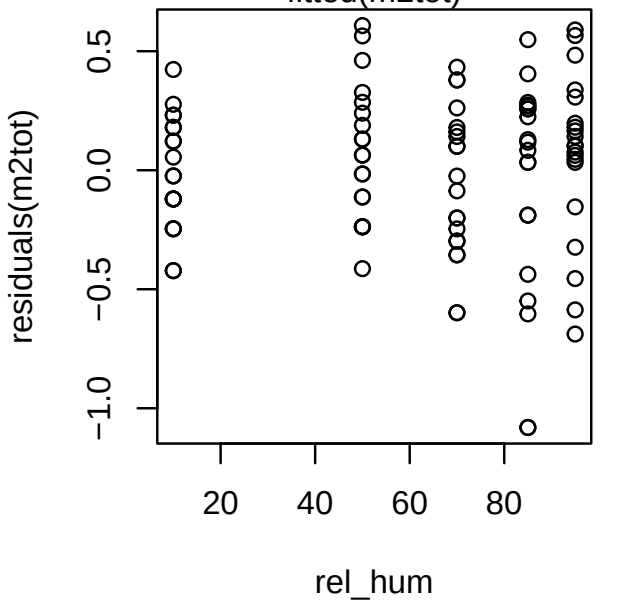
```
summary(m2tot)
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + I(rel_hum^2), data = dtot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0808 -0.2374  0.0636  0.2313  0.6077
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.29e-01   1.09e-01    8.55  1.8e-13 ***
## rel_hum      -2.47e-02   4.87e-03   -5.08  1.8e-06 ***
## I(rel_hum^2)  4.09e-04   4.58e-05    8.92  2.9e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.34 on 97 degrees of freedom
## Multiple R-squared:  0.767,  Adjusted R-squared:  0.762
## F-statistic:  159 on 2 and 97 DF,  p-value: <2e-16
```

Réponse à la question 4 - résidus





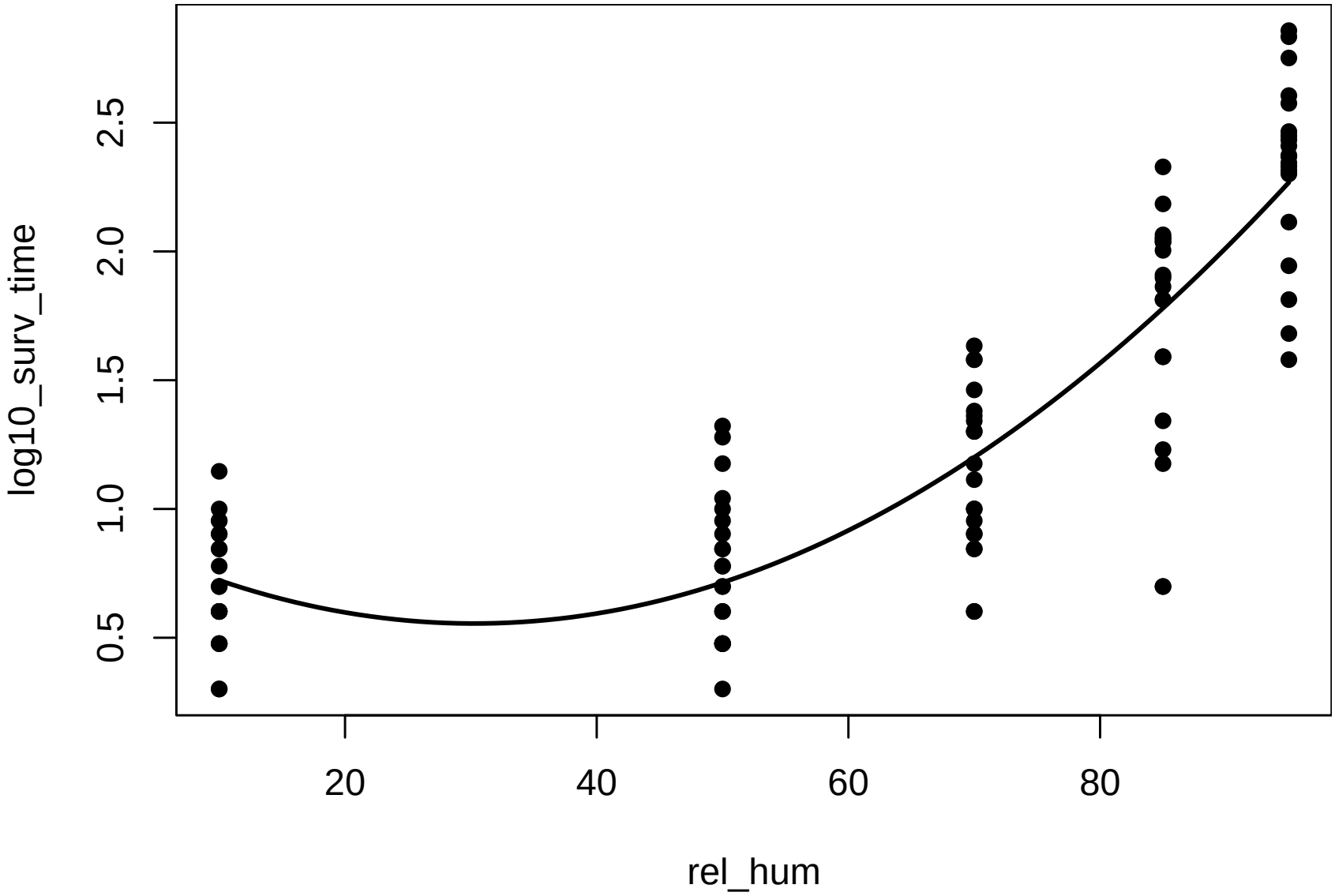


Réponse à la question 5 - code R

```r
# Tracés des points observés
plot(log10_surv_time ~ rel_hum,
     data = dtot, pch = 16)
# Création d'un jeu de données avec 50 valeurs
# de rel_hum  entre 10 et 95
data4pred <- data.frame(rel_hum =
                  seq(10, 95, length.out = 50))
# Utilisation du modèle en prédiction sur
# ce jeu de données
pred <- predict(mm2tot, newdata = data4pred)
# Tracé des points prédits en les reliant par
# des segments de droite
lines(pred ~ data4pred$rel_hum, lwd = 2)
```

Réponse à la question 5 - le graphe

Un minimum de temps de survie entre 20% et 40% d'humidité est-il réellement attendu d'un point de vue biologique ?



Prise en compte de variable(s) explicative(s) qualitative(s)

Le modèle d'ANOVA1

Certaines **variables explicatives peuvent être qualitatives** (par exemple le sexe) ou peuvent être **transformées en une variable qualitative** pour faire face à la violation de la condition de linéarité. Les variables explicatives qualitatives sont appelées aussi des **facteurs**.

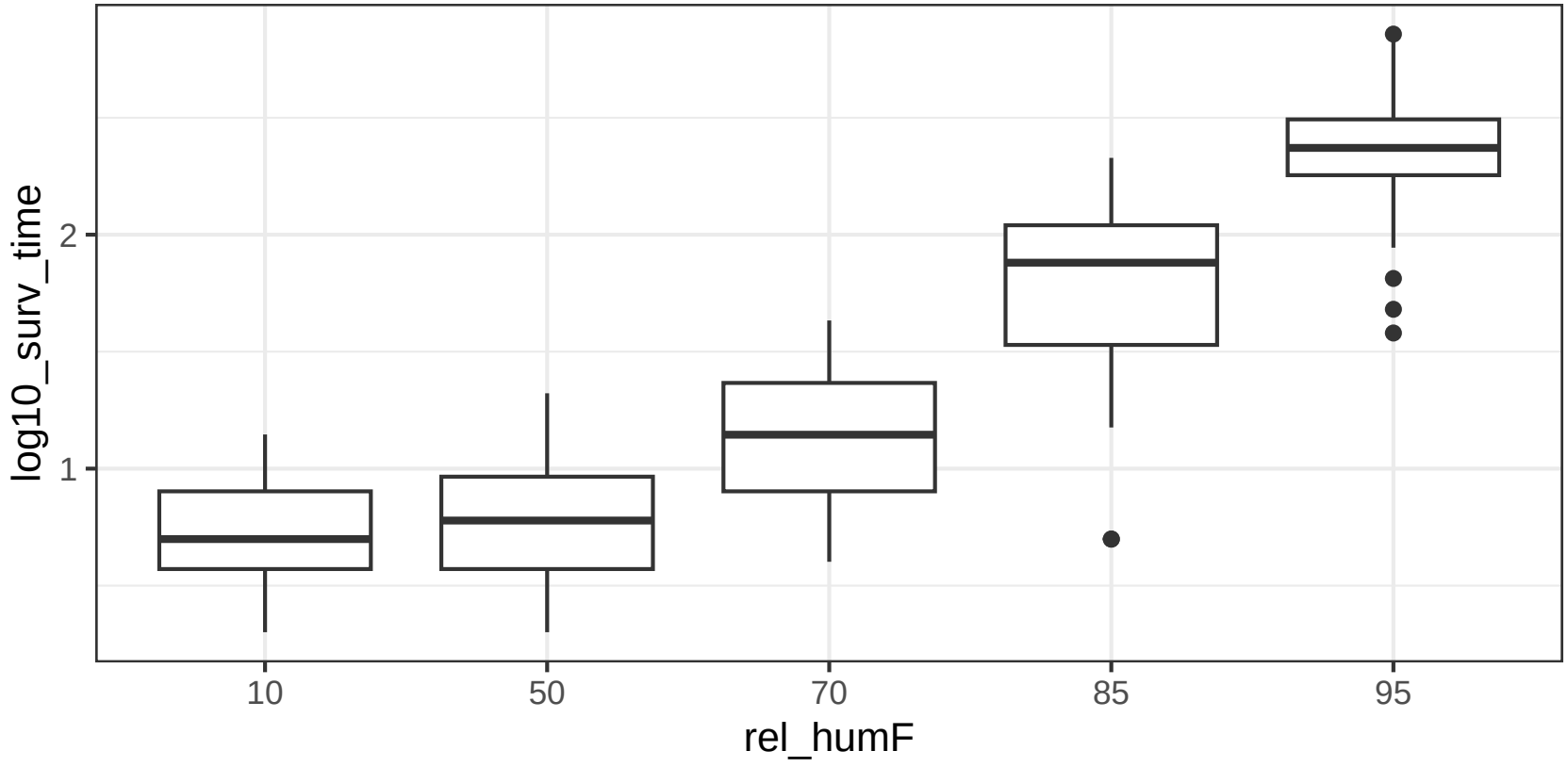
Imaginons que nous voulions modéliser l'**impact de l'humidité relative** sur le temps de survie (en log_{10}) en la considérant **comme un facteur à 5 modalités (= conditions testées)**, en négligeant l'impact potentiel de la température.

```
# Definition of the qualitative variable
dtot$rel_humF <- as.factor(dtot$rel_hum)
levels(dtot$rel_humF)
```

```
## [1] "10" "50" "70" "85" "95"
```

