The linear model: theory and practice using R

Marie Laure Delignette-Muller - VetAgro Sup - LBBE

2022-10-11

Introduction

Introduction

Linear models form the basis of much of statistical practice.

My objectives:

- to make you well **understand the main concepts** behind linear models,
- to introduce you to the **handling of linear models using R**,
- to develop your **critical faculties with regard to the published works** using linear models.

Our example

We will present various models on a unique data set extracted from an old publication describing the observed **survival time** (in days) of **adult ticks** as a function of **temperature** and **relative humidity**:

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 of the whole data set

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

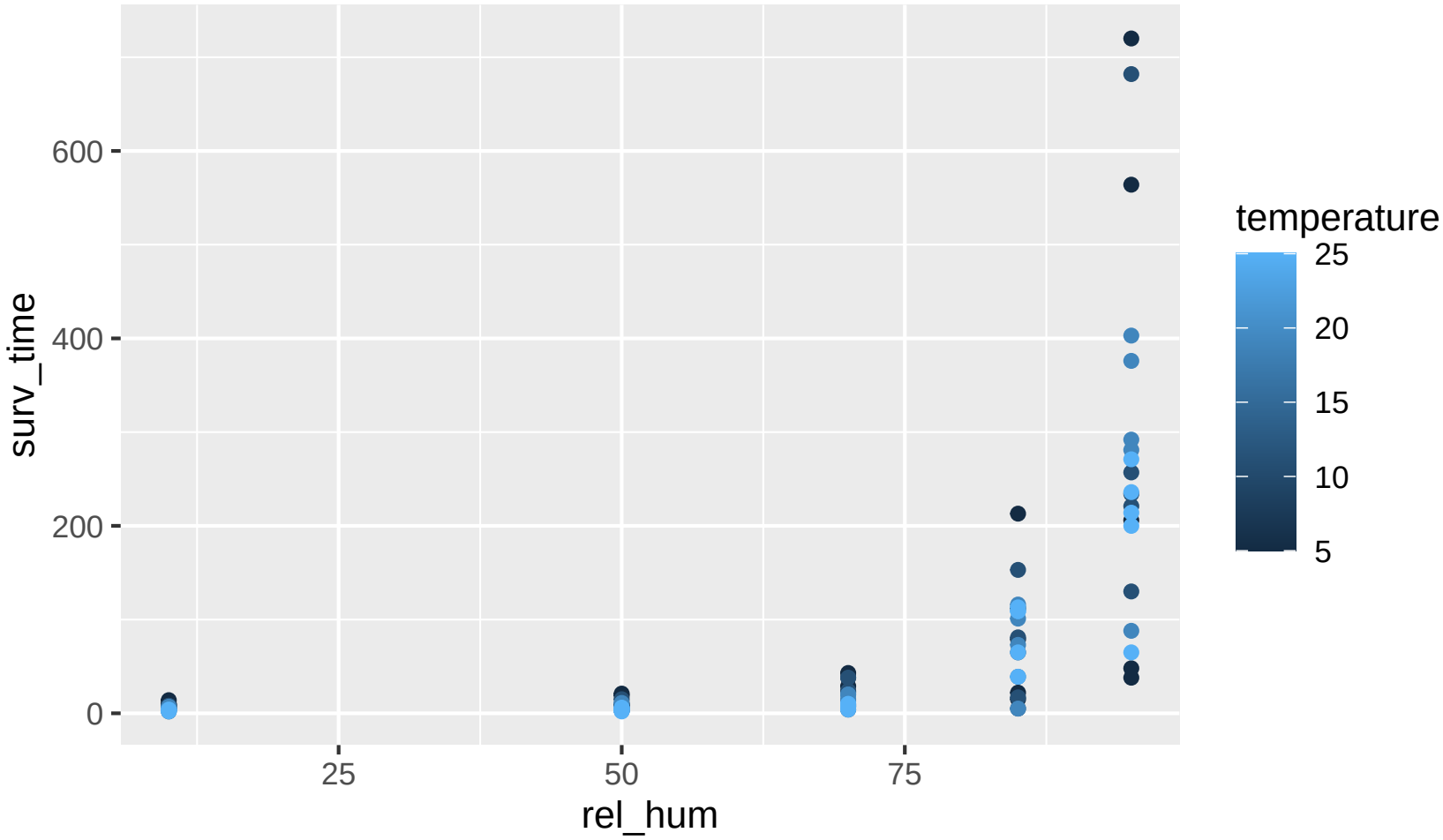
```
## 'data.frame':    100 obs. of  3 variables:
##  $ rel_hum    : int  0 50 70 85 95 0 50 70 85 95 ...
##  $ surv_time  : int  7 7 22 15 38 9 9 23 22 48 ...
##  $ temperature: int  5 5 5 5 5 5 5 5 5 5 ...
```

```
# 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)
```

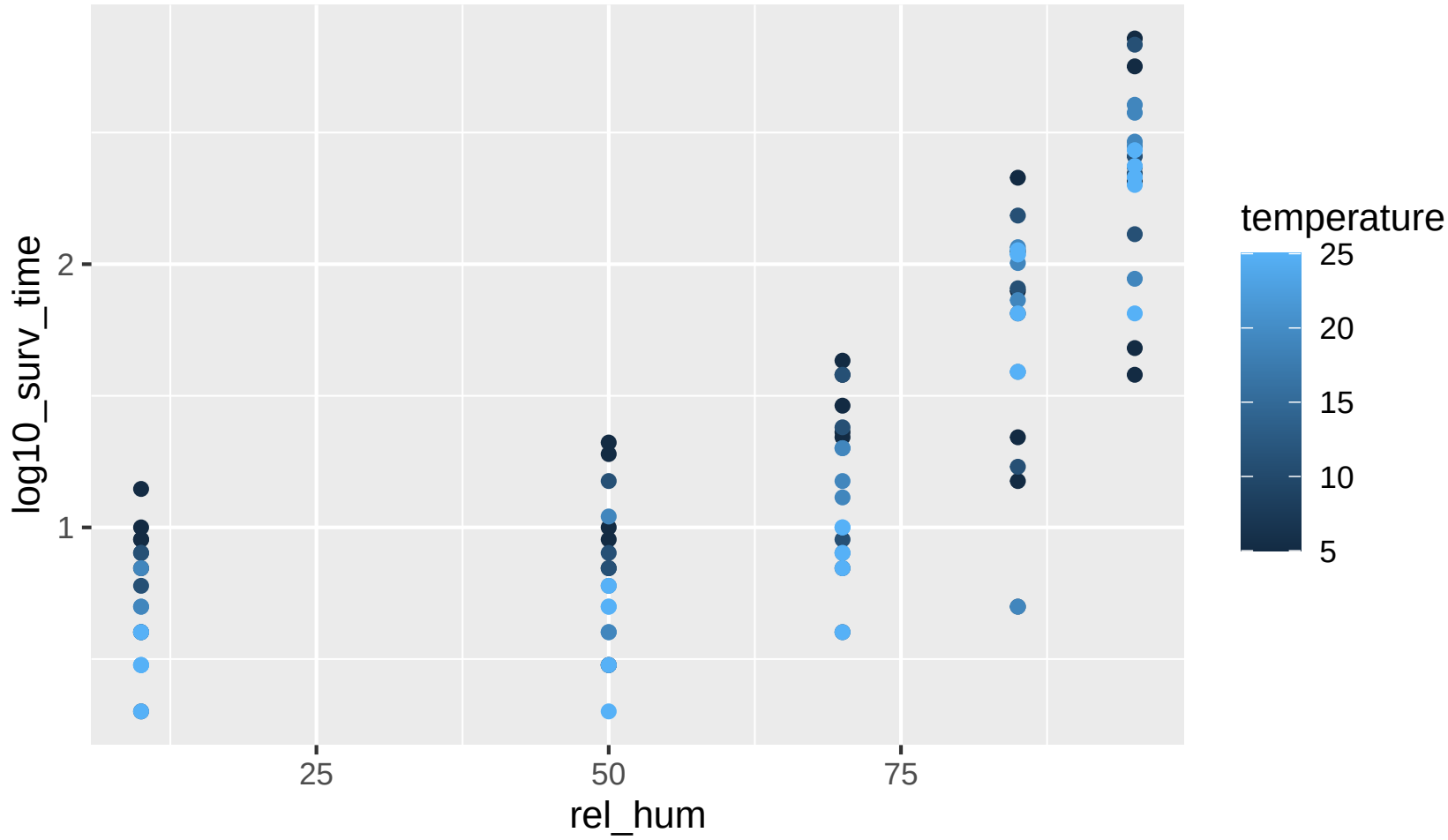
Plot of data using ggplot2

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



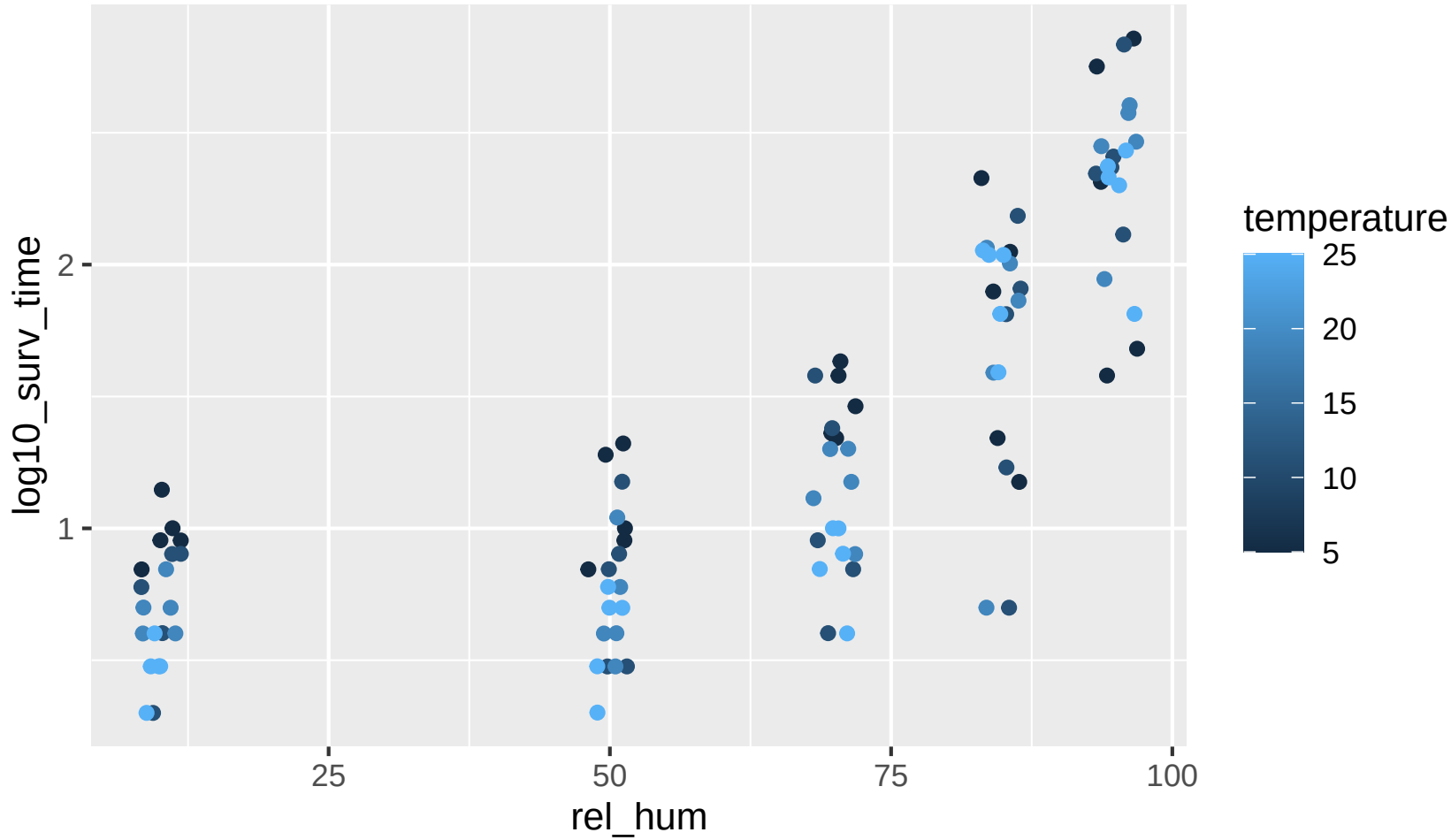
Plot of data after log transformation of the outcome

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



Add jitter

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



Definition of basic terms

The outcome = the **dependent variable** = the survival time (in days) = a **continuous variable**

No censoring here as the experiment was pursued to reach the death for each tick

The explicative variables = the **independent variables**:

- the relative humidity (in %)
- the temperature (in Celsius degrees)

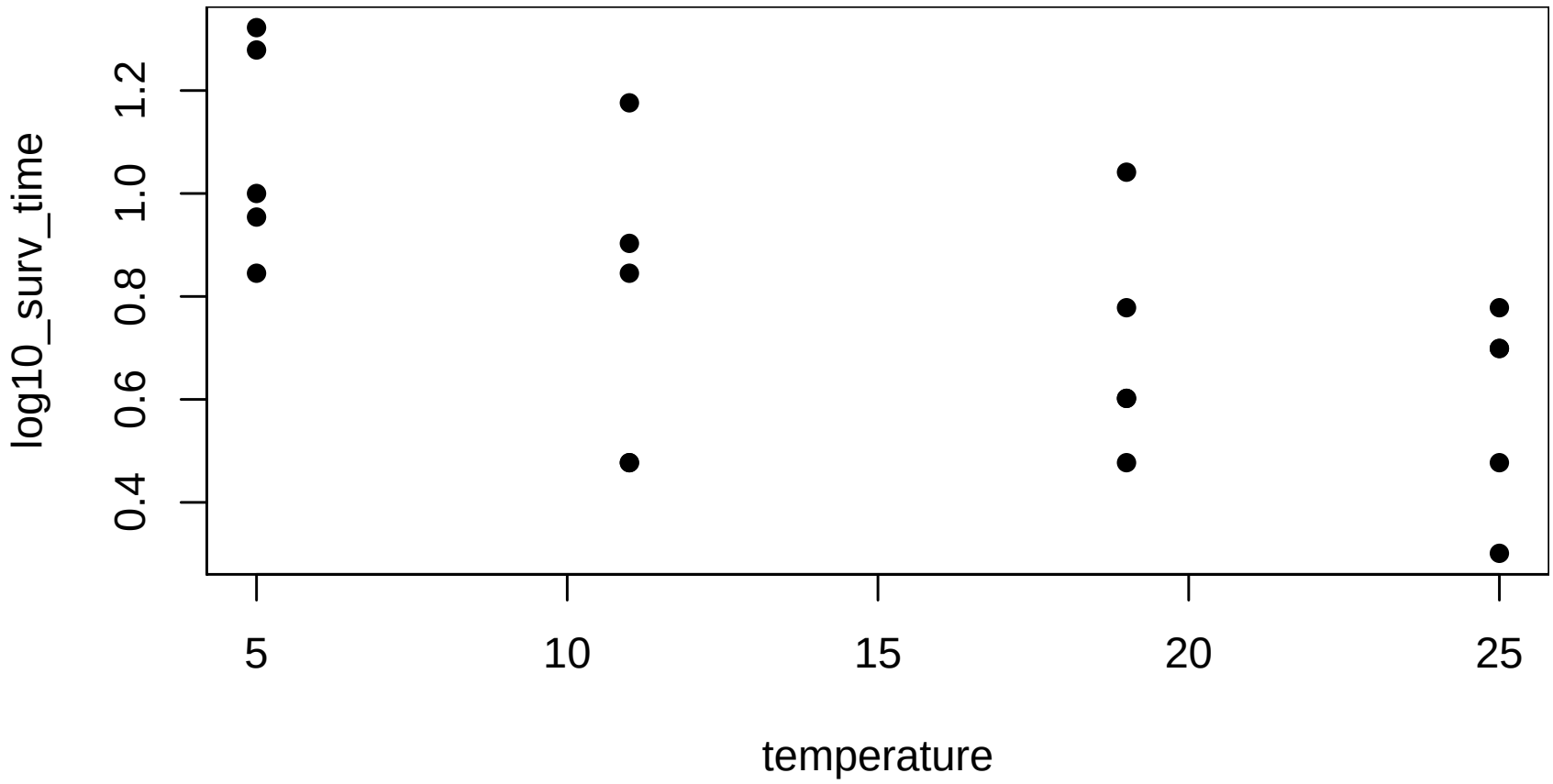
In this example the two explicative variables are controlled (**experimental study**).