Représentation classique des données en diagrammes en boîte

```
ggplot(data = dtot, aes(x = rel_humF,
                        y = log10_surv_time)) + geom_boxplot()
```



Formalisation du modèle d'ANOVA1

- **Formalisation classique**: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ avec $\epsilon_{ij} \sim N(0, \sigma)$ et $\sum \alpha_i = 0$
- Formalisation utilisant $p-1$ **variables muettes** X_1 à X_{p-1} codant pour **l'appartenance de chaque observation aux** p **groupes à l'exception du groupe de référence (ici groupe p)** : $Y_k = \beta_0 + \beta_1 X_{1,k} + \cdots + \beta_{p-1} X_{p-1,k} + \epsilon_k$ avec $\epsilon_k \sim N(0, \sigma)$
- lien entre les deux formalisations :
  - moyenne du groupe 1 $= \mu + \alpha_1 = \beta_0 + \beta_1$
  - moyenne du groupe 2 $= \mu + \alpha_2 = \beta_0 + \beta_2$
  - moyenne du groupe i $= \mu + \alpha_i = \beta_0 + \beta_i$
  - moyenne du groupe p $= \mu + \alpha_p = \beta_0$

Ainsi, chaque coefficient du modèle linéaire correspond à la **différence entre la moyenne de la classe correspondante et la moyenne de la classe de référence**.

Ajustement du modèle linéaire

```
(manova1 <- lm(log10_surv_time ~ rel_humF, data = dtot))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_humF, data = dtot)
##
## Coefficients:
## (Intercept)   rel_humF50   rel_humF70   rel_humF85   rel_humF95
##      0.7084       0.0783       0.4359       1.0107       1.6188
```

On voit ici, d'après l'appellation des coefficients, que le groupe de référence choisi est le groupe à 10% d'humidité relative (il prend la première modalité du facteur comme modalité de référence).

```
levels(dtot$rel_humF)
```

```
## [1] "10" "50" "70" "85" "95"
```

Tableau d'ANOVA1

```
anova(manova1)
```

```
## Analysis of Variance Table
##
## Response: log10_surv_time
##            Df Sum Sq Mean Sq F value Pr(>F)
## rel_humF    4   37.2    9.31      81 <2e-16 ***
## Residuals 95   10.9    0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 
```

Comme attendu dans cet exemple, il montre un effet significatif de l'humidité relative sur le temps de survie (en log_{10})

Alternative pour tester l'effet du facteur

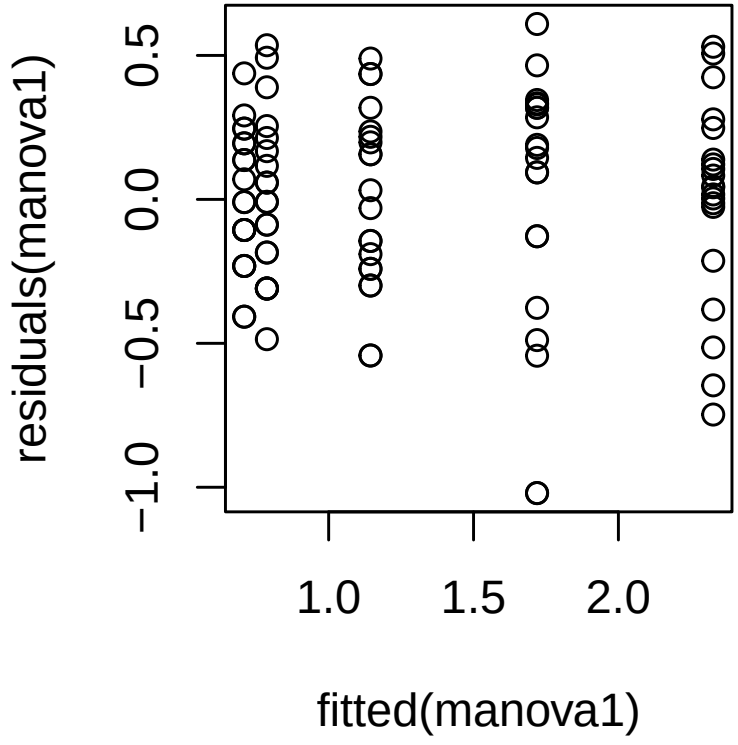
Alternative qui fonctionne même avec les modèles plus complexes pour tester l'impact du facteur (alors que l'utilisation de la fonction `anova()`sur un modèle plus complexe est très délicate)

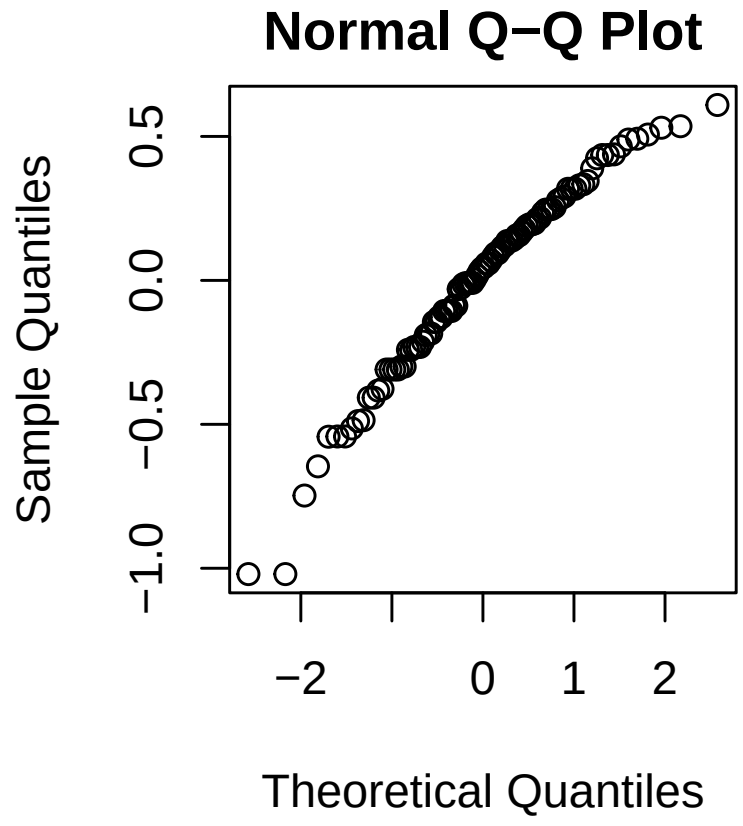
```
drop1(manova1, test = "F")
```

```
## Single term deletions
##
## Model:
## log10_surv_time ~ rel_humF
##           Df Sum of Sq  RSS    AIC F value Pr(>F)
## <none>                 10.9 -211.4
## rel_humF  4      37.2 48.2  -71.1      81 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Vérification des conditions d'utilisation

```
par(mar = c(4, 4, 2, 2), mfrow = c(1,2))
plot(residuals(manova1) ~ fitted(manova1))
qqnorm(residuals(manova1))
```





Interprétation des coefficients

```
# Observed means
tapply(dtot$log10_surv_time, dtot$rel_humF, mean)
```

```
##    10    50    70    85    95
## 0.708 0.787 1.144 1.719 2.327
```

```
# estimated coefficients
coef(manova1)
```

```
## (Intercept)  rel_humF50  rel_humF70  rel_humF85  rel_humF95
##      0.7084      0.0783      0.4359      1.0107      1.6188
```

Chacun correspond à la **différence entre la moyenne de la classe correspondante et la moyenne de la classe de référence** (premier niveau du facteur, par défaut avec classement alphabétique des niveaux).

Coefficients et intervalles de confiance à 95%

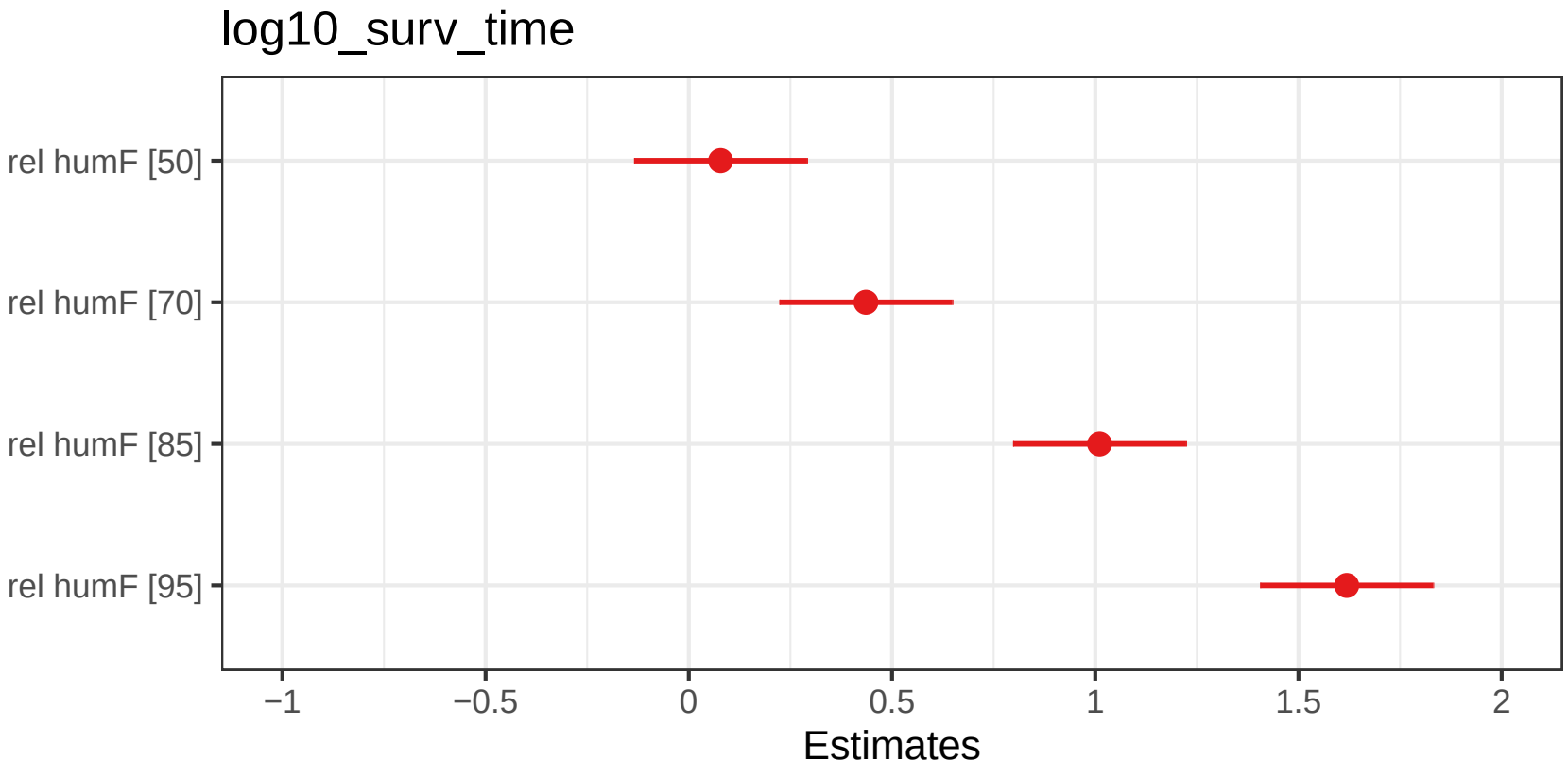
(ou estimations ponctuelles et par intervalle)

```
cbind(estimate = coef(manova1), confint(manova1))
```

```
##                estimate  2.5 % 97.5 %
## (Intercept)      0.7084  0.558  0.859
## rel_humF50       0.0783 -0.135  0.291
## rel_humF70       0.4359  0.223  0.649
## rel_humF85       1.0107  0.798  1.224
## rel_humF95       1.6188  1.406  1.832
```

“Forest plot” pour visualiser les estimations ponctuelles et par intervalle

```
library(sjPlot)
plot_model(manova1, type = "est")
```



Un modèle à deux facteurs (fixes et croisés)

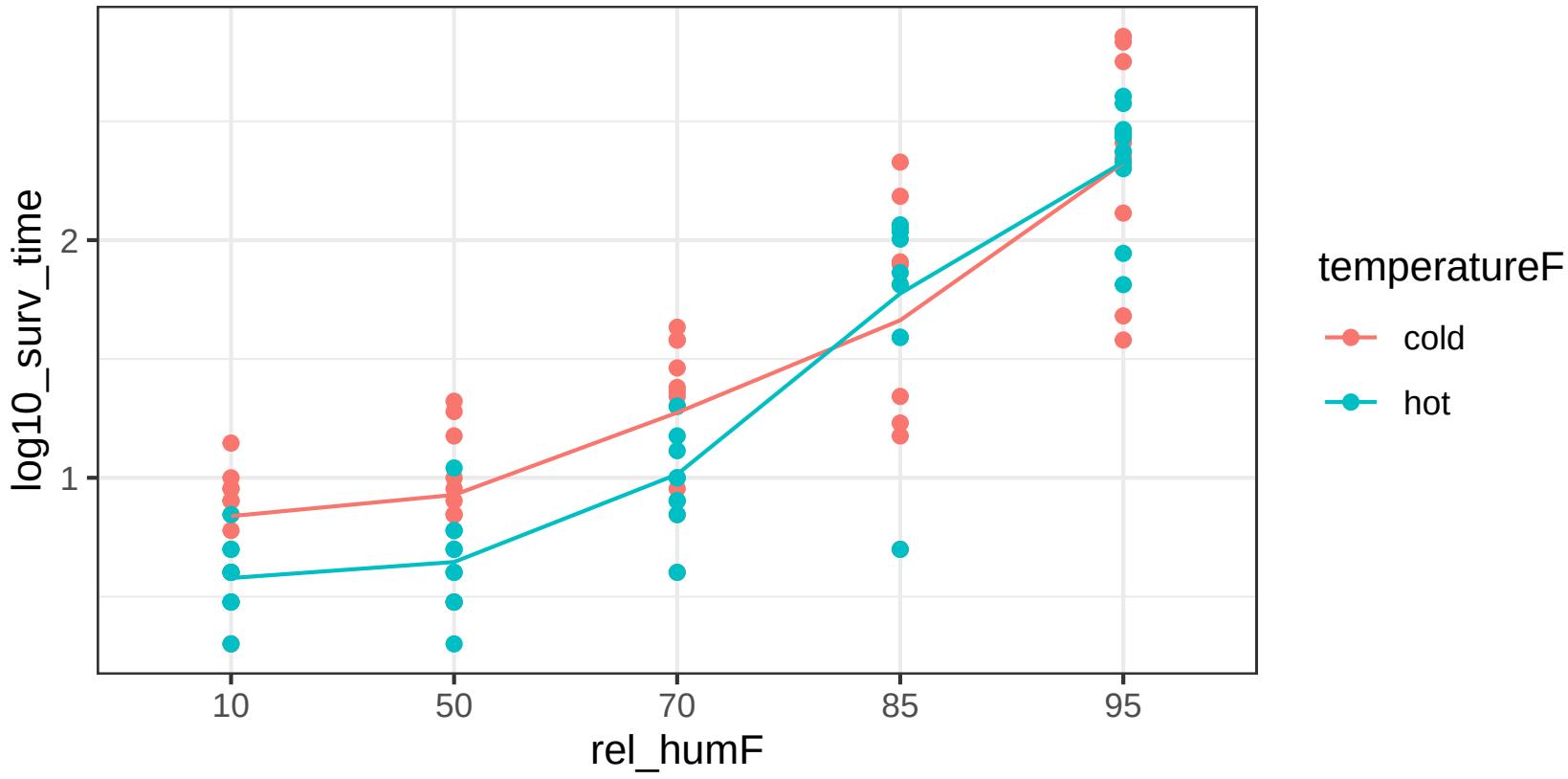
Imaginons maintenant que nous voulions **ajouter** dans ce modèle l'impact de la **température**, également **transformée en une variable qualitative** à deux modalités, froid ($< 15°C$) ou chaud.

```
dtot$temperatureF <- as.factor(ifelse(dtot$temperature < 15,
                                      "cold", "hot"))
# Look at the experimental design
xtabs(data = dtot, ~ rel_humF + temperatureF)
```

```
##          temperatureF
## rel_humF cold hot
##       10   10  10
##       50   10  10
##       70   10  10
##       85   10  10
##       95   10  10
```

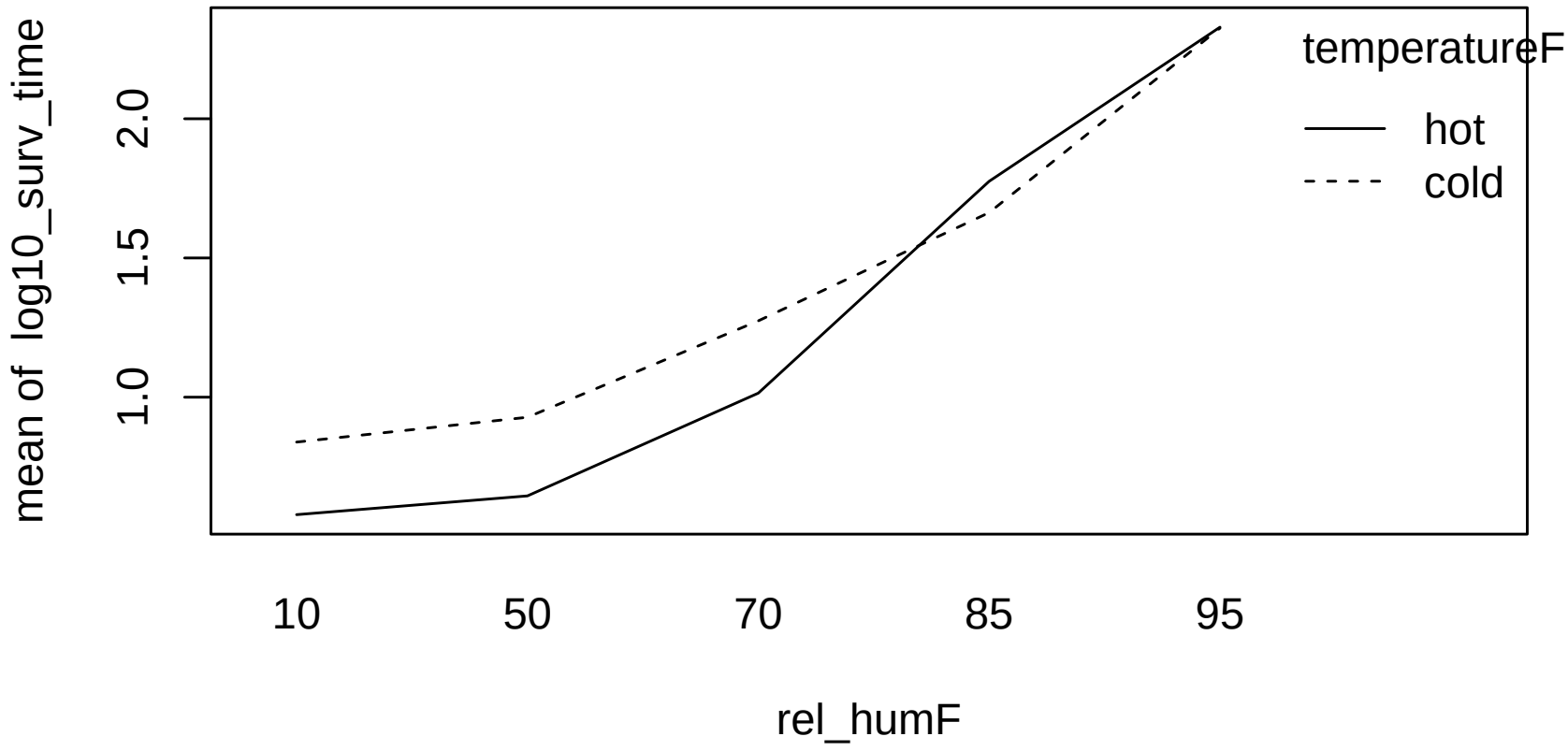
Un graphe d'interaction utilisant ggplot2