Reminder on simple linear regression

Impact of the temperature for mild humidity conditions

using a **subset** of the whole data set **at a humidity level of** 50%

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



Basic concepts and fitting method

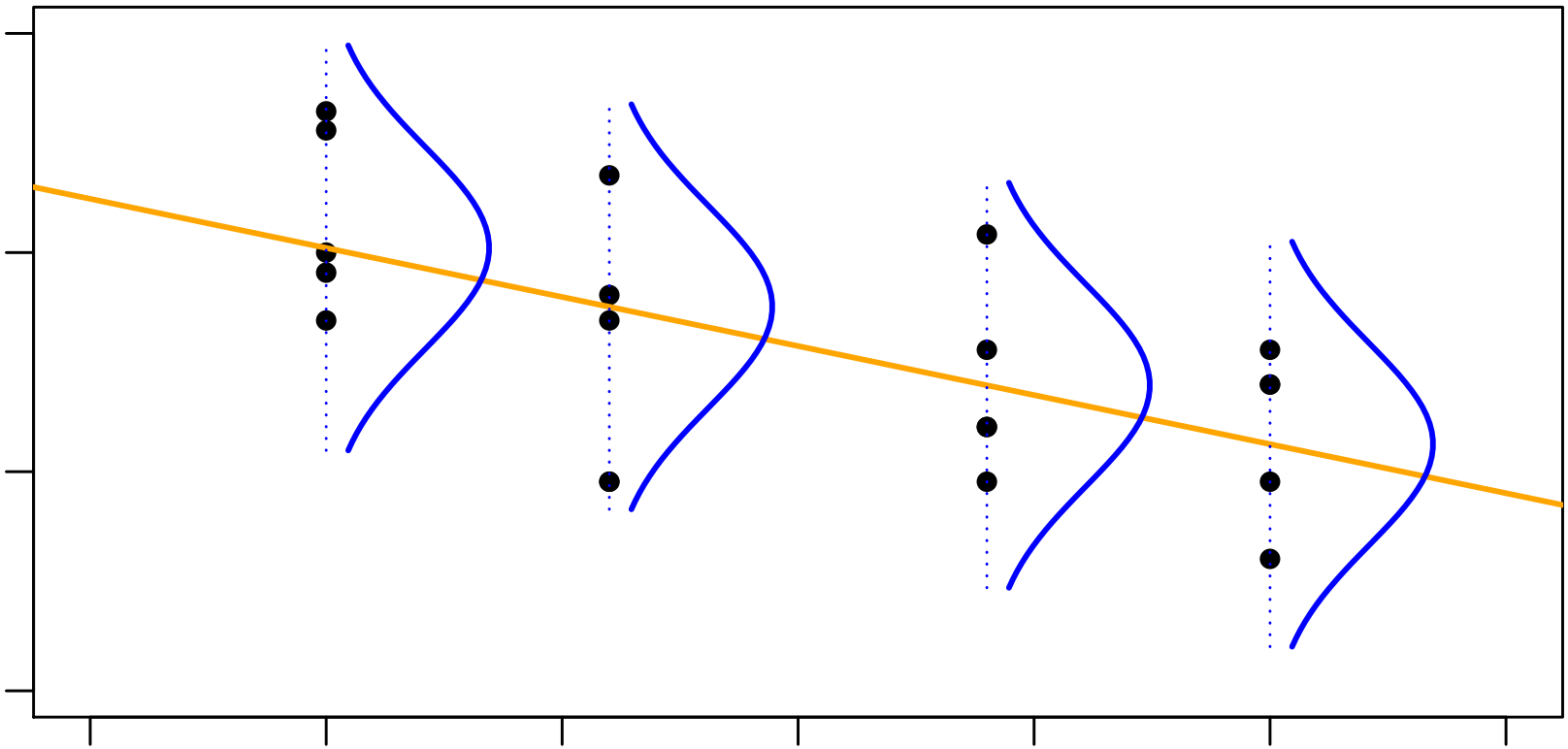
Basic concepts and fitting method

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

Deterministic part: linear link

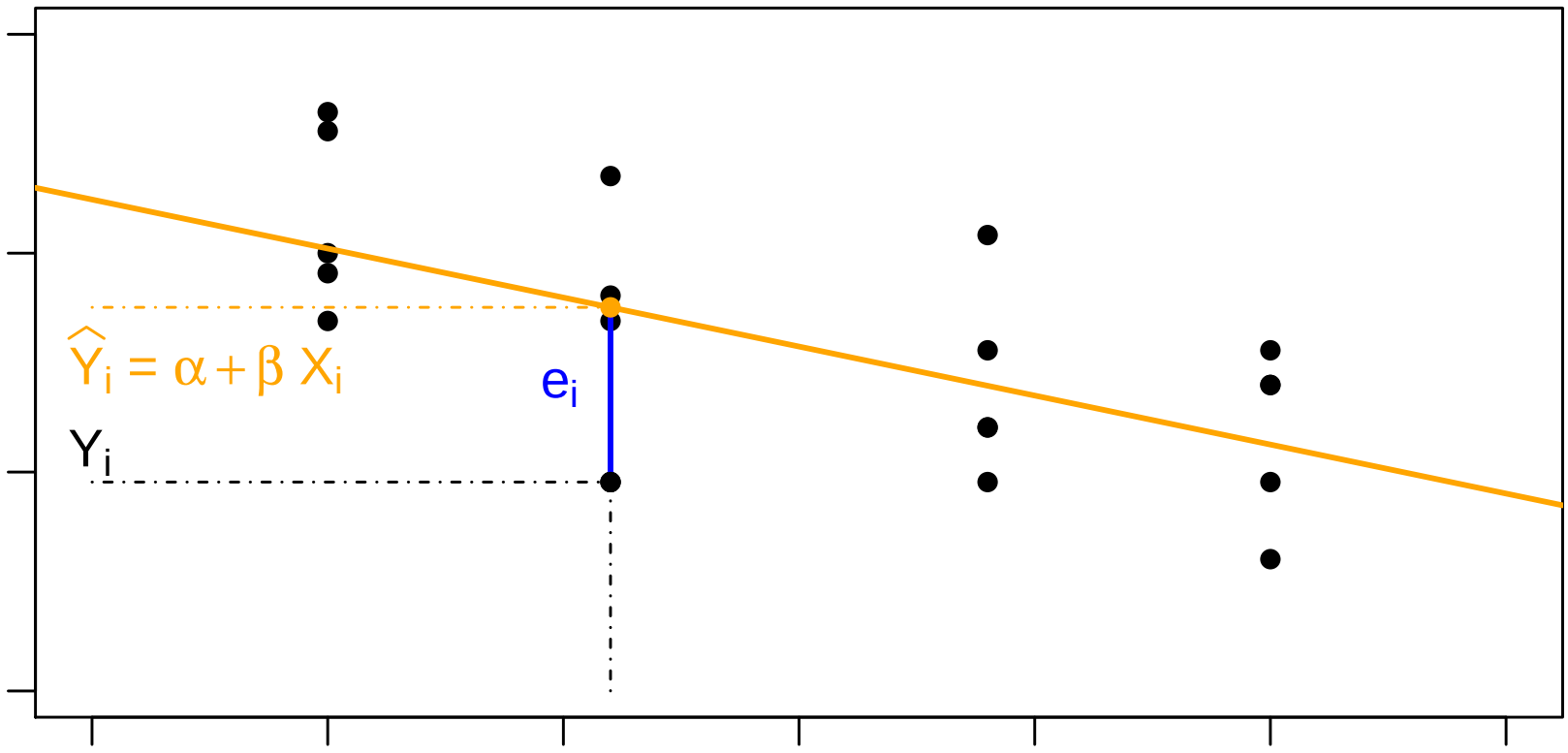
Stochastic part : Gaussian model

assuming **random**, **independent** residuals ϵ_i following a **Gaussian** (normal) distribution of constant variance σ^2.



The least squares estimation of parameters

In the case of this model, the **maximum likelihood estimation** (maximizing $Pr(Y|\alpha, \beta, \sigma)$) corresponds to the **least squares estimation** minimizing $SCE = \sum_{i=1}^{n} e_i^2$ with $e_i = Y_i - \hat{Y}_i$



Estimation of parameters using 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
```

R summary for a linear model (will be detailed later)

```
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 ' '
##
## 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
```

How to interpret the p-value associated to the regression coefficient (slope) ?

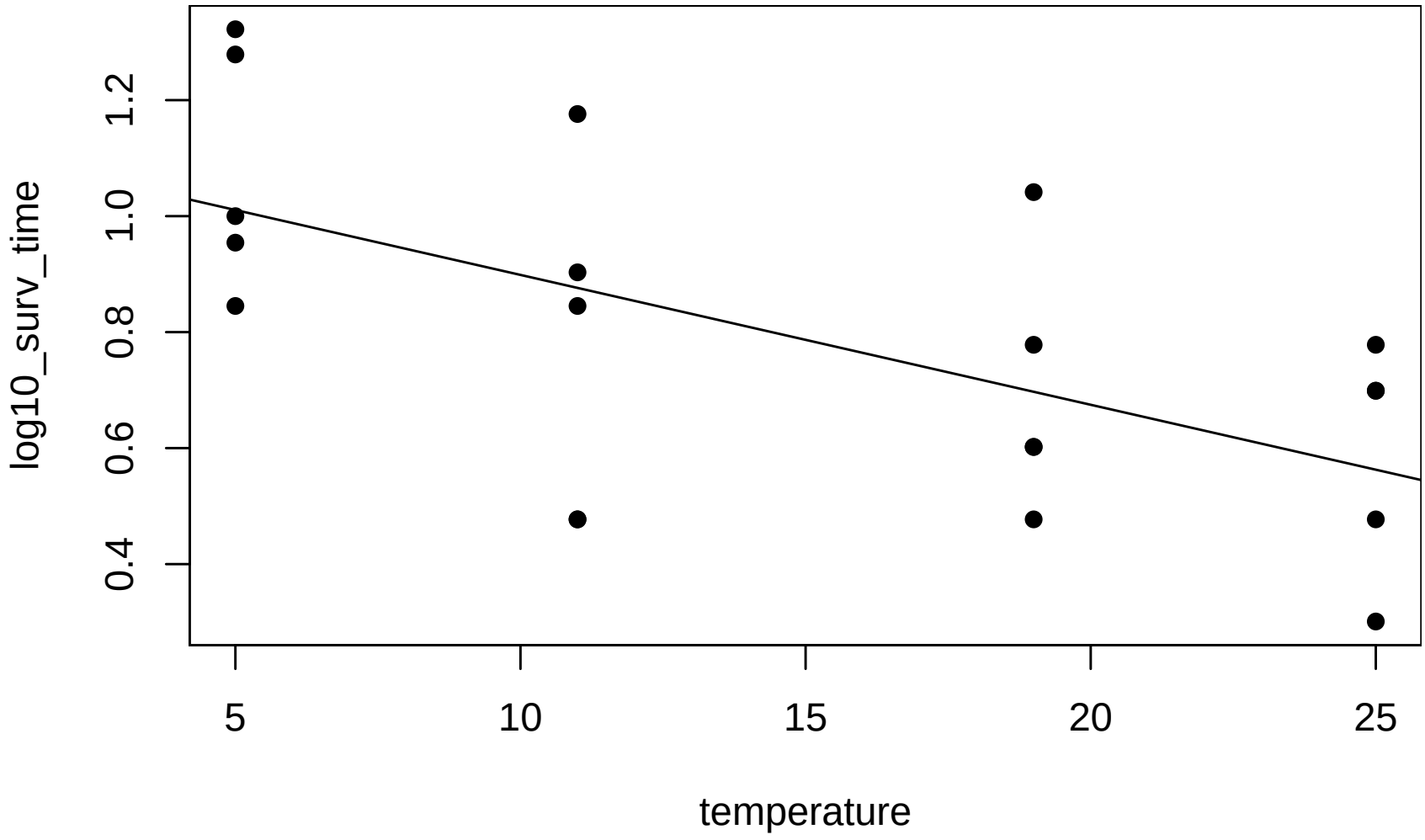
```
summary(m)$coefficients
```

```
##                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
```

It corresponds to the **significance test of the slope** with H_0 the hypothesis of null slope. So it answers to the question: "**is there a significant linear relationship between** the independent variable X **and** the dependent variable Y ?"

Look at the fitted model

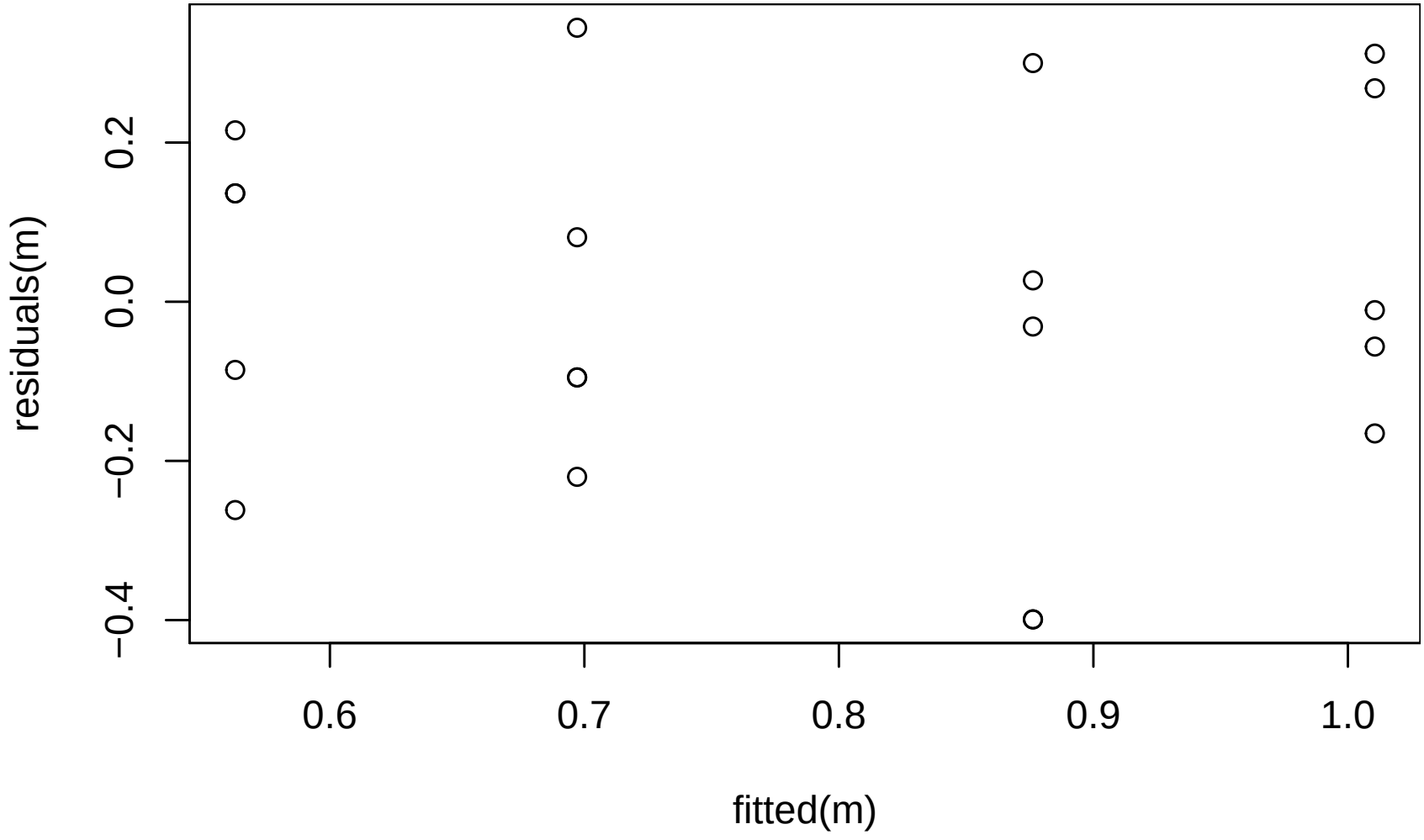
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ temperature, data = dRH50, pch = 16)
abline(m)
```



Check of use conditions

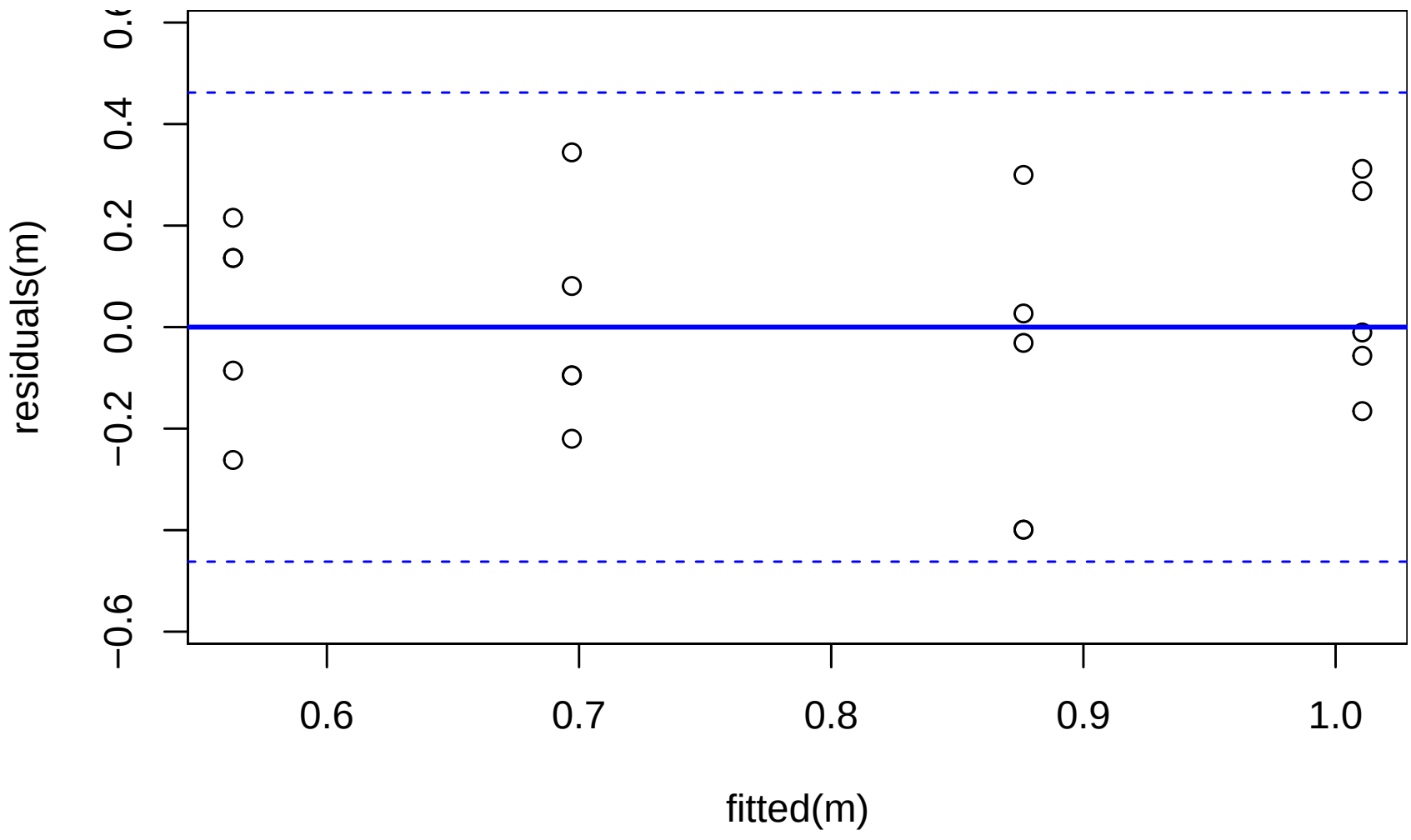
Plot of residuals against fitted values

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



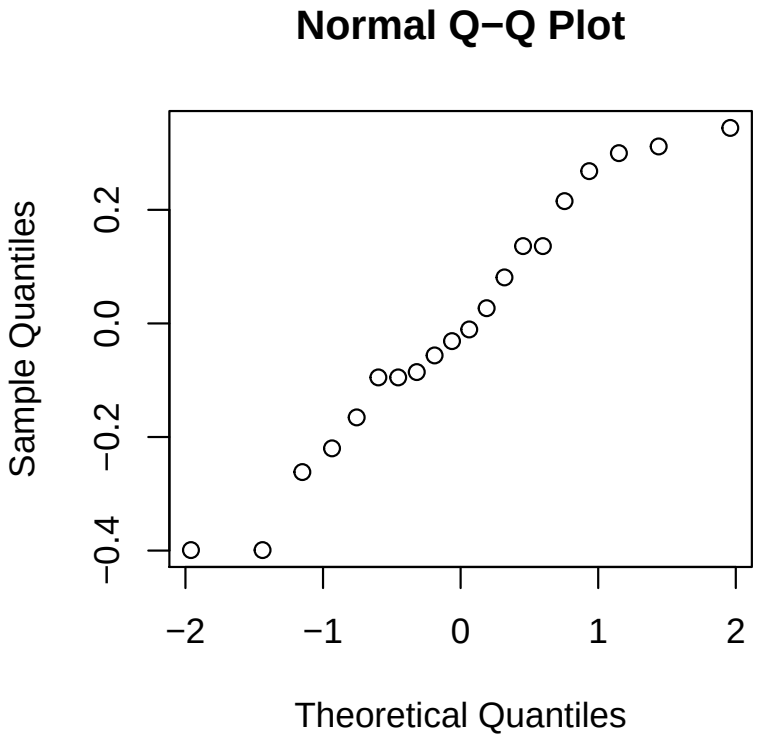
Expected residuals under the model conditions

Centered on 0, 95% within $[-2\sigma; 2\sigma]$, constant variance, no trend.



Quantile-quantile plot of residuals

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



Gaussian (normal) expected global distribution of residuals:

roughly aligned points in the Q-Q plot of residuals

Your turn to play with three other examples

Fit a simple linear model and especially look at the residuals in the three following cases:

1. Fit a simple linear model on the same example **without log transformation** of the survival time
2. Model the **impact of the relative humidity on the survival time** (after log_{10} tranformation) **at** $25°C$.
3. Model the **impact of the relative humidity on the survival time** (after log_{10} tranformation) **at** $19°C$, excluding data at the driest condition.

You can use the R script "intro2linmodel.R" to help you.

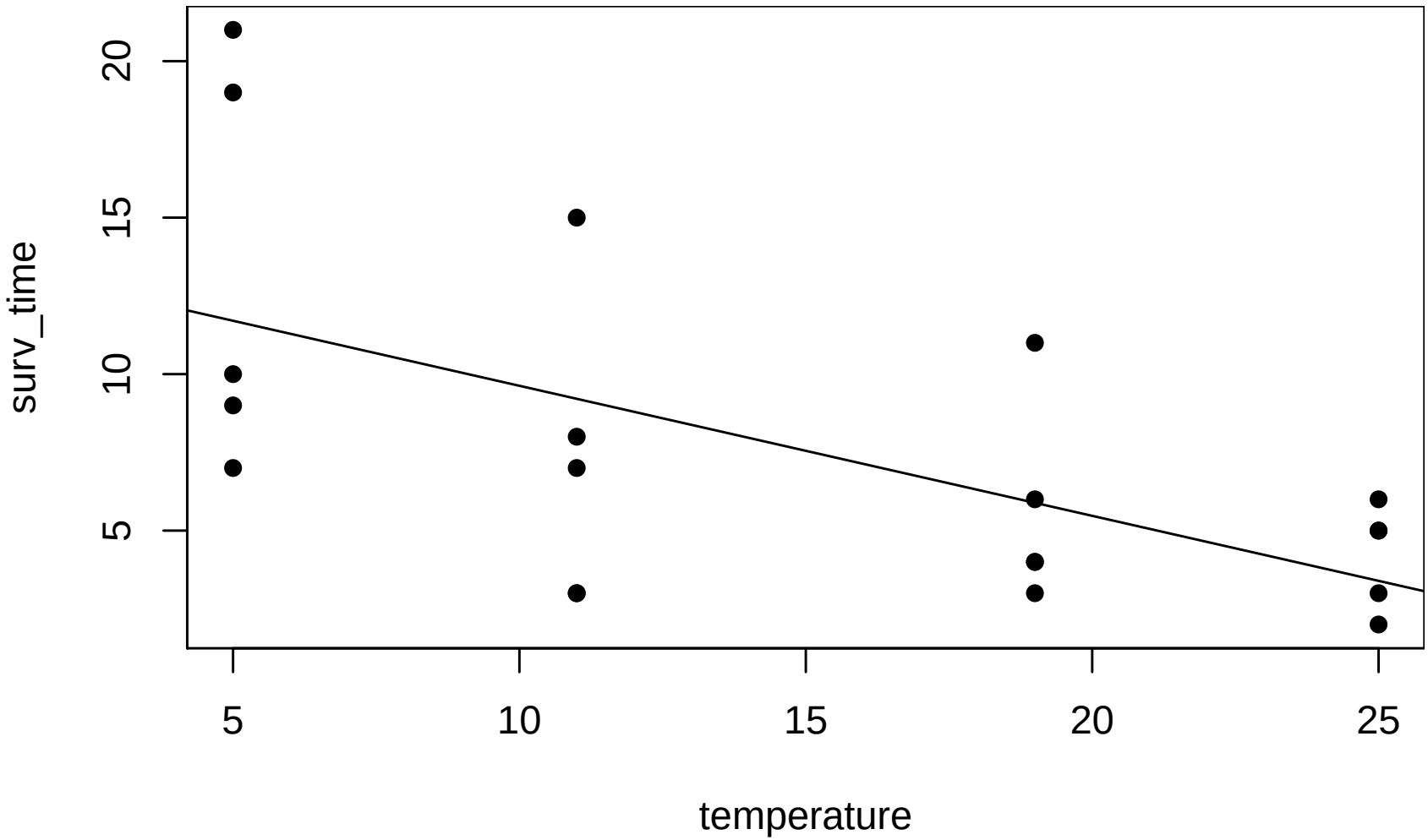
Example 1 - fit of the model

```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
```