```
ggplot(data = dtot, aes(x = rel_humF, y = log10_surv_time,
        col = temperatureF)) + geom_point() +
 stat_summary(fun = mean, geom = "line", aes(group = temperatureF))
```



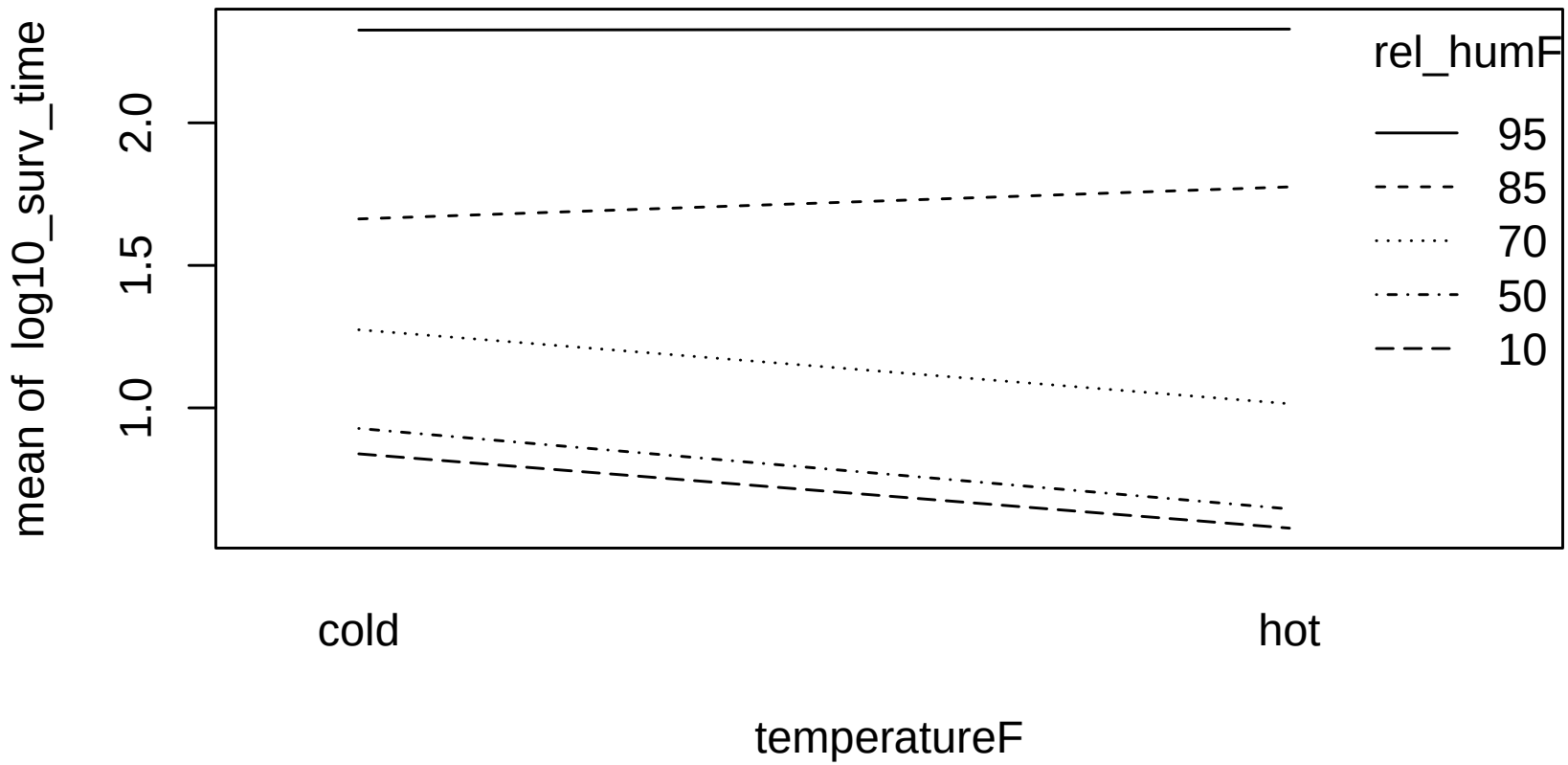
Un graphe d'interaction utilisant graphics

```
par(mar = c(4, 4, 1, 1))
within(dtot, interaction.plot(rel_humF, temperatureF,
                              log10_surv_time))
```



Une seconde version du graphe d'interaction utilisant graphics

```
par(mar = c(4, 4, 1, 1))
within(dtot, interaction.plot(temperatureF, rel_humF,
                              log10_surv_time))
```



Modèles sans ou avec interaction entre les facteurs A et B

- Deux facteurs A et B peuvent chacun avoir un effet sur la variable dépendante **sans interaction**. Les **effets** sont alors dits **additifs**.

L'effet de *A* ne dépend pas de la modalité de *B*, et *vice versa*.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim N(0, \sigma)$$

- Deux facteurs peuvent interagir. Le **modèle n'est plus additif**. Il incorpore des termes d'interaction γ_{ij}.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim N(0, \sigma)$$

Le choix entre les deux modèles doit être guidé par les connaissances biologiques *a priori* et l'observation des données.

Ajustement du modèle sans interaction

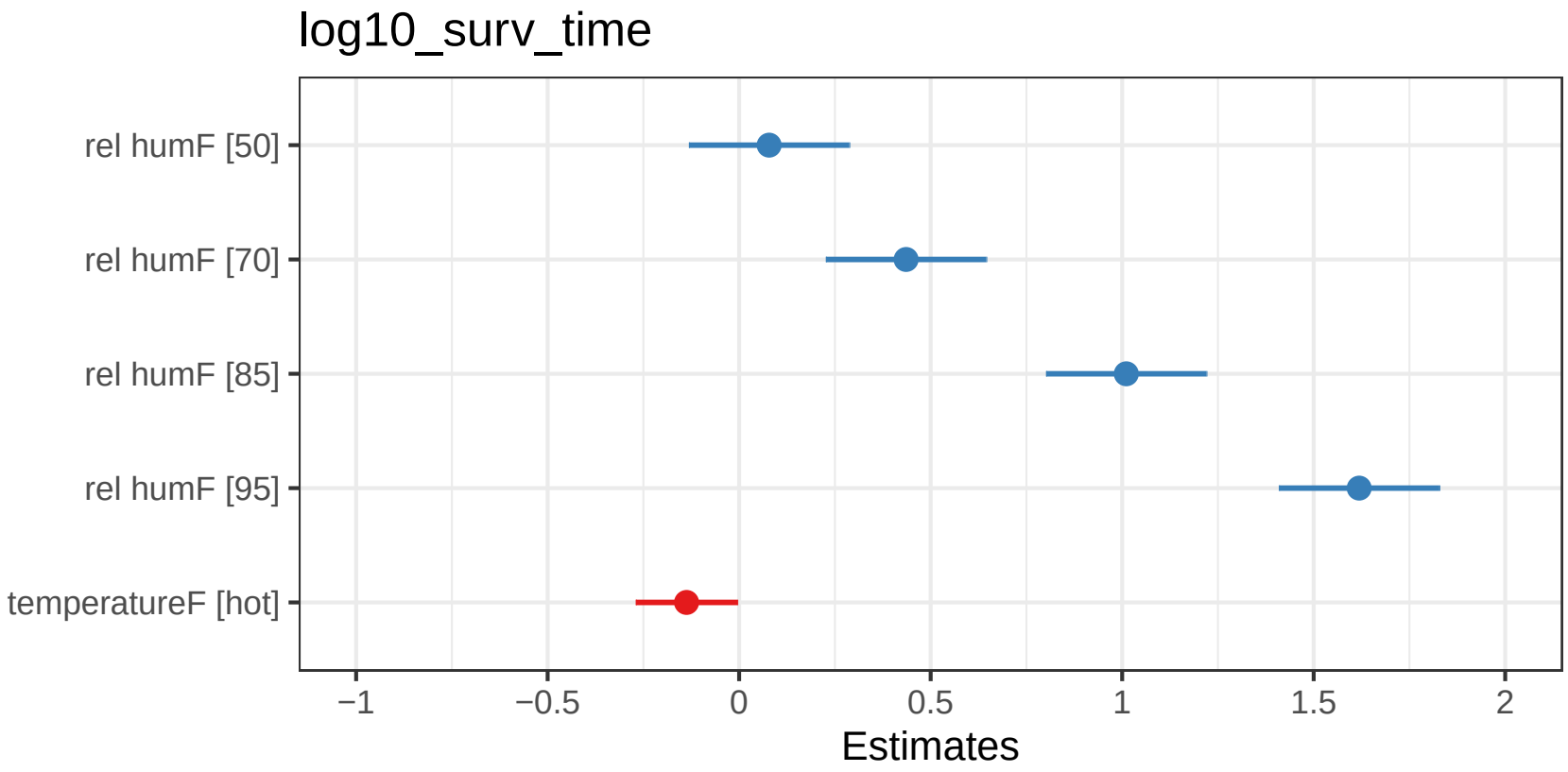
```
(manova2 <- lm(log10_surv_time ~ rel_humF + temperatureF,
               data = dtot))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_humF + temperatureF, data = dtot)
##
## Coefficients:
##     (Intercept)       rel_humF50       rel_humF70       rel_humF85
##          0.7771           0.0783           0.4359           1.0107
##      rel_humF95  temperatureFhot
##          1.6188          -0.1373
```

Ex. de la moyenne prédite à 70% d'humidité et à température élevée : 0.777 + 0.436 - 0.137

Interprétation assez simple des coefficients