Example 1 - plot of the fitted model

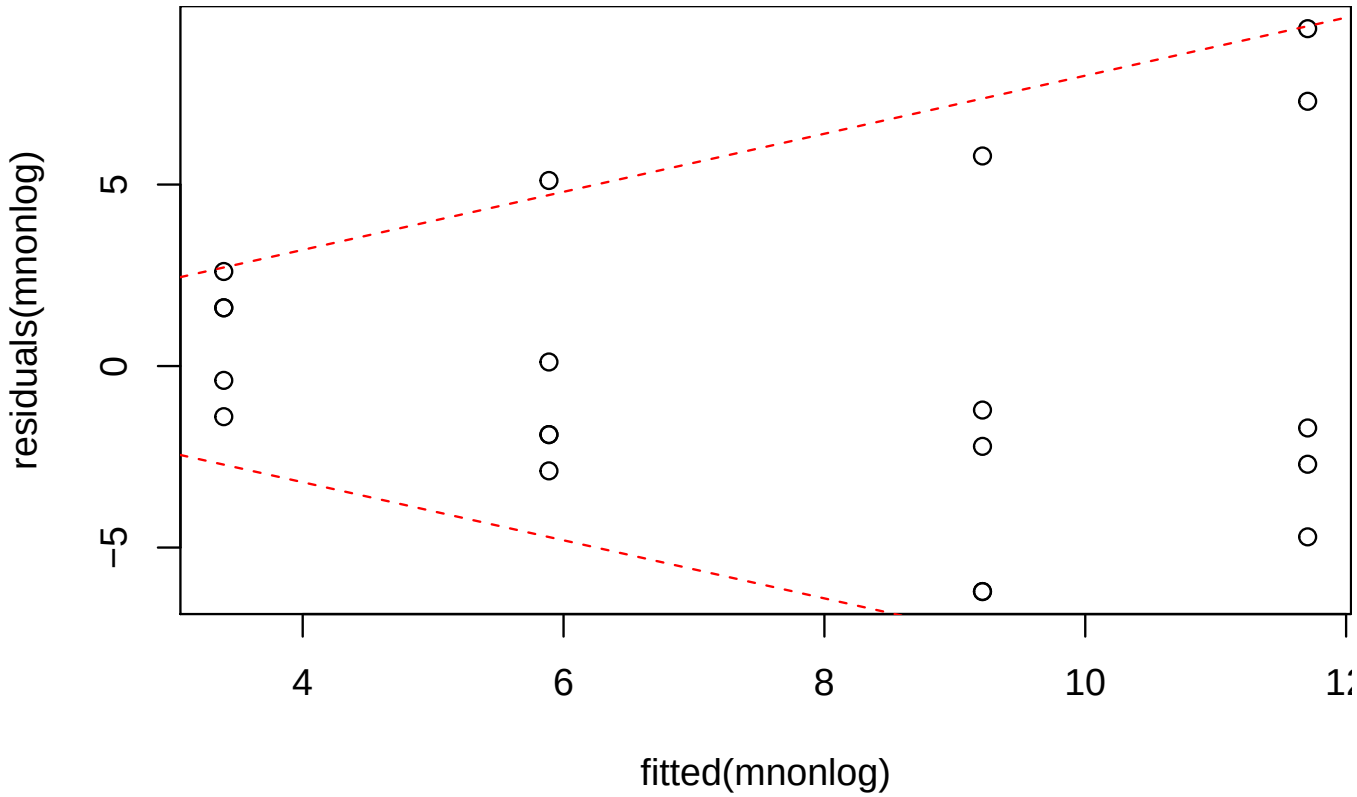
```
par(mar = c(4, 4, 0, 0))
plot(surv_time ~ temperature, data = dRH50, pch = 16)
abline(mnonlog)
```



Example 1 - Plot of residuals

The bottleneck effect = heteroscedasticity = σ is not constant

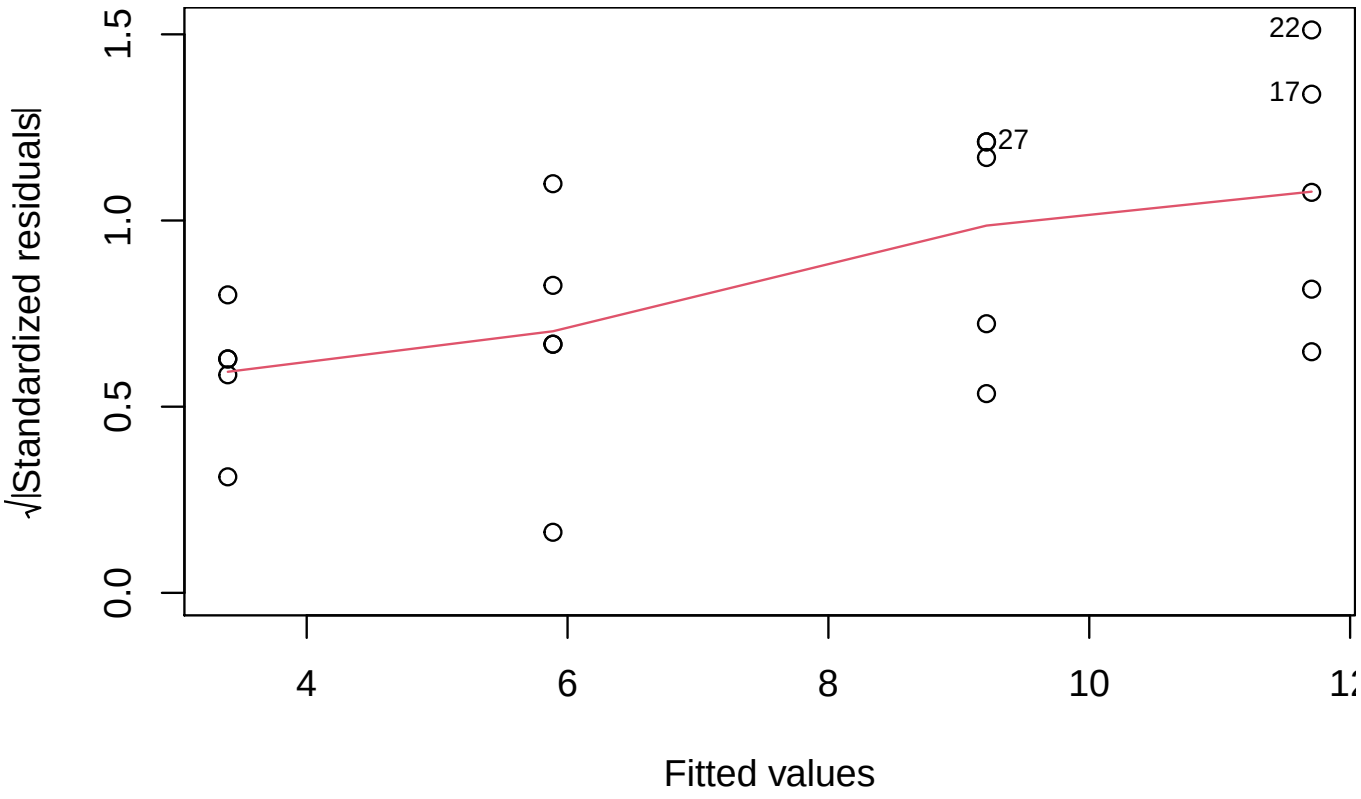
```
plot(residuals(mnonlog) ~ fitted(mnonlog))
```



Example 1 - An R variant to see the problem

Scale-Location plot

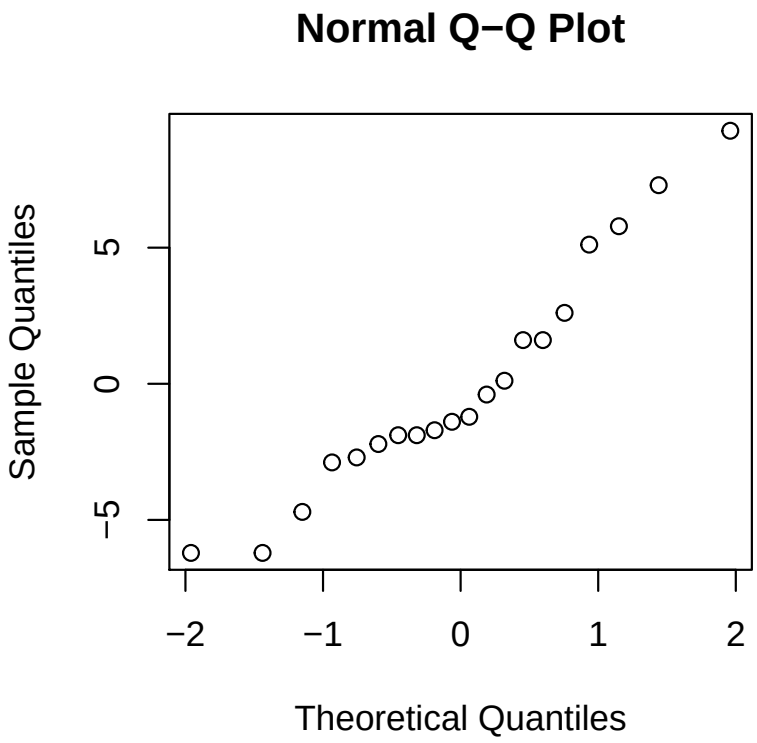
```
par(mar = c(4, 4, 0, 0))
plot(mnonlog, which = 3)
```



Example 1 Q-Q plot of residuals

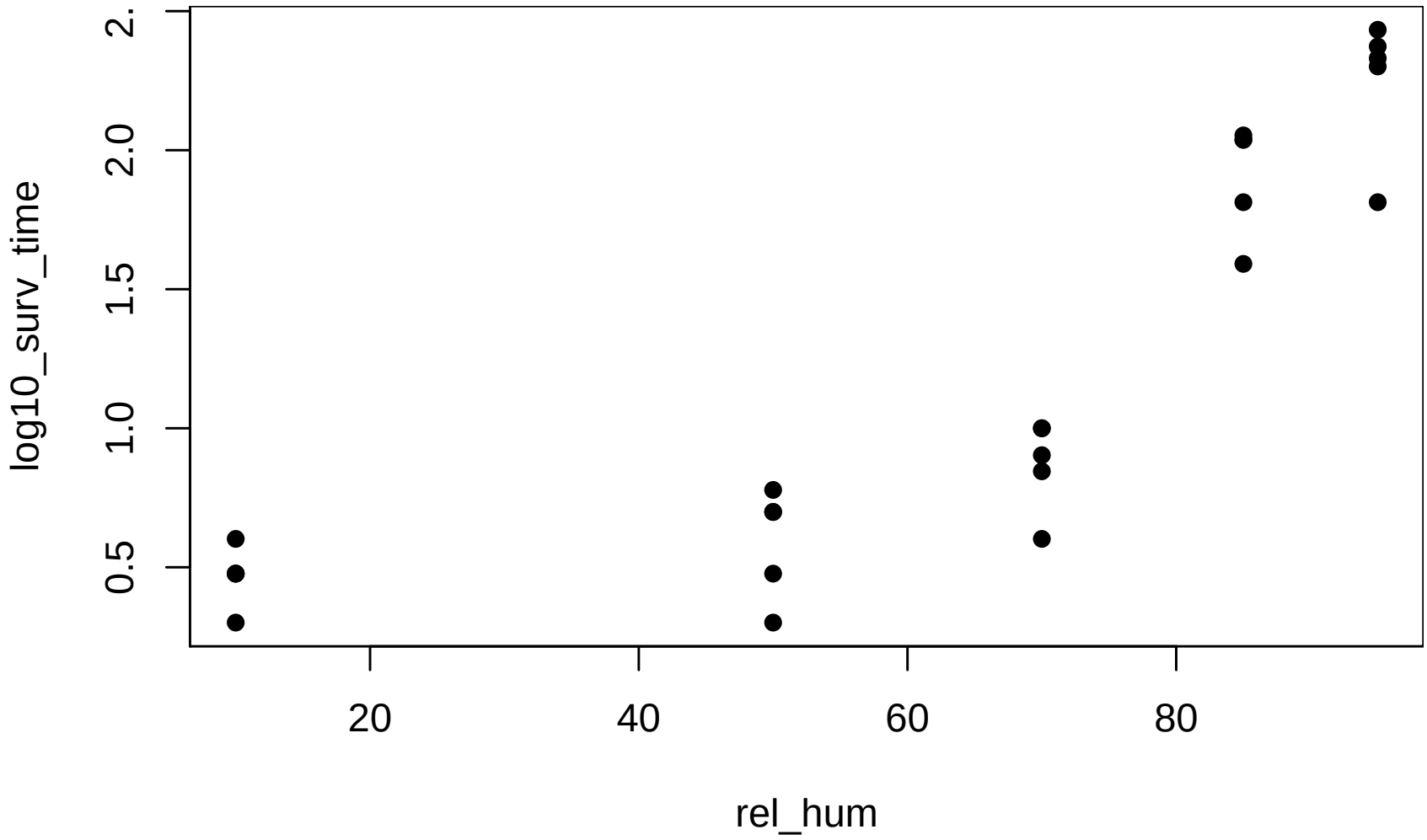
The problem is not detectable on the Q-Q plot in this example.

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



Example 2 - build and look at the data set

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



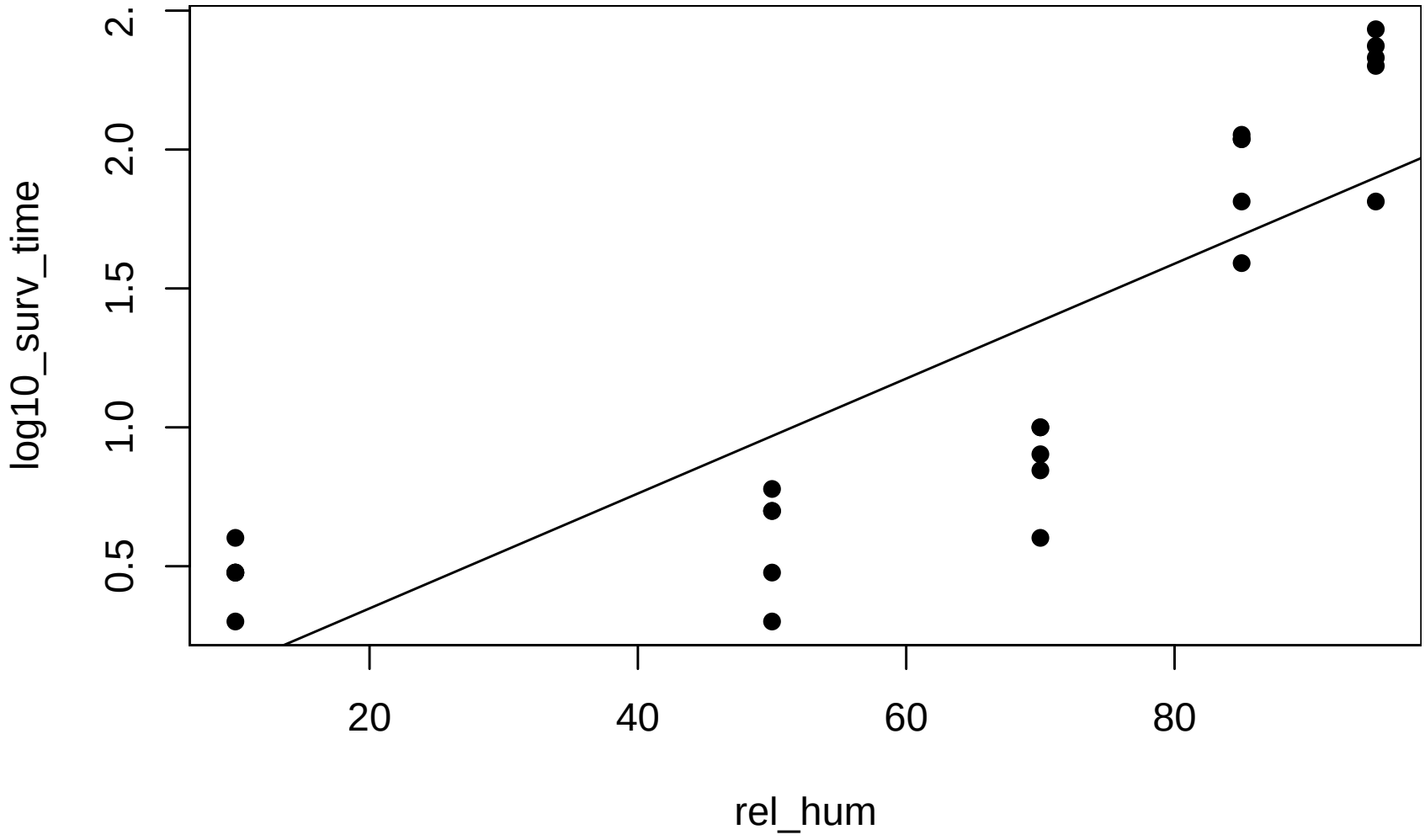
Example 2 - fit of the model

```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
```

Example 2 - plot of the fitted model

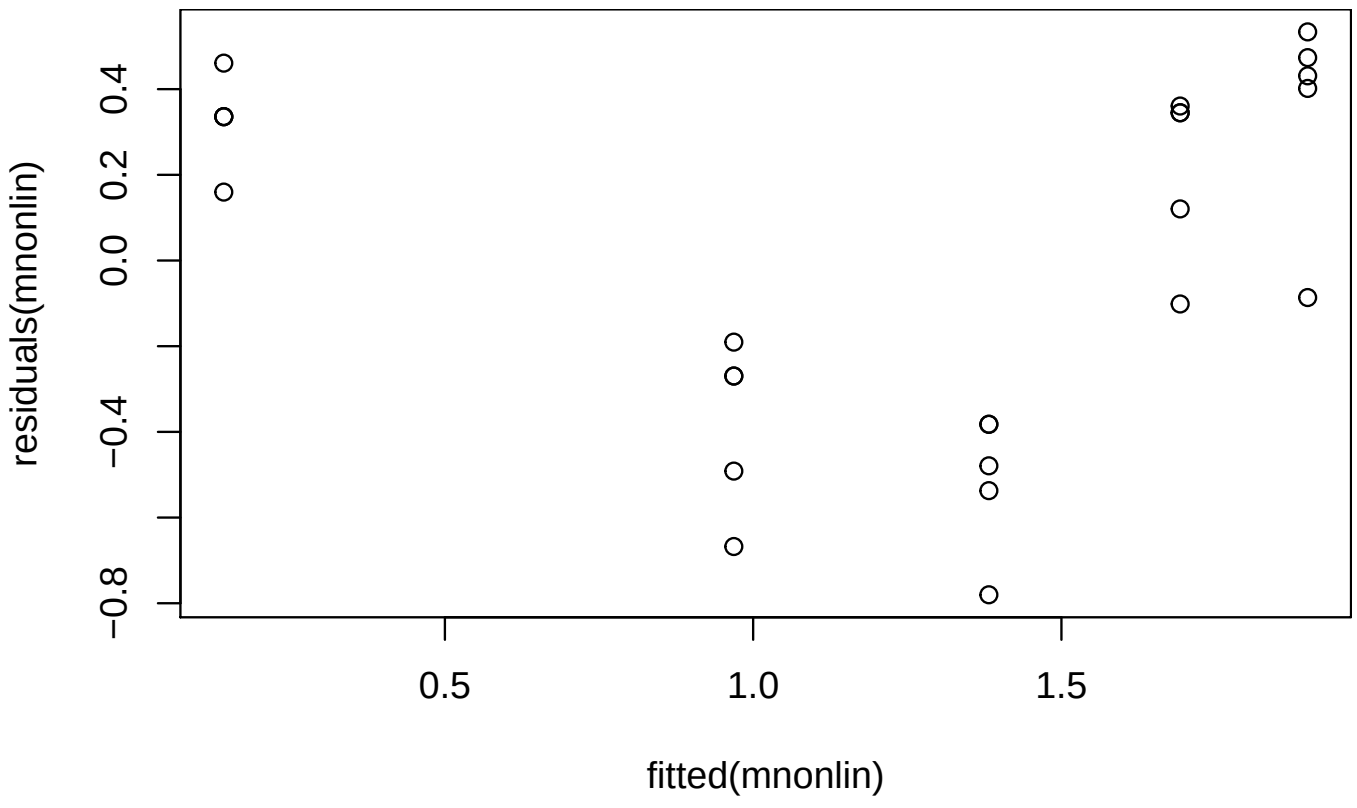
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT25, pch = 16)
abline(mnonlin)
```



Example 2 - Plot of residuals

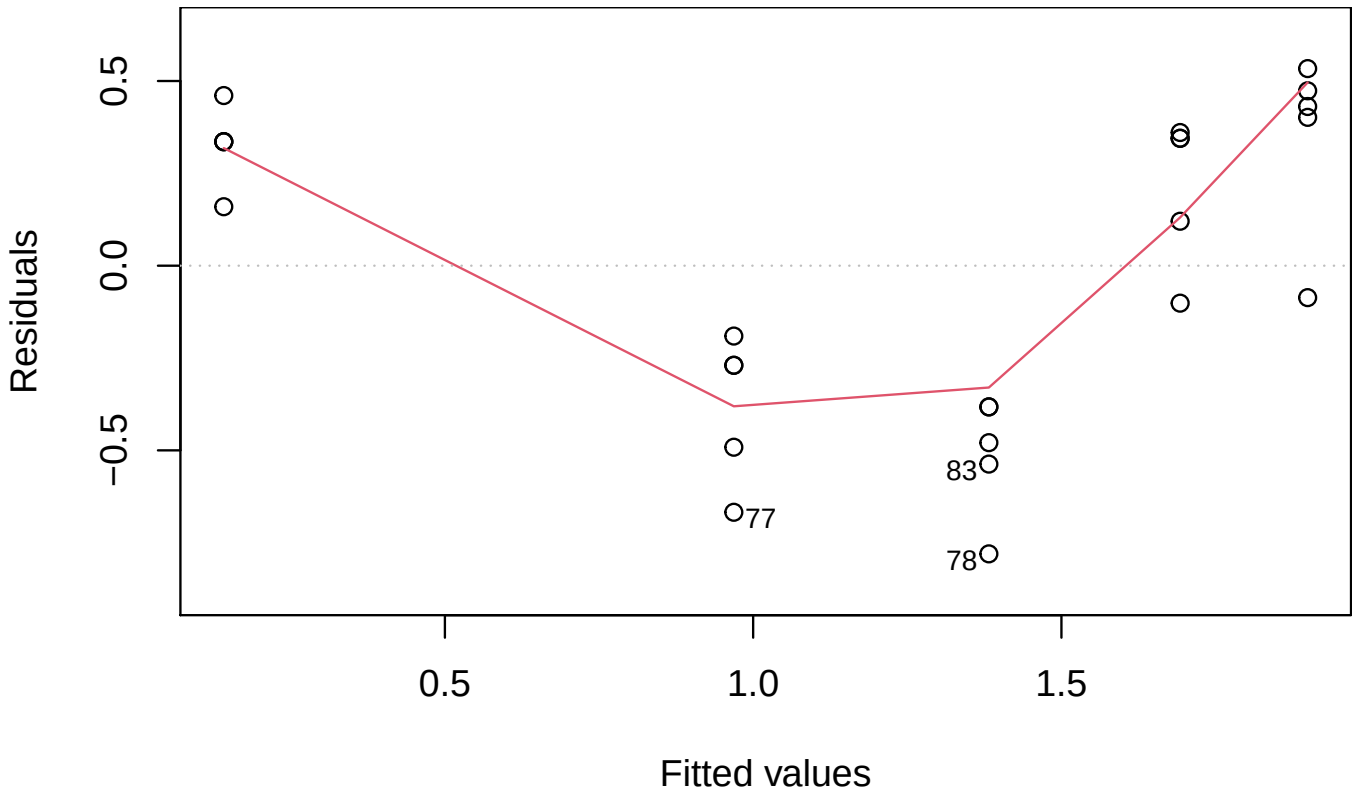
Trend in the residuals = residuals are not independent, in this example due to non linearity of the relation

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



Example 2 - An R variant helping to see the problem

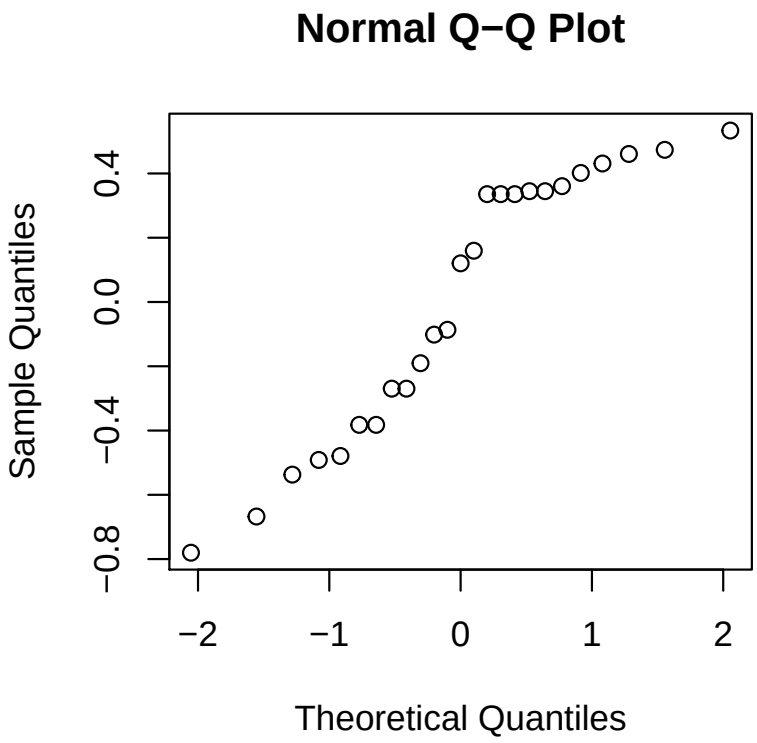
```
par(mar = c(4, 4, 0, 0))
plot(mnonlin, which = 1)
```



Example 2 - Q-Q plot of residuals

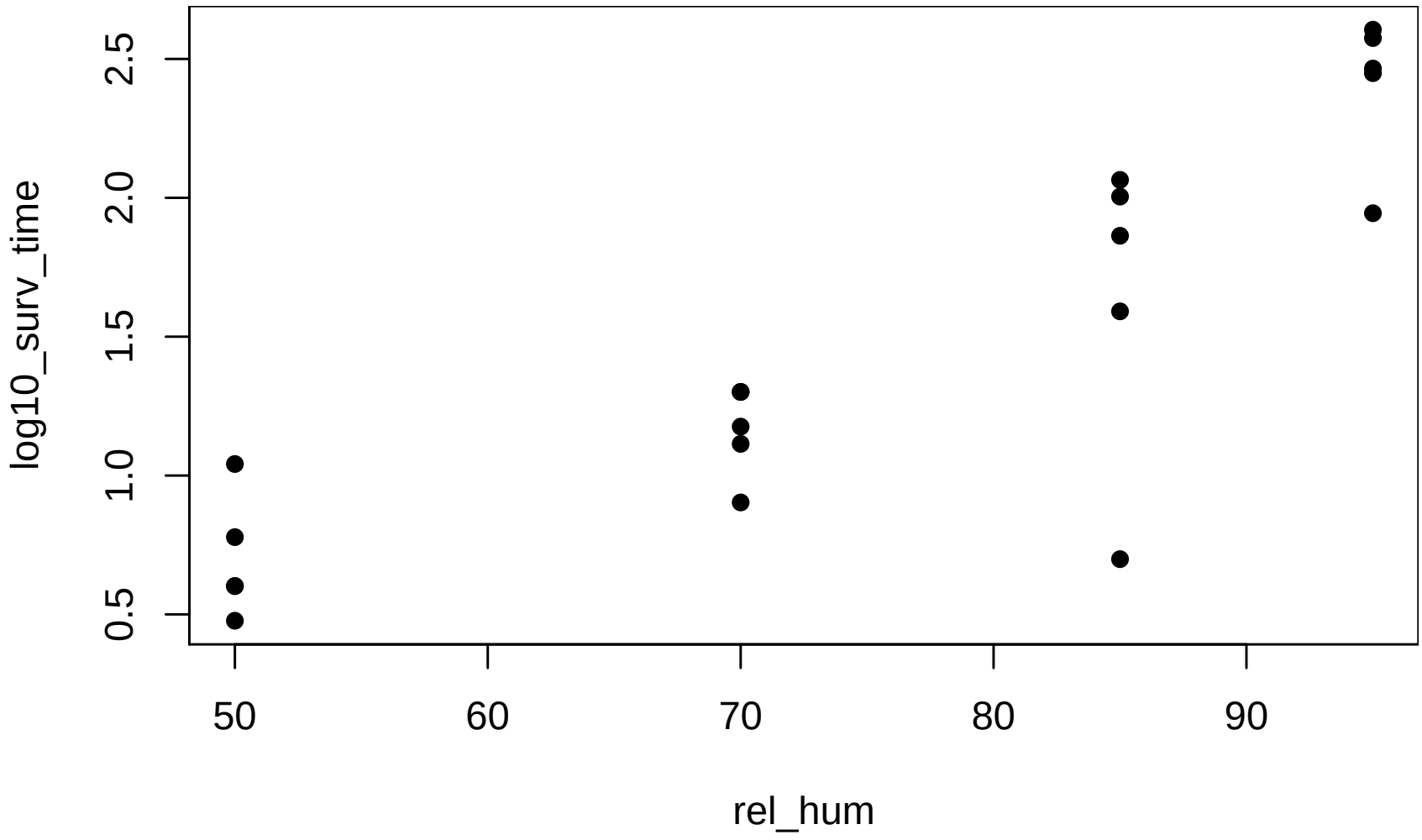
In this example the problem is also detectable on the Q-Q plot.

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



Example 3 - build and look at the data set