```
plot_model(manova2)
```



Ajustement du modèle avec interaction

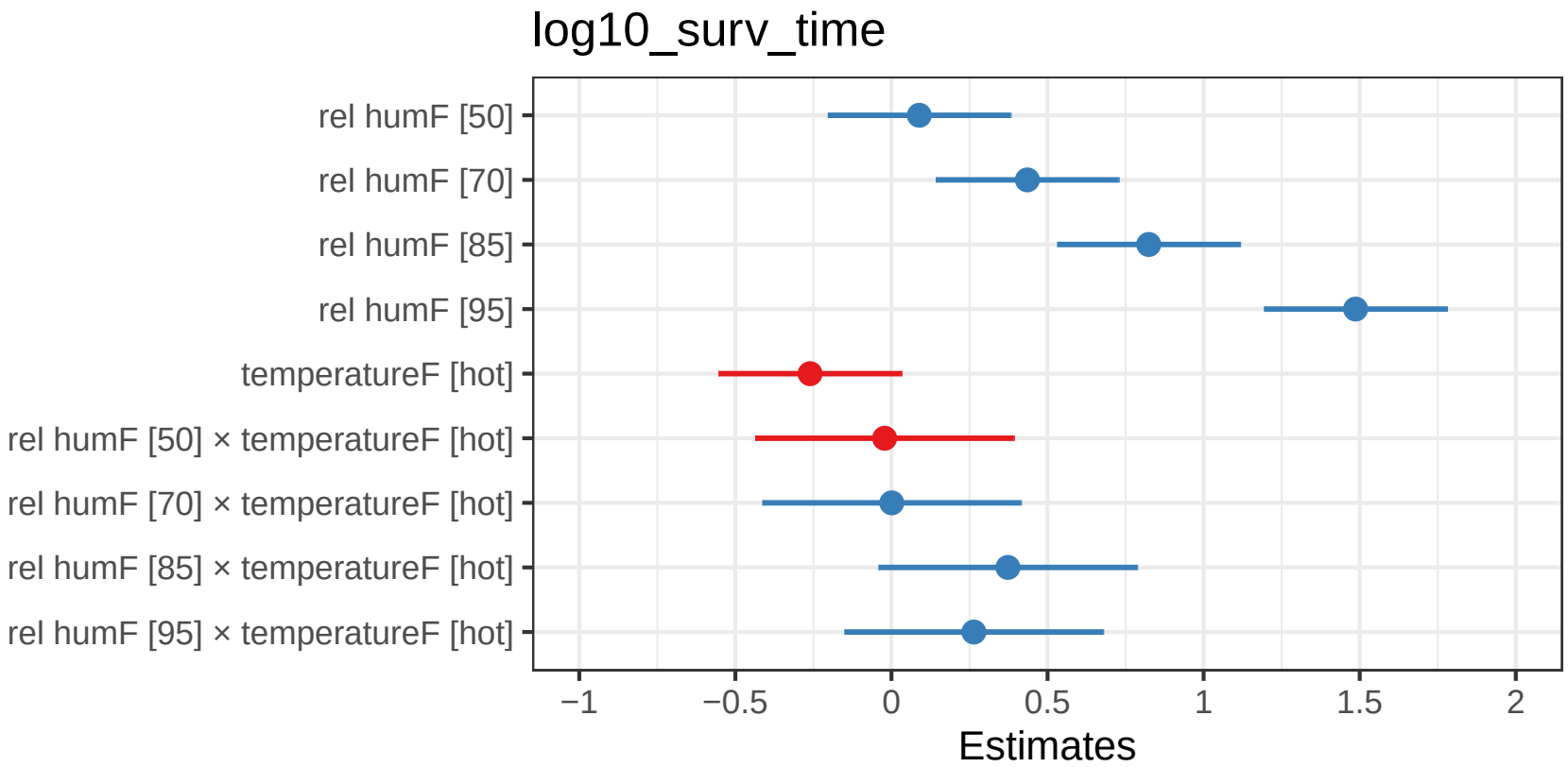
```
(manova2int <- lm(log10_surv_time ~ rel_humF + temperatureF +
                    rel_humF:temperatureF, data = dtot))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_humF + temperatureF + rel_humF:temperatur
##     data = dtot)
##
## Coefficients:
##                 (Intercept)                     rel_humF50
##                    0.838713                       0.089170
##                  rel_humF70                     rel_humF85
##                    0.435406                       0.824212
##                  rel_humF95              temperatureFhot
##                    1.486764                      -0.260552
## rel_humF50:temperatureFhot  rel_humF70:temperatureFhot
##                   -0.021829                      0.000976
## rel_humF85:temperatureFhot  rel_humF95:temperatureFhot
##                    0.373032                      0.264005
```

Ex. de la moyenne prédite à 70% d'humidité et à température élevée : 0.839 + 0.435 - 0.261 + 0.001

Interprétation plus complexe des coefficients

```
plot_model(manova2int)
```



Comparaison des modèles avec et sans interaction

Ces deux modèles emboîtés (l'un étant une simplification de l'autre) peuvent être comparés à l'aide d'un test F.

```
anova(manova2int, manova2)
```

```
## Analysis of Variance Table
##
## Model 1: log10_surv_time ~ rel_humF + temperatureF + rel_humF:temperatureF
## Model 2: log10_surv_time ~ rel_humF + temperatureF
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     90  9.78
## 2     94 10.45 -4    -0.667 1.53    0.2
```

Dans cet exemple, l'interaction n'est pas significative (mais ce n'est pas une raison qui à elle seule justifie le choix du modèle le plus simple).

Prise en compte de variable(s) explicative(s) qualitative(s) et quantitatives(s)

Un modèle linéaire avec des variables explicatives qualitatives et quantitatives

Imaginons que nous voulions modéliser l'impact sur le temps de survie (en log_{10})

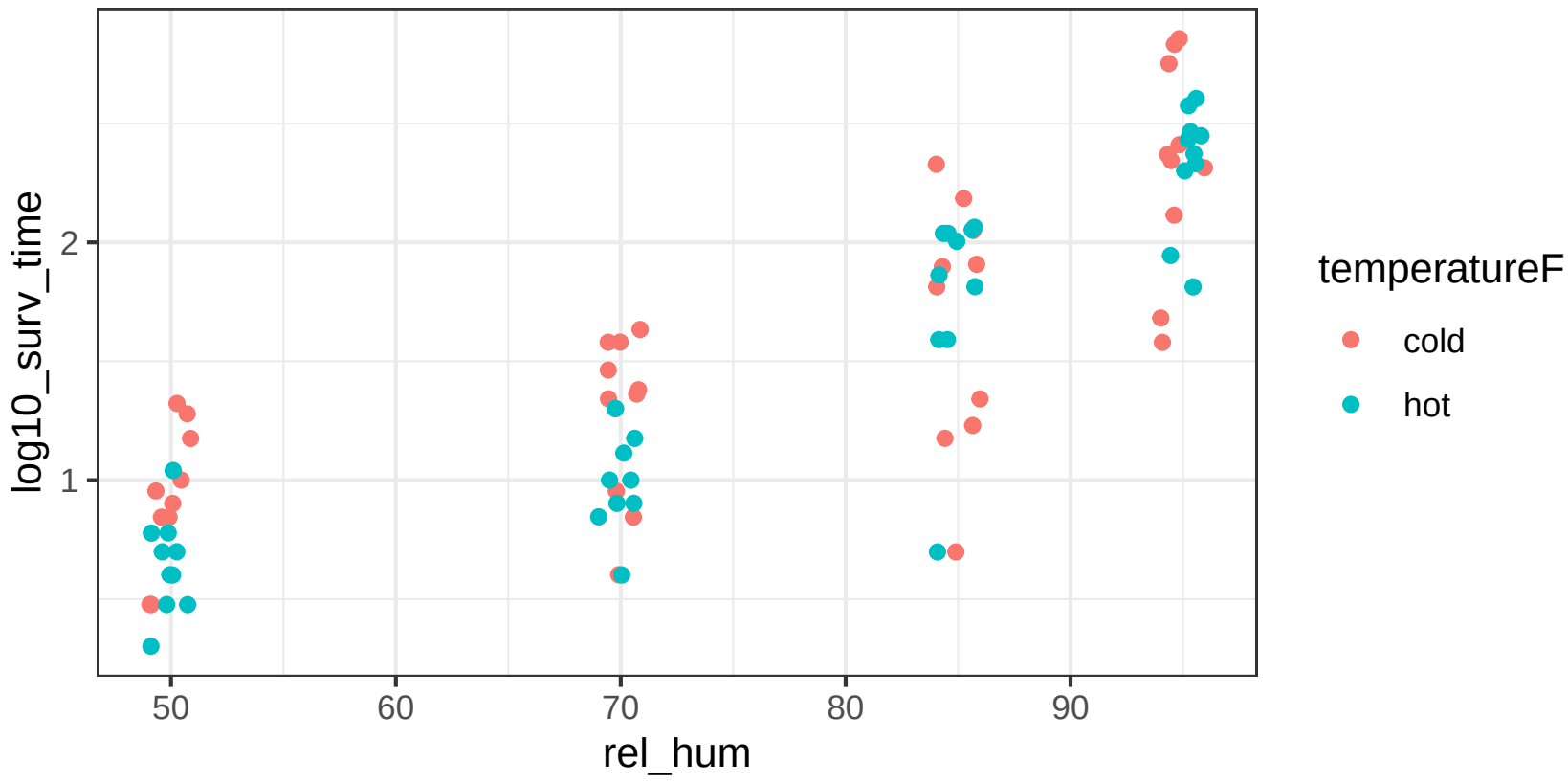
- de l'**humidité relative considérée comme une variable quantitative**
- et de la température **comme variable qualitative à 2 modalités**, froid ($< 15°C$) ou chaud,

en excluant les données de la condition la plus arride.