```
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)
```



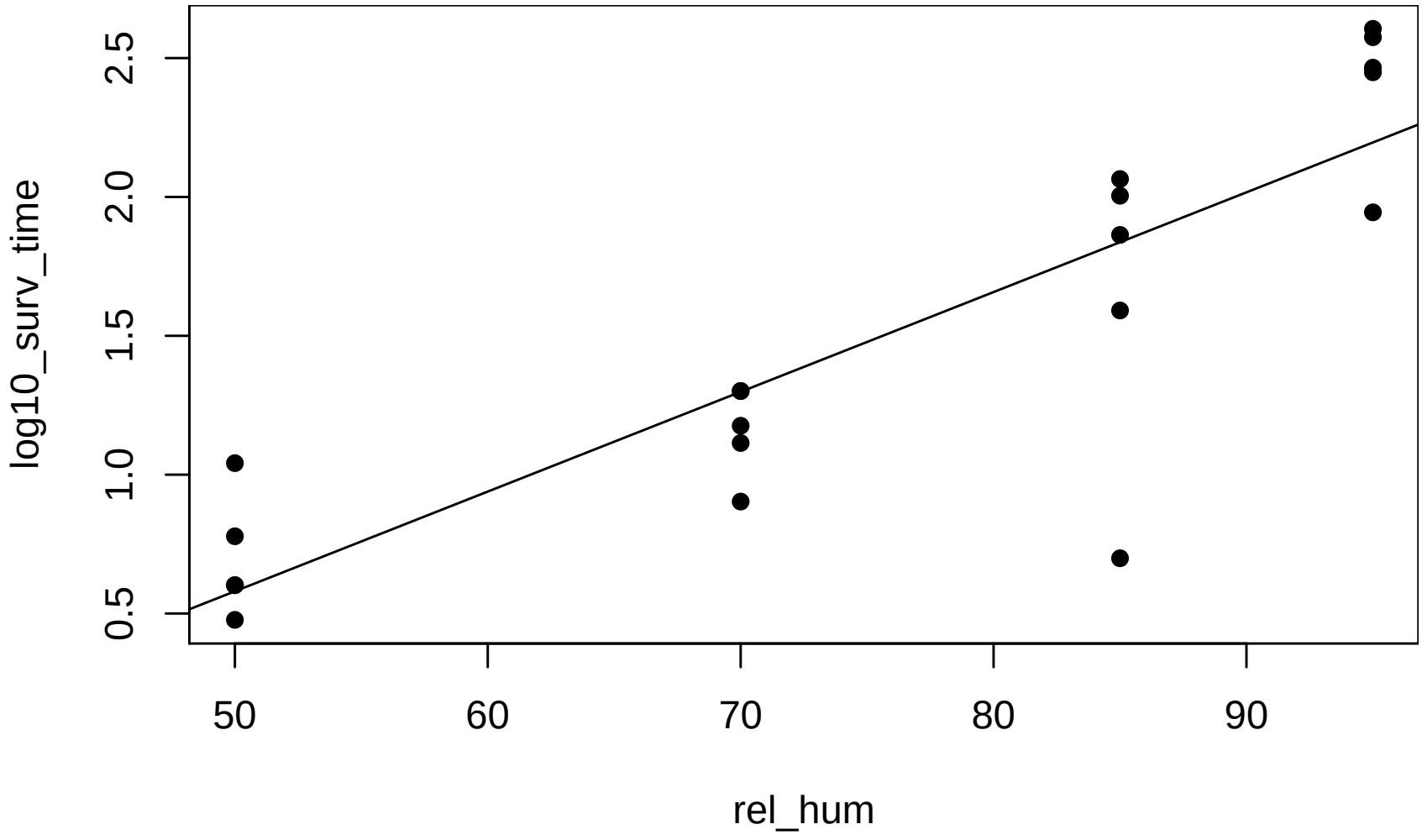
Example 3 - fit of the model

```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
```

Example 3 - plot of the fitted model

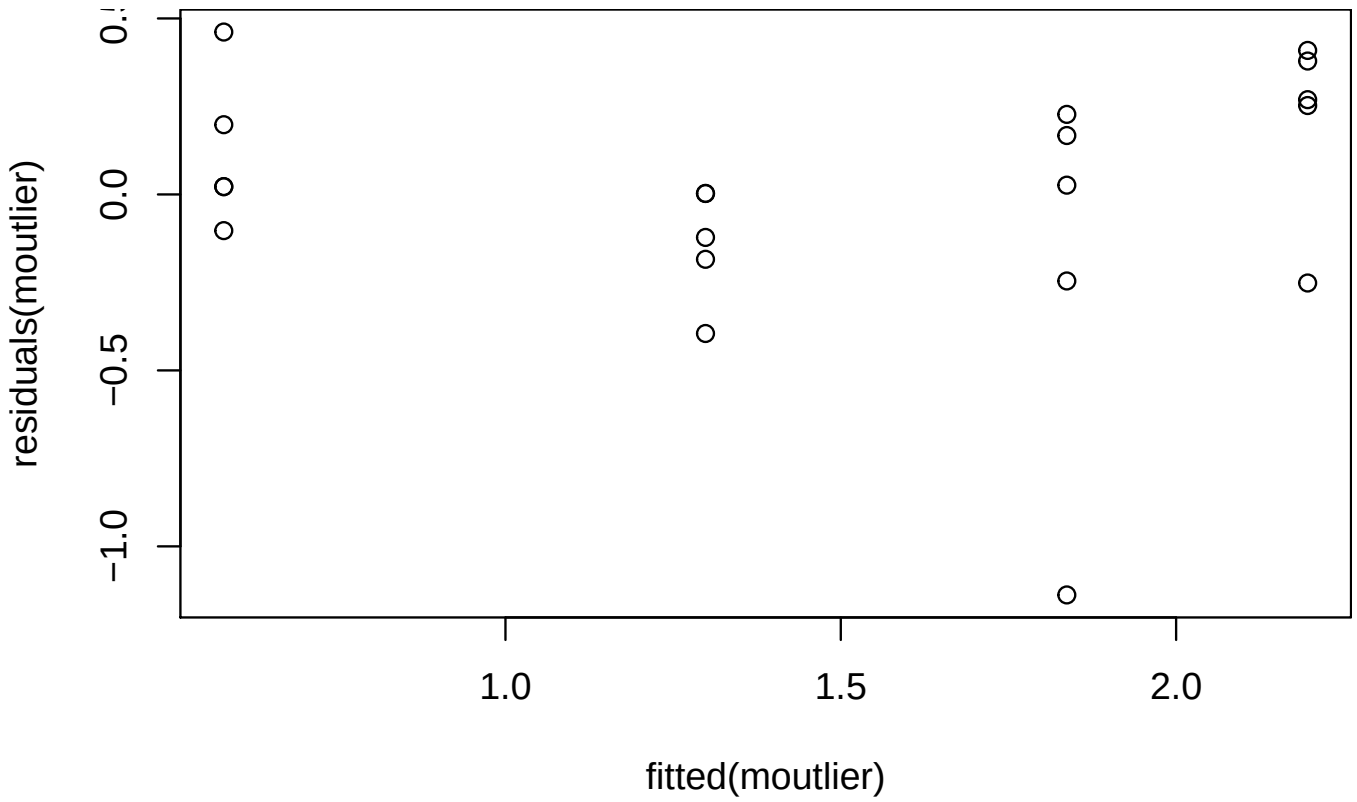
```
par(mar = c(4, 4, 0, 0))
plot(log10_surv_time ~ rel_hum, data = dT19, pch = 16)
abline(moutlier)
```



Example 3 - Plot of residuals

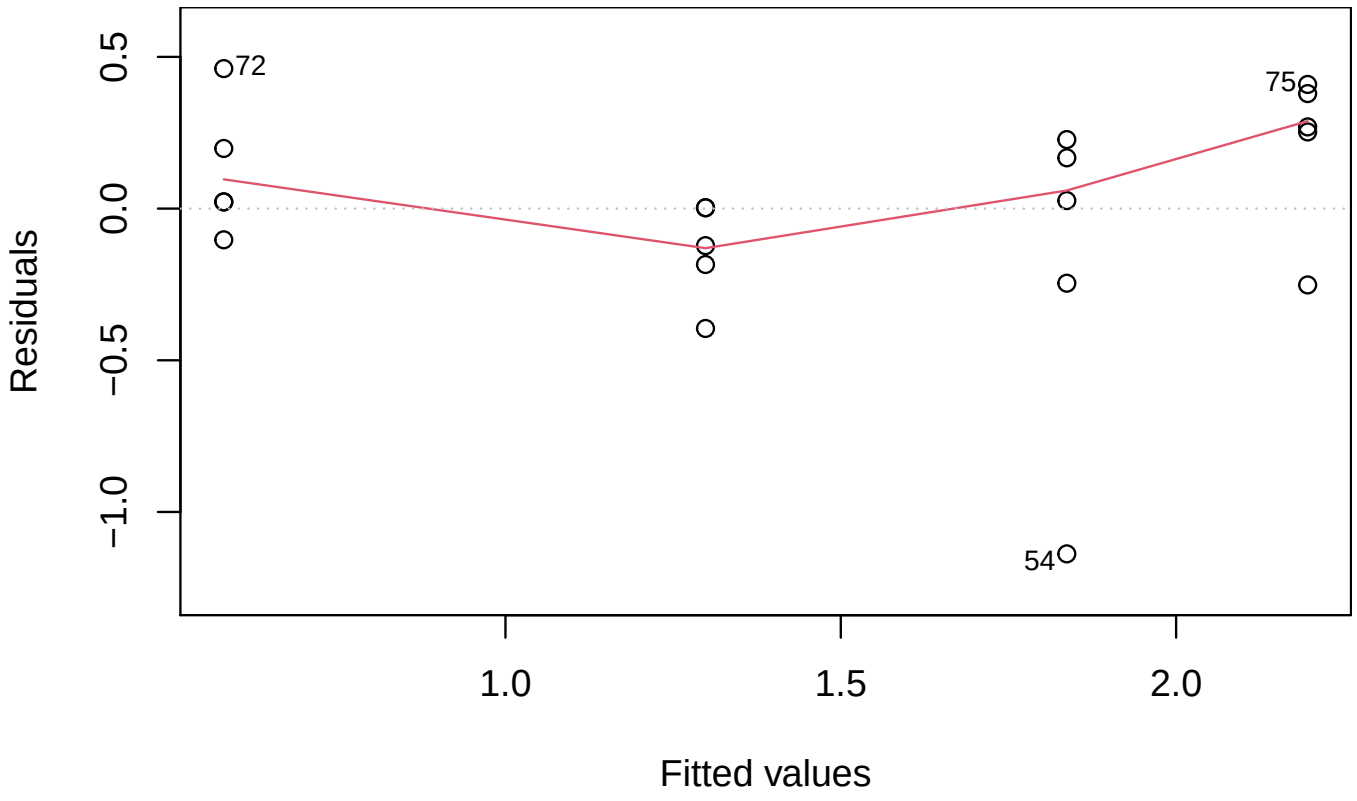
Trend in the residuals = one outlier

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



Example 3 - An R variant identifying outliers

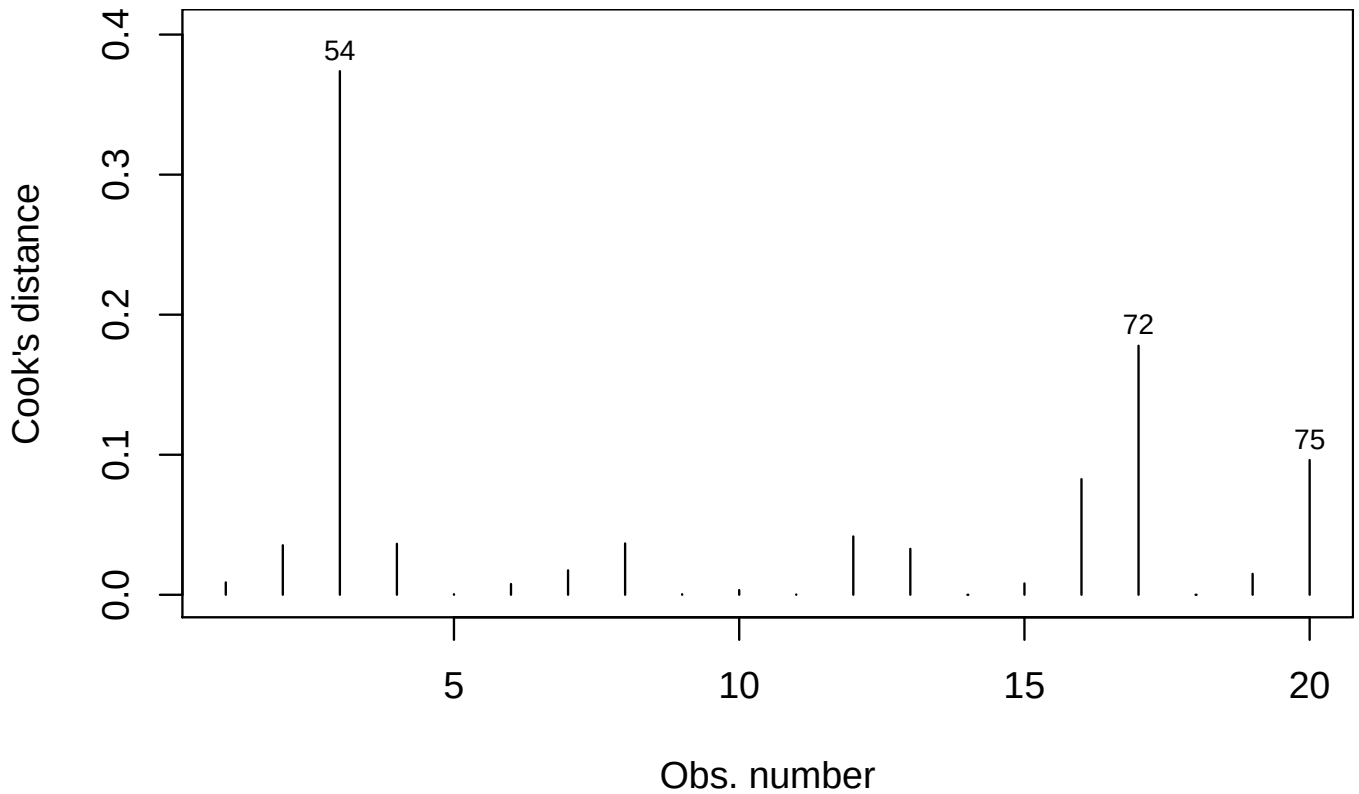
```
par(mar = c(4, 4, 0, 0))
plot(moutlier, which = 1)
```



Example 3 - Cook's distances: impact of outliers ?

To identify **influential observations**: impact of the removing of each observation on the parameter estimates.

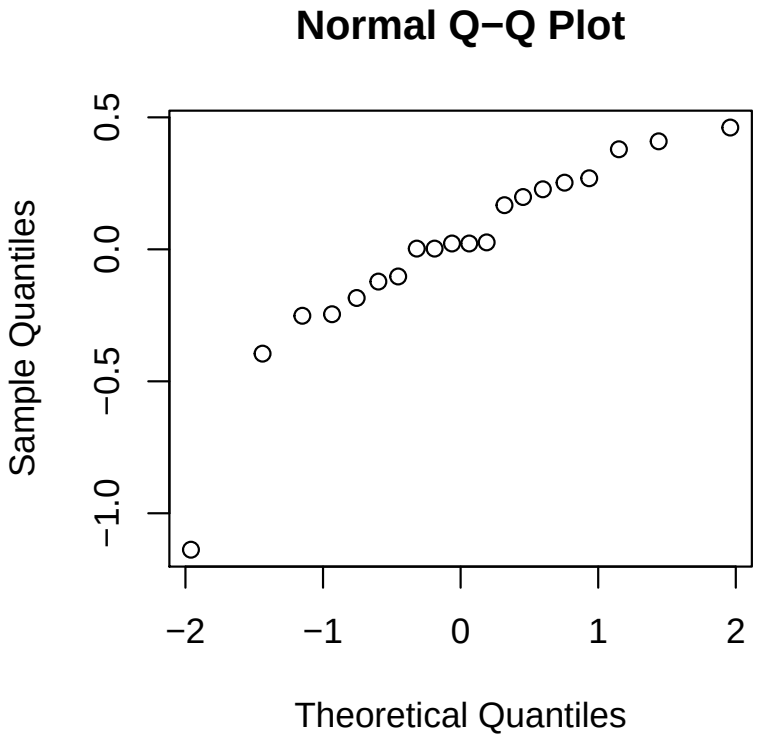
```
par(mar = c(4, 4, 0, 0))
plot(moutlier, which = 4)
```



Example 3 - Q-Q plot of residuals

In this example the problem is also detectable on the Q-Q plot.

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



Inference using simple linear regression

Inference using simple linear regression

Possible if your model is not invalidated by plots of residuals.

Let us go back to our example respecting the linear regression conditions

Interpretation of the estimates with their confidence intervals