```
dhum$temperatureF <- as.factor(ifelse(dhum$temperature < 15,
                                      "cold", "hot"))
```

Représentation graphique des données

```
ggplot(data = dhum, aes(x = rel_hum, y = log10_surv_time,
       col = temperatureF)) + geom_jitter(width = 1)
```



Modèles avec ou sans interaction entre le facteur A et la covariable X

- Le modèle sans interaction

La pente β ne dépend pas de la modalité de A.

$$Y_{ij} = \mu + \alpha_i + \beta \times X_{ij} + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim N(0, \sigma)$$

- Le modèle avec interaction

Les pentes β_i sont différentes

$$Y_{ij} = \mu + \alpha_i + \beta_i \times X_{ij} + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim N(0, \sigma)$$

Le choix entre les deux modèles doit être guidé par les connaissances biologiques *a priori* et l'observation des données.

Ajustement du modèle sans interaction

```
(mancova <- lm(log10_surv_time ~ rel_hum + temperatureF,
                  data = dhum))
```

```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperatureF, data = dhum)
##
## Coefficients:
##     (Intercept)          rel_hum  temperatureFhot
##         -0.9533           0.0333          -0.1065
```

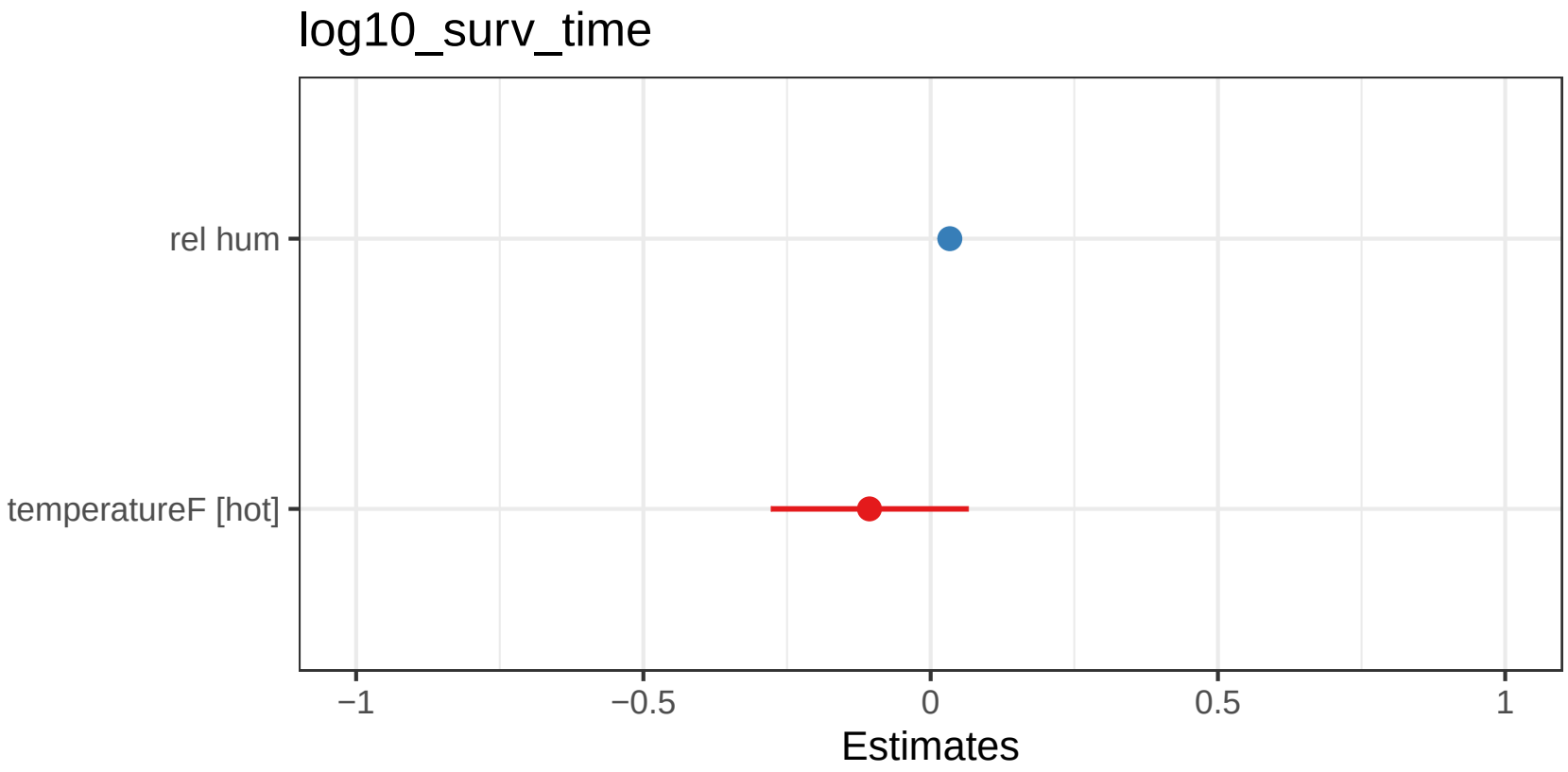
Notez que l'ordonnée à l'origine n'a pas de sens biologique dans ce cas (0 n'est pas dans l'intervalle des valeurs d'humidité relative observées).

Ex. de la moyenne prédite à 65% d'humidité et température chaude :

$-0.9533 + 0.0333 \times 65 - 0.1065.$

Interprétation difficile des coefficients pris tels quels

```
plot_model(mancova)
```



Normalisation des coefficients de régression pour faciliter leur interprétation

Les **coefficients de régression β (ou β_i) dépendent de l'ordre de grandeur de la covariable** X. Pour les **rendre comparables** entre eux et comparables aux coefficients correspondant aux facteurs, il est recommandé de les diviser par $2 \times SD_X$ avec SD_X l'écart-type des valeurs de X.

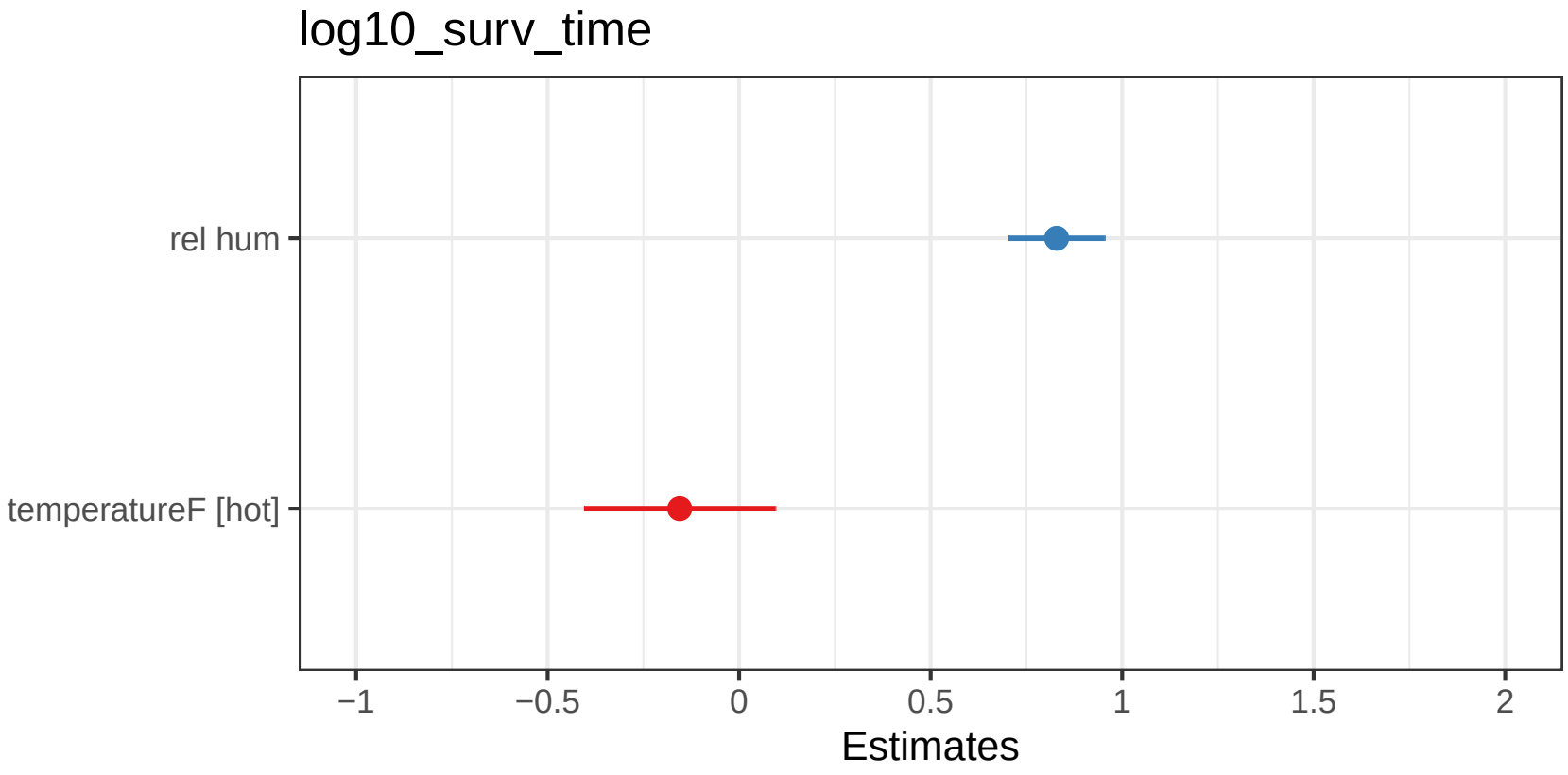
cf. Gelman A (2008) "Scaling regression inputs by dividing by two standard deviations." Statistics in Medicine 27: 2865-2873. pour une justification théorique.

Cela peut être facilement fait en utilisant l'argument `type` de la fonction `plot_model()` comme ci-dessous.

```
plot_model(mancova, type = "std2")
```

Interprétation des coefficients standardisés du modèle sans interaction

```
plot_model(mancova, type = "std2")
```



Ajustement du modèle avec interaction

```
(mancovaint <- lm(log10_surv_time ~ rel_hum + temperatureF +
                   rel_hum:temperatureF, data = dhum))
```

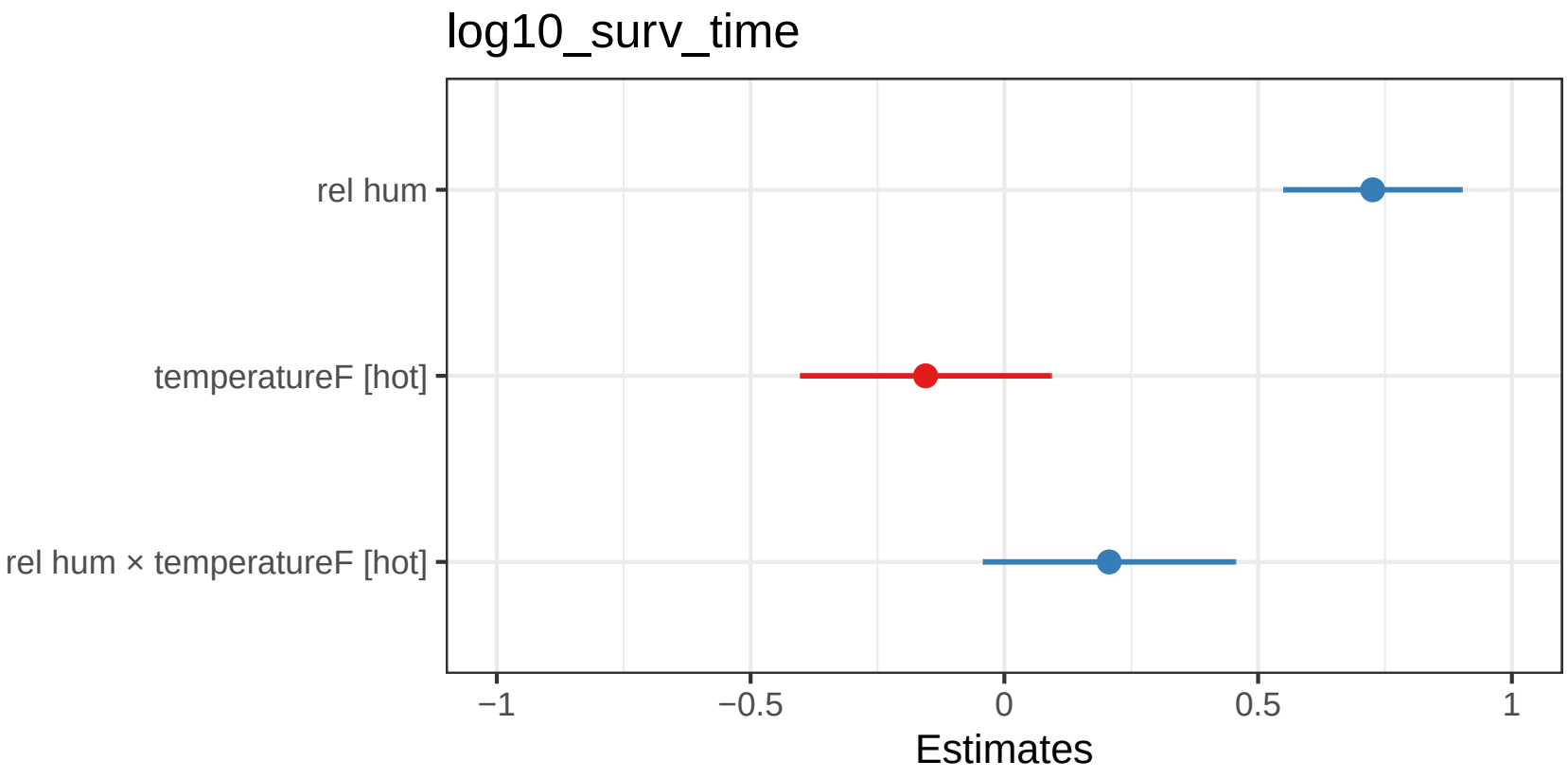
```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperatureF + rel_hum:temperatureF
##     data = dhum)
##
## Coefficients:
##              (Intercept)                  rel_hum          temperatureFhot
##                  -0.64182                  0.02919                  -0.72941
## rel_hum:temperatureFhot
##                  0.00831
```

Ex. de la moyenne prédite à 65% d'humidité et à température basse :
$-0.6418 + 0.0292 \times 65$.

Et pour la moyenne prédite à 65% d'humidité à température élevée ?

Interprétation des coefficients standardisés du modèle avec interaction

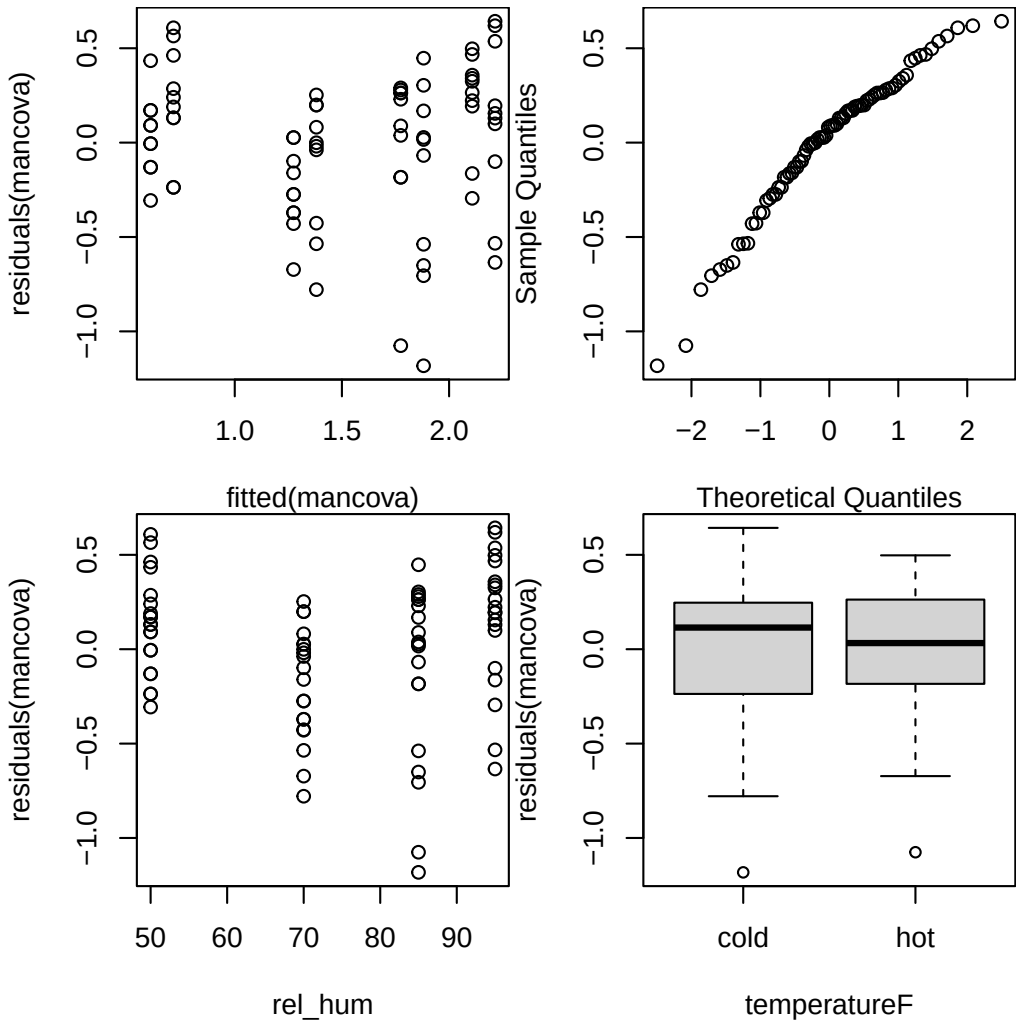
```
plot_model(mancovaint, type = "std2")
```



A votre tour de manipuler variables qualitatives et quantitatives

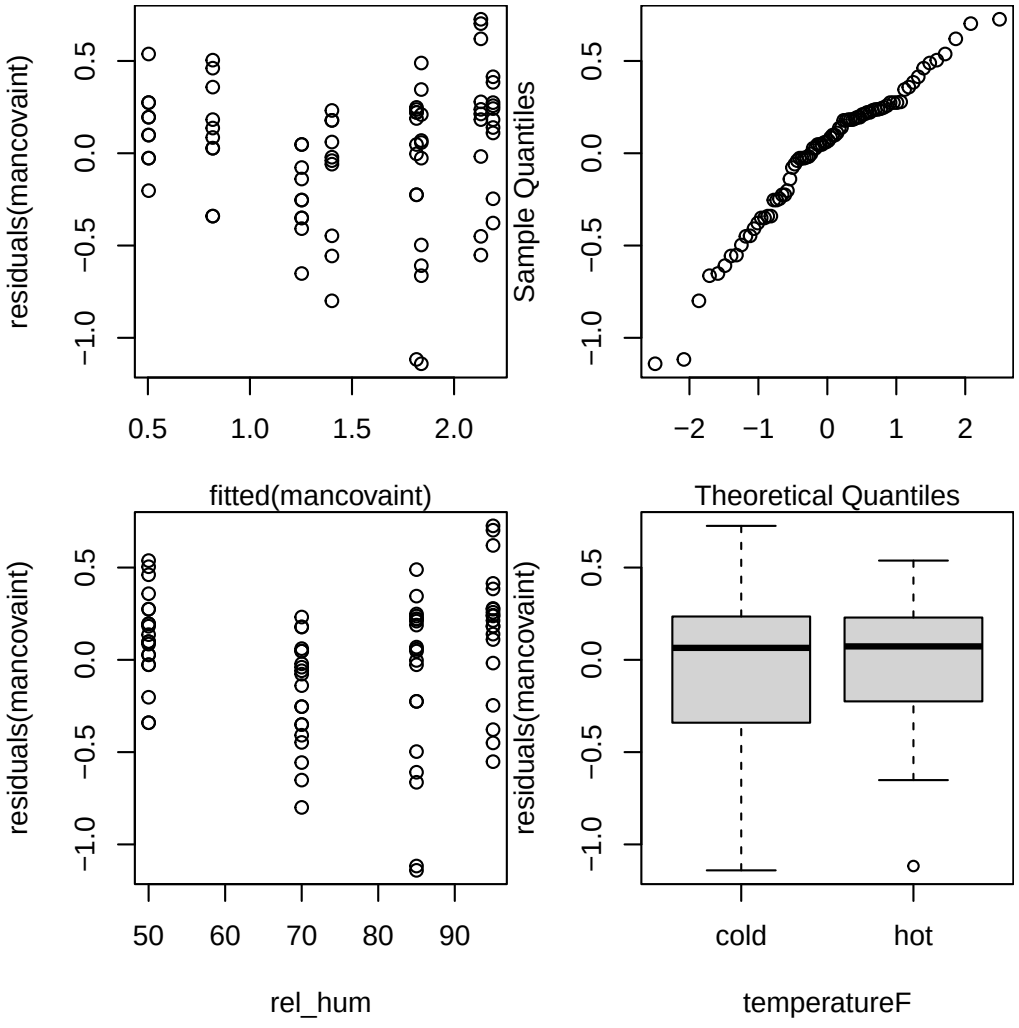
1. Examinez les résidus pour chacun des deux modèle précédents (avec et sans interaction).
2. Prédire le temps de survie d'une tique exposée à une humidité de 65% et à une température comprise entre $15°C$ et $25°C$ avec son intervalle de confiance à 95%, en utilisant la fonction `predict()` avec chacun des deux modèles.
3. Tracer chaque modèle sur les données à l'aide de la fonction `abline()`.

Réponse à la question 1 - modèle sans interaction



La principale tendance observée est celle due à la non-linéarité de la relation entre le résultat et l'humidité relative.

Réponse à la question 1 - modèle avec interaction



Comme précédemment, la principale tendance observée est celle due à la non-linéarité de la relation entre le résultat et l'humidité relative.

Réponse à la question 2

```r
# Definition of the new data frame for prediction
data4pred <- data.frame(rel_hum = 65, temperatureF = "hot")