```
coef(m)
```

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

```
confint(m)
```

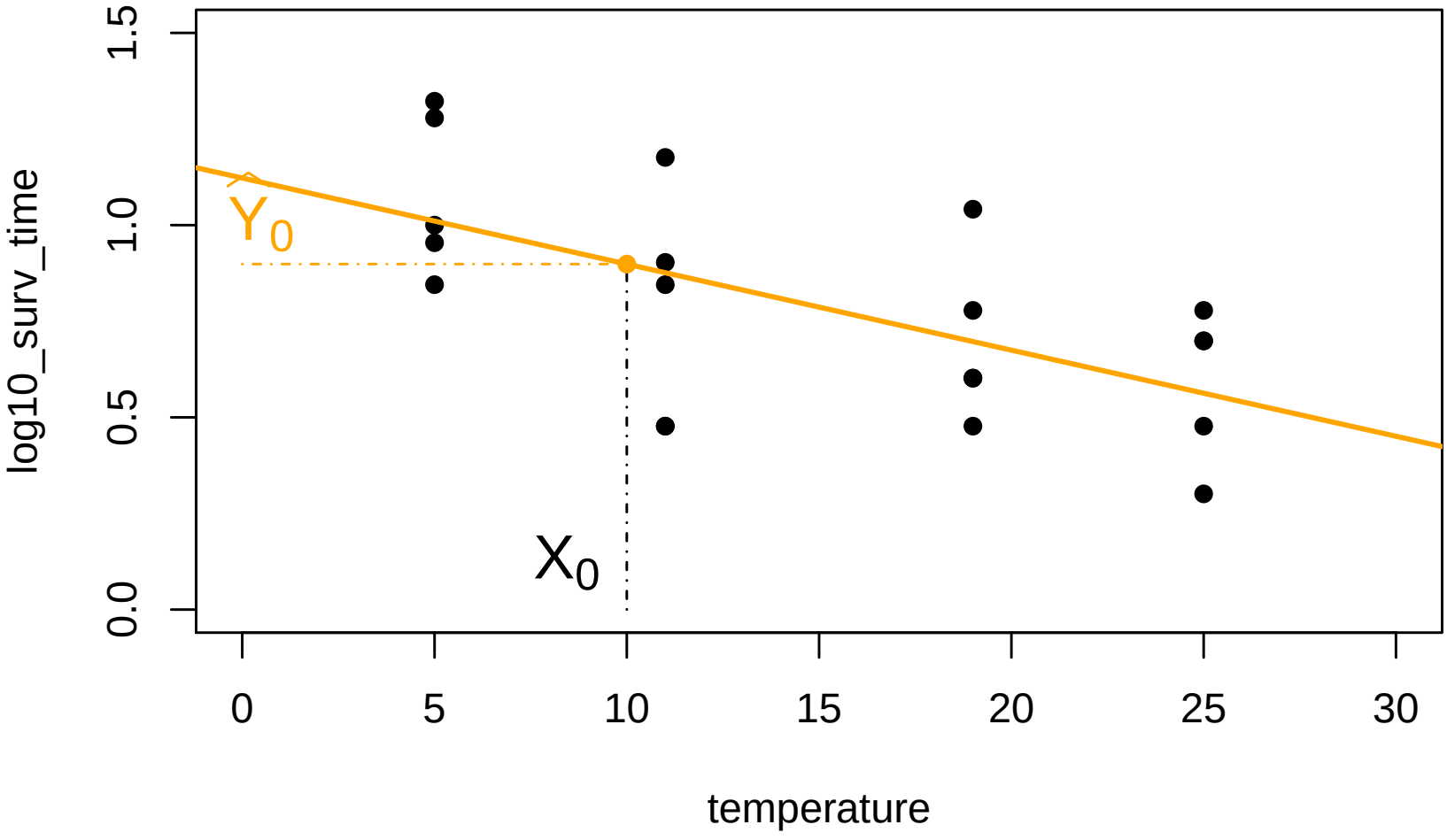
```
##                2.5 %   97.5 %
## (Intercept)  0.8828  1.36228
## temperature -0.0366 -0.00814
```

- **intercept**: estimated value of Y for $X = 0$ (meaningful if $X = 0$ has a biological meaning and stands in the range of observed values)
- **slope** (**regression coefficient**): **change in the dependent variable** corresponding to **a unit change in the independent variable**

Prediction using the model

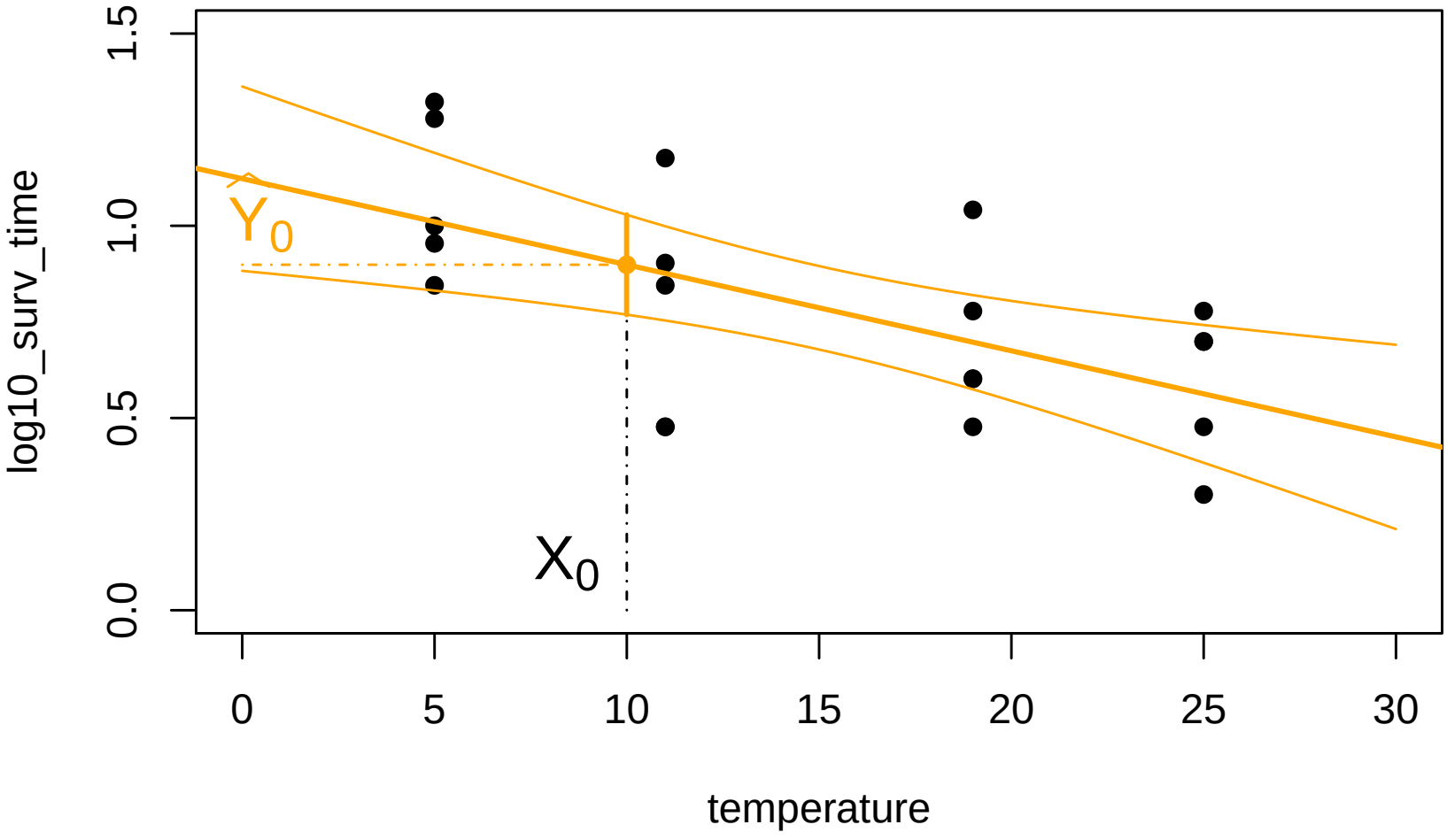
Prediction of Y_0 pour $X = X_0$ within the observed area.

BE CAREFUL: no extrapolation !



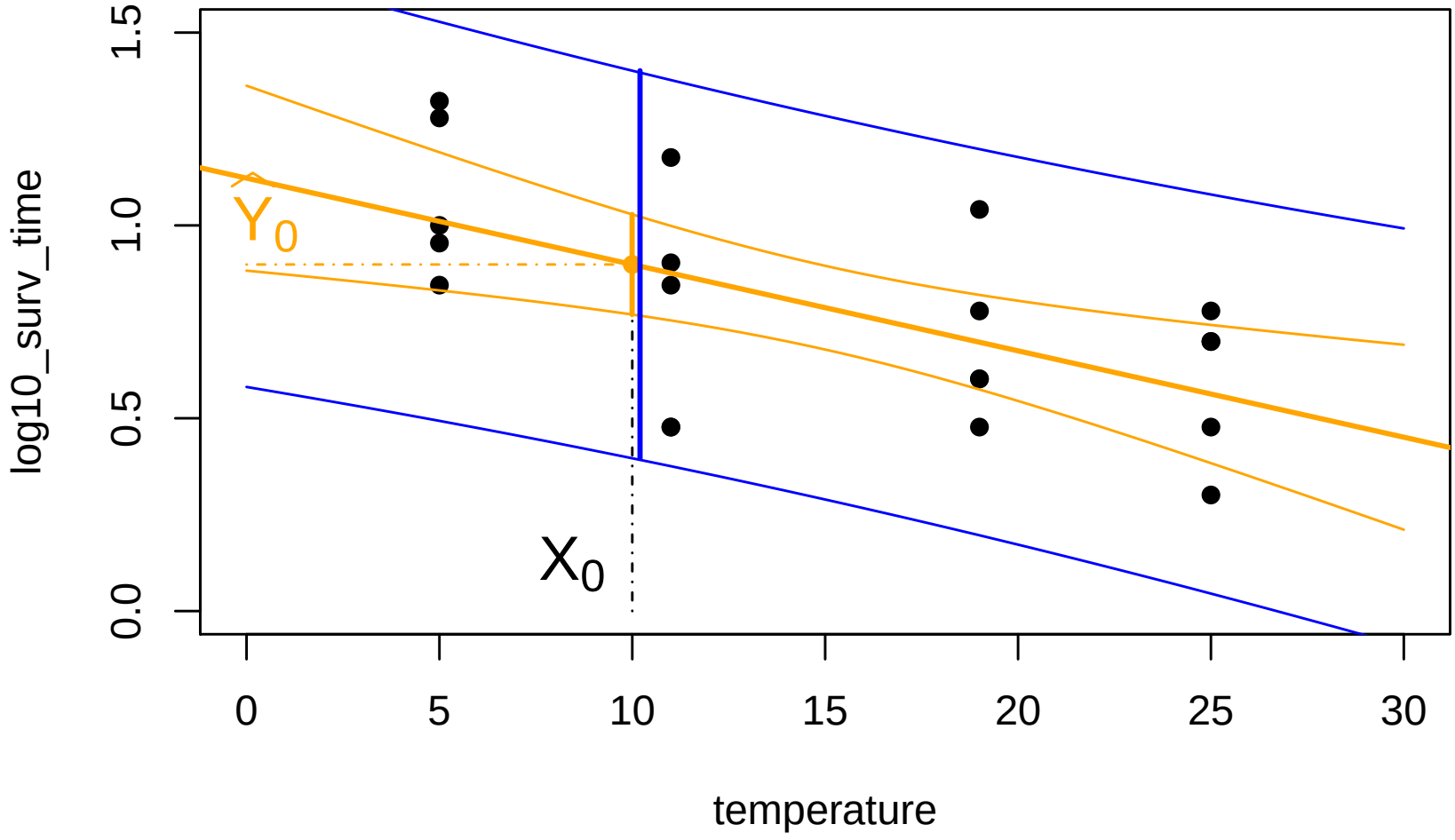
Confidence interval on the mean predicted value

Uncertainty on the deterministic part of the model (line)



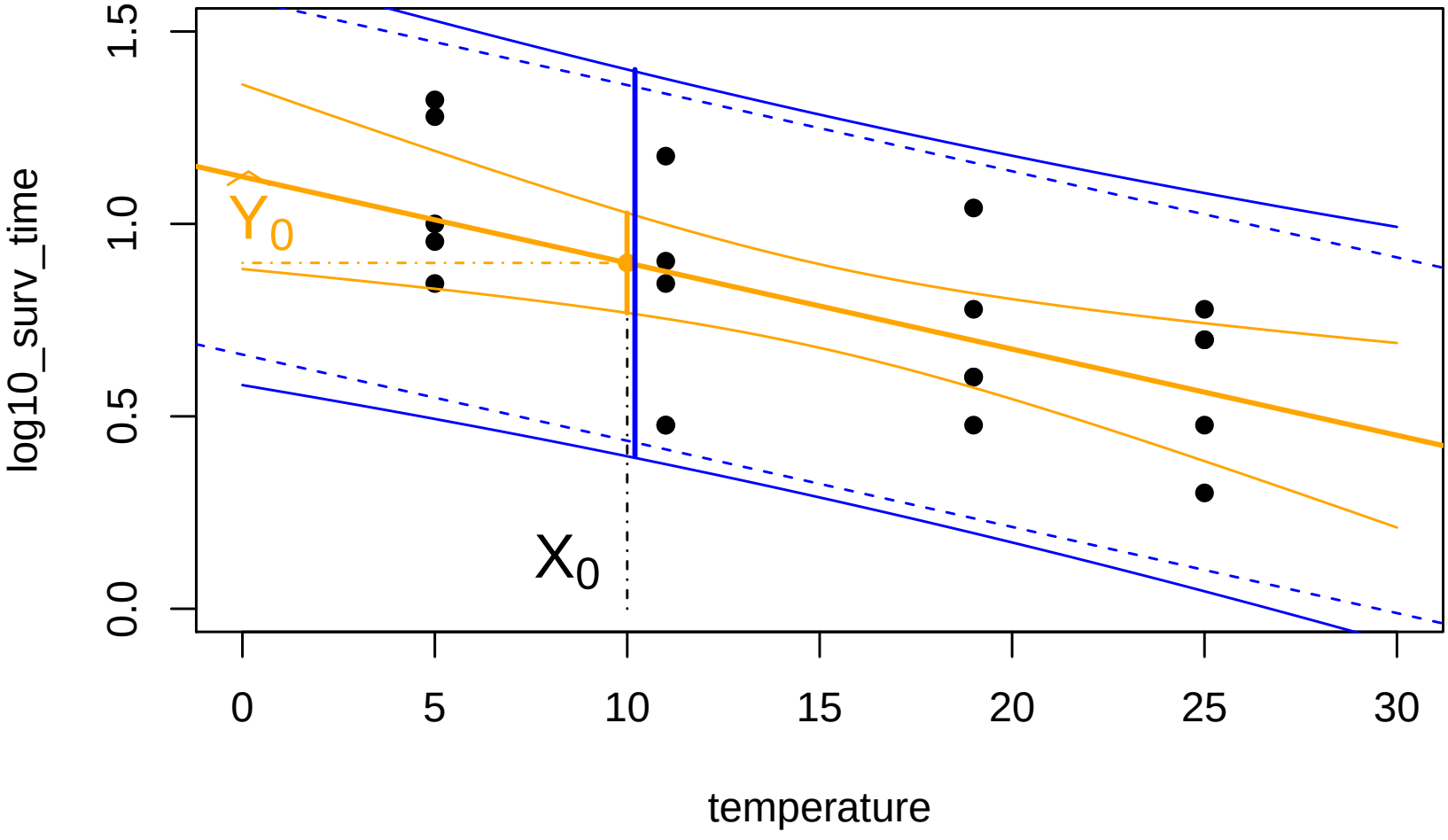
Prediction interval = confidence interval on an indiviual prediction

Uncertainty on the deterministic part of the model + stochastic part



Prediction interval = confidence interval on an indiviual prediction

Uncertainty on the deterministic part of the model + stochastic part often approximated by $\hat{Y}_0 \pm 2 \times \sigma$



Calculation of confidence and prediction intervals using R

Confidence interval

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

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

Prediction interval

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

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

Goodness-of-fit measures

Interpretation of the R-squared (r^2, coefficient of determination)

It corresponds to the **proportion of the variance in the dependent variable that the independent variable explains** (by the linear deterministic relation).

r^2 given in % is sometimes named **the percent of variance accounted for**.

The standard deviation σ is also a goodness-of-fit measure, but that must be interpreted with regard to the order of magnitude of Y.

R summary for the fit a linear model

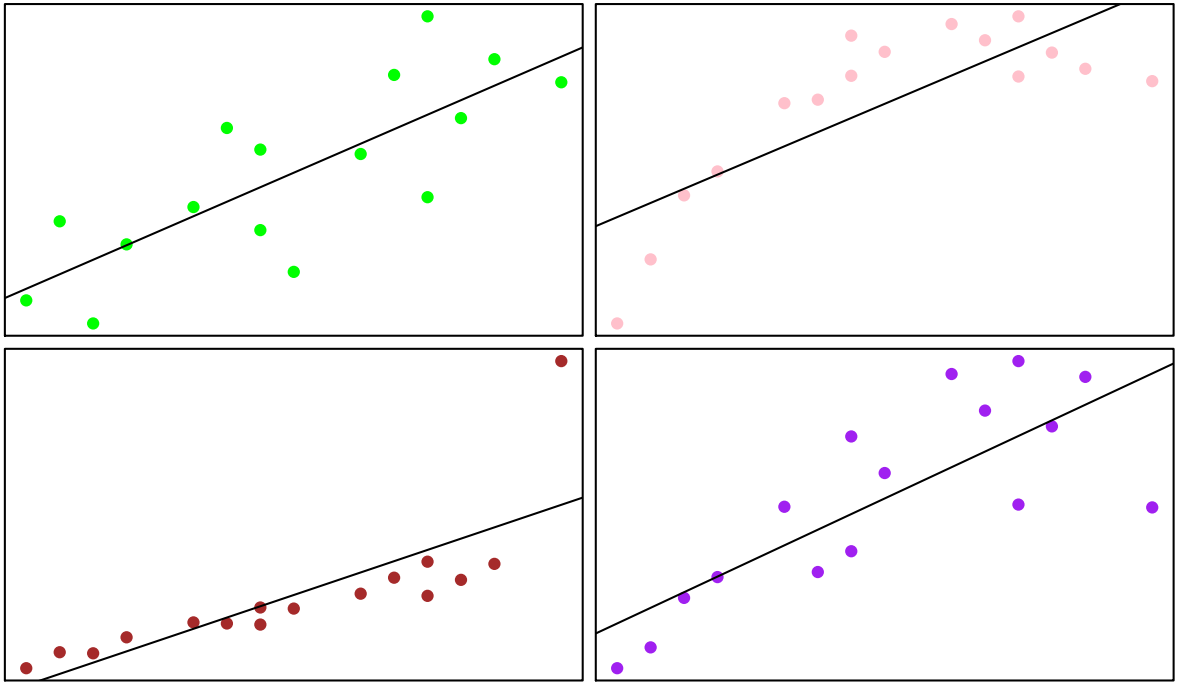
```
summary(m)
```