# Prediction using the model without interaction
predict(mancova, newdata = data4pred, interval = "prediction")
```

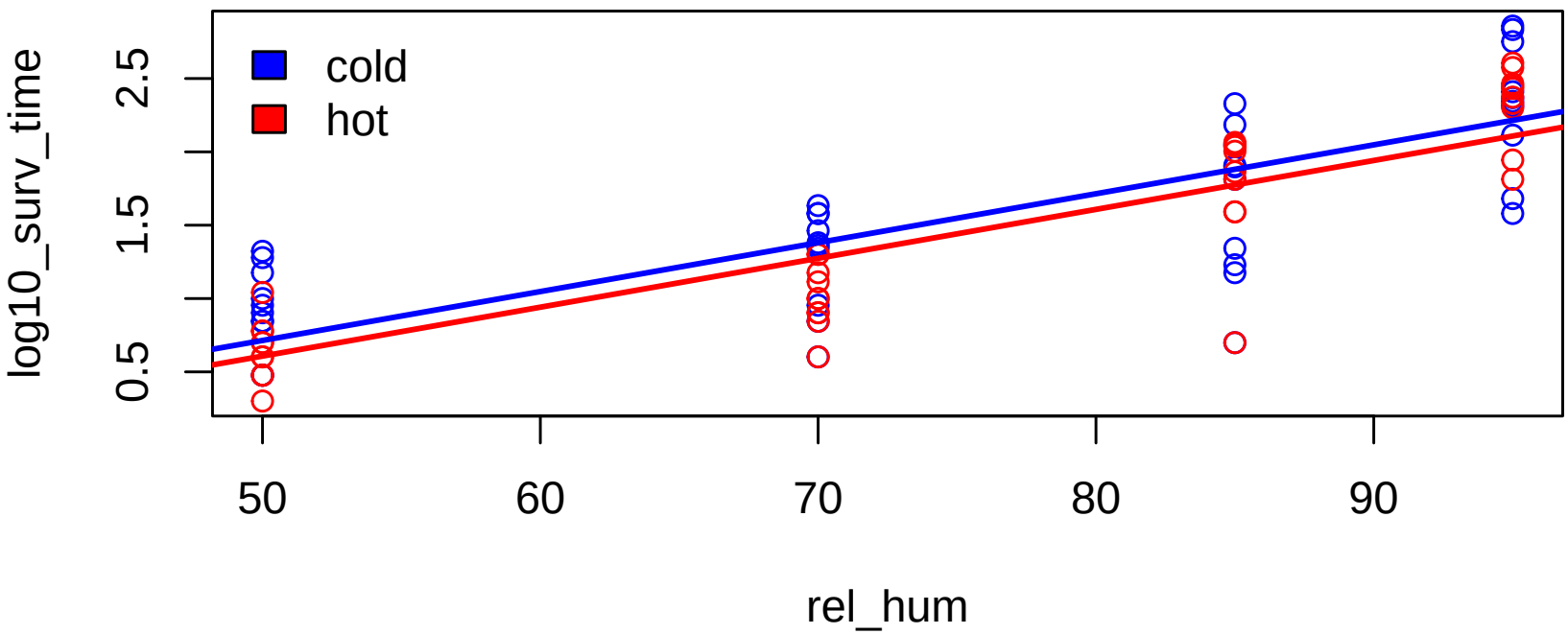
```
##    fit  lwr  upr
## 1 1.11 0.33 1.89
```

```r
# Prediction using the model with interaction
predict(mancovaint, newdata = data4pred, interval = "prediction")
```

```
##    fit   lwr  upr
## 1 1.07 0.295 1.84
```

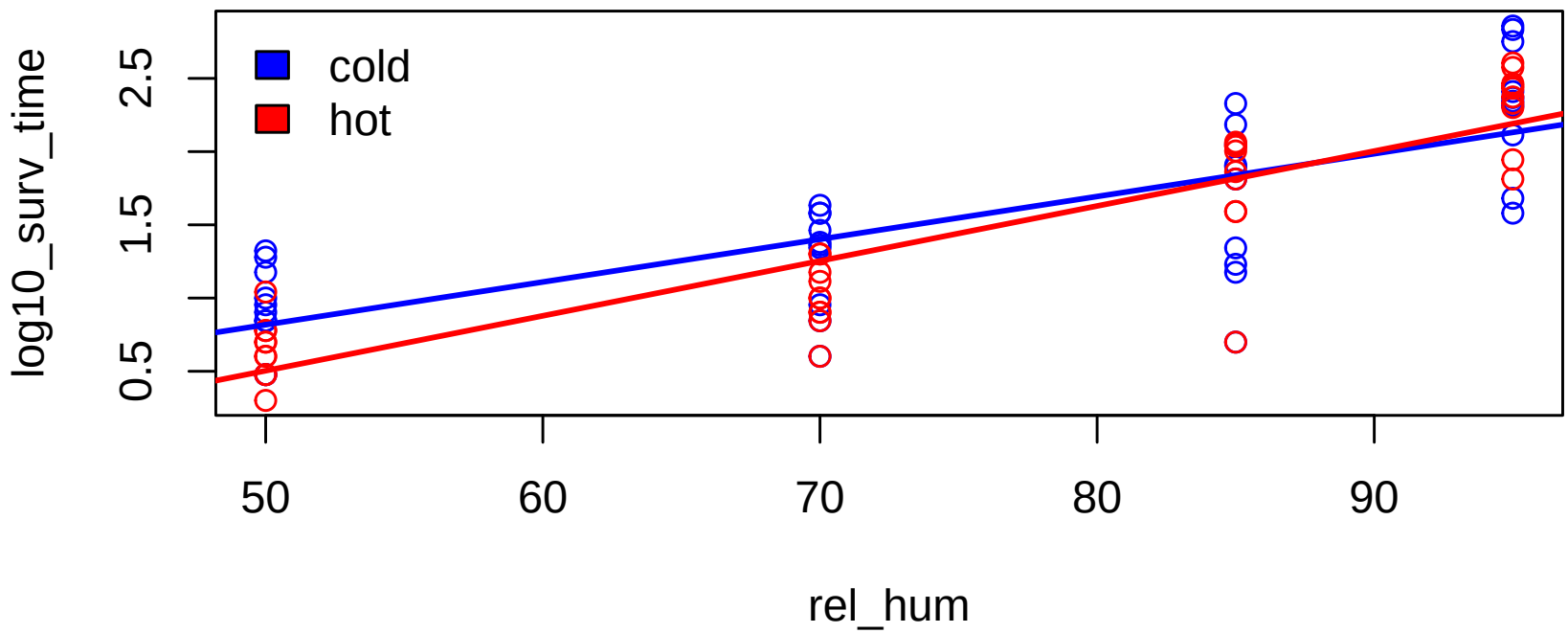
Réponse à la question 3 - modèle sans interaction

```
a <- coef(mancova); par(mar = c(4, 4, 1, 1))
plot(log10_surv_time ~ rel_hum , data = dhum,
     col = ifelse(temperatureF == "cold", "blue", "red"))
abline(a = a[1], b = a[2],lwd = 2, col = "blue")
abline(a = a[1] + a[3], b = a[2], lwd = 2, col = "red")
legend("topleft", fill = c("blue", "red"),
       legend = c("cold", "hot"), bty = "n")
```



Réponse à la question 3 - modèle avec interaction

```
ai <- coef(mancovaint); par(mar = c(4, 4, 1, 1))
plot(log10_surv_time ~ rel_hum , data = dhum,
     col = ifelse(temperatureF == "cold", "blue", "red"))
abline(a = ai[1], b = ai[2],lwd = 2, col = "blue")
abline(a = ai[1] + ai[3], b = ai[2] + ai[4], lwd = 2, col = "red")
legend("topleft", fill = c("blue", "red"),
       legend = c("cold", "hot"), bty = "n")
```



Extensions du modèle linéaire gaussien

Extensions du modèle linéaire gaussien

Vous avez maintenant toutes les pièces de lego pour **construire des modèles linéaires plus complexes**.

Et vous avez les bases pour **comprendre les différentes extensions du modèle linéaire**.

Régression non linéaire

Si le **modèle est une fonction non linéaire des paramètres**.

Dans notre exemple, si vous souhaitez modéliser la relation non linéaire entre le temps de survie (en logarithme) et l'humidité relative à l'aide d'un modèle plus réaliste que le polynôme du second ordre.

Regression logistique

La **régression logistique** (cas particulier du modèle linéaire généralisé - GLM) est utilisée lorsque le résultat n'est plus continu, mais qu'il s'agit d'un **résultat binaire**.

Dans notre exemple, si vous prenez comme variable à expliquer la survie ou non à un moment spécifique.

La régression logistique est très souvent utilisée pour **identifier les facteurs de risque**, par par exemple avec comme variable à expliquer la présence ou non d'une maladie dans les élevages.

Modèles mixtes

Les **modèles mixtes** sont des modèles qui prennent en compte les **effets aléatoires** des facteurs aléatoires ($\neq$ effets déterministes des facteurs fixes),

tels que l'effet *exploitation*, lorsqu'il y a plus d'une observation par exploitation, ou l'effet *animal*, lorsqu'il y a plus d'une observation par animal, ..,

Modèles de survie

Notre exemple était très spécifique puisque le temps de survie était connu pour tous les individus.

Plus classiquement, les données de survie incluent des données censurées (par exemple, pour les individus qui ne sont pas morts à la fin de l'étude, leur **temps de survie est censuré à droite** : on sait seulement qu'il est supérieur à une valeur).

De plus, la distribution des temps de survie n'est pas nécessairement log-normale (hypothèse que nous avons faire sur notre exemple fil rouge).

Ces problèmes peuvent être pris en compte en utilisant une **approche semi-paramétrique** (modèle de Cox) ou une **approche paramétrique** (voir Wongnak *et al.* 2022 sur notre exemple).

Wongnak, P., *et al.* (2022). A hierarchical Bayesian approach for incorporating expert opinions into parametric survival models: a case study of female Ixodes ricinus ticks exposed to various temperature and relative humidity conditions. *Ecological Modelling*, 464, 109821.

Perspectives

Dans les présentations suivantes nous aborderons ou approfondirons les points suivants :

- la **stratégie de construction des modèles** (notamment quelles variables indépendantes incorporer),
- la **compréhension de leurs coefficients** et leur **présentation** dans des publications scientifiques,
- **quelques limites de l'approche**