```
##
## 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 ' '
##
## 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
```

A value of r^2 close to 1 does not inform you about compliance with the use conditions of the model.

To convince you look at the four following examples sharing exactely the same r^2 value of 62%

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



Your turn to use this model for inference purposes

Using the same model (`m`), answer the following questions:

1. At a relative humidity of 50%, what is the expected **change of the survival rate (in log_{10})** due to an **increase of the temperature of** $1°C$.
2. Traduce this change in a **mutiplicative coefficient on the survival time** and in its **relative diminution** due to the same increase of temperature of $1°C$.
3. Give a prediction (with its 95% confidence interval) of the **mean survival time in log scale** of ticks at a relative humidity of 50% and a **temperature of** $22°C$ and its traduction in the raw scale (in days).
4. Give a prediction (with its 95% confidence interval) of **the survival time of one tick** exposed at a relative humidity of 50% and a **temperature of** $22°C$.
5. Give a prediction (with its 95% confidence interval) of **the survival time of one tick** exposed at a relative humidity of 50% and a **temperature of** $40°C$.

Answer to question 1

At a relative humidity of 50%, what is the expected **change of the survival rate (in log_{10})** due to an **increase of the temperature of** $1°C$.

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

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

Answer to question 2

Traduce this change in a **mutiplicative coefficient on the survival time** and in its **relative diminution** due to the same increase of temperature of $1°C$.

If we name st_0 the initial survival time, and st_c the survival time after the change, we expect $log_{10}(st_c) - log_{10}(st_0) = b$ with b the regression coefficient, so $log_{10}(\frac{st_c}{st_0}) = b$ so $st_c = 10^b \times st_0$.

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

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

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

Answer to question 3

Give a prediction (with its 95% confidence interval) of the **mean survival time in log scale** of ticks at a relative humidity of 50% and a **temperature of** $22°C$ and its traduction in the raw scale (in days).

```
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
```

Answer to question 4

Give a prediction (with its 95% confidence interval) of **the survival time of one tick** exposed at a relative humidity of 50% and a **temperature of** $22°C$.

```
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
```

Answer to question 5

Give a prediction (with its 95% confidence interval) of **the survival time of one tick** exposed at a relative humidity of 50% and a **temperature of** $40°C$.

```
# 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 !!!!!!!!!!!!!!
```

NO EXTRAPOLATION !

This data set does not inform you about what happens for temperature above $25°C$!

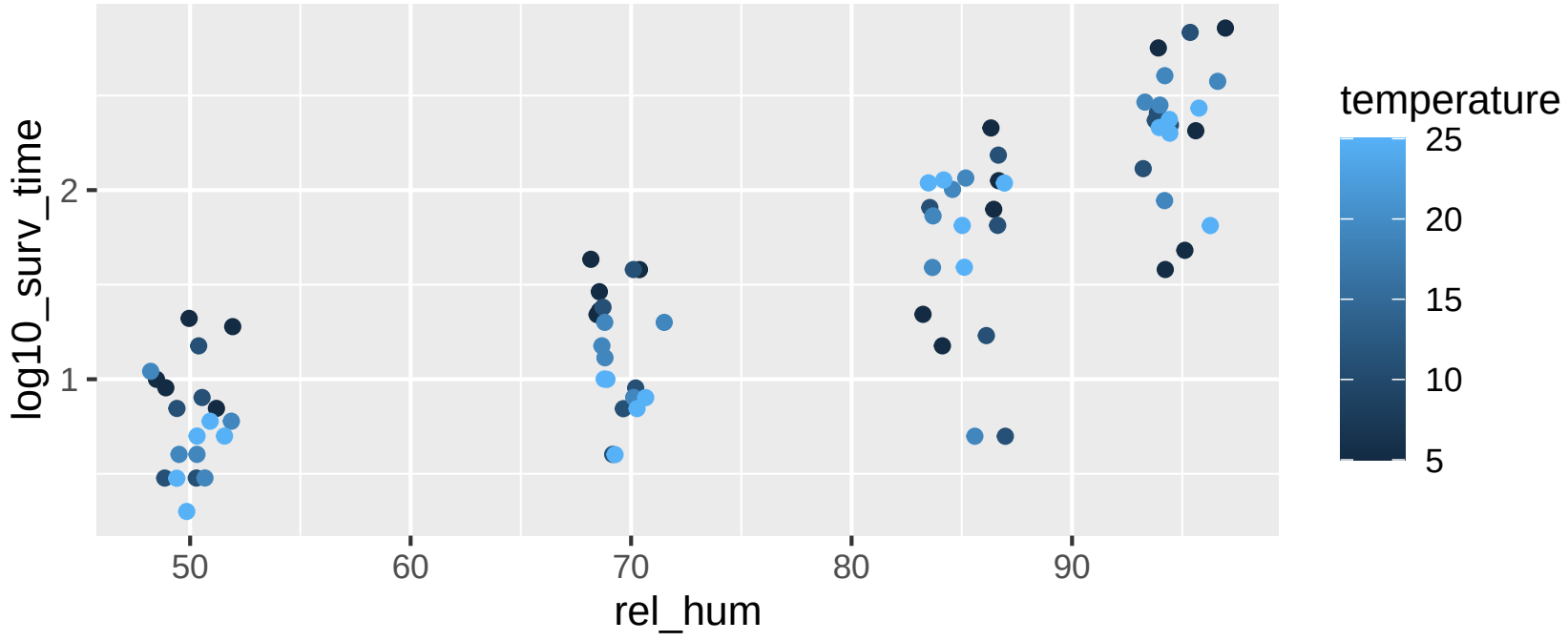
Multiple linear regression

Basic concepts and fitting method

Modeling of the impact of both relative humidity and temperature in non dried conditions

using a subset of the **whole data set** at all humidity conditions **except the driest condition**.

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



The theoretical model

Very similar to the simple linear model,

with **more than one continuous independent variable** (called regressors),

so **more than one regression coefficient** (no more called slope).

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

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

Deterministic part: linear link
Stochastic part : Gaussian model

assuming **random**, **independent** residuals ϵ_i following a **Gaussian** (normal) distribution of constant variance σ^2.

The least squares estimation of parameters

As in the case of the simple linear model,

the **maximum likelihood estimation**

(maximizing $Pr(Y|\beta_0, \beta_1, \ldots, \beta_p, \sigma)$) still

corresponds to the **least squares estimation** minimizing $SCE = \sum_{i=1}^{n} e_i^2$ with $e_i = Y_i - \hat{Y}_i$

Estimation of parameters using 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
```

R summary for the fit of a multiple linear model

```
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
```

How to interpret the p-values associated to each regression coefficient ?

```
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
```

Each p-value corresponds to the **significance test of each regression coefficient** with H_0 the null hypothesis of each coefficient, the other ones being kept in the model. So it answers to the question: "**is there a significant linear relationship between** the regressor X_i **and** the outcome Y **when the other regressors** X_j with $j \neq i$ **have already been taken into account**?"

In this example we see an significant impact of the relative humidity on the survival rate, but the add of the temperature as a second regressor does not significantly improve the model.

Another equivalent way to compare two nested models using an F test

```
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 ' '
```

The add of the temperature in the model does not significantly improve the fit.

Another way to compare two nested models using an F test - impact of the relative humidity

```
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 ' '
```

The add of the relative humidity in the model significantly improves the fit.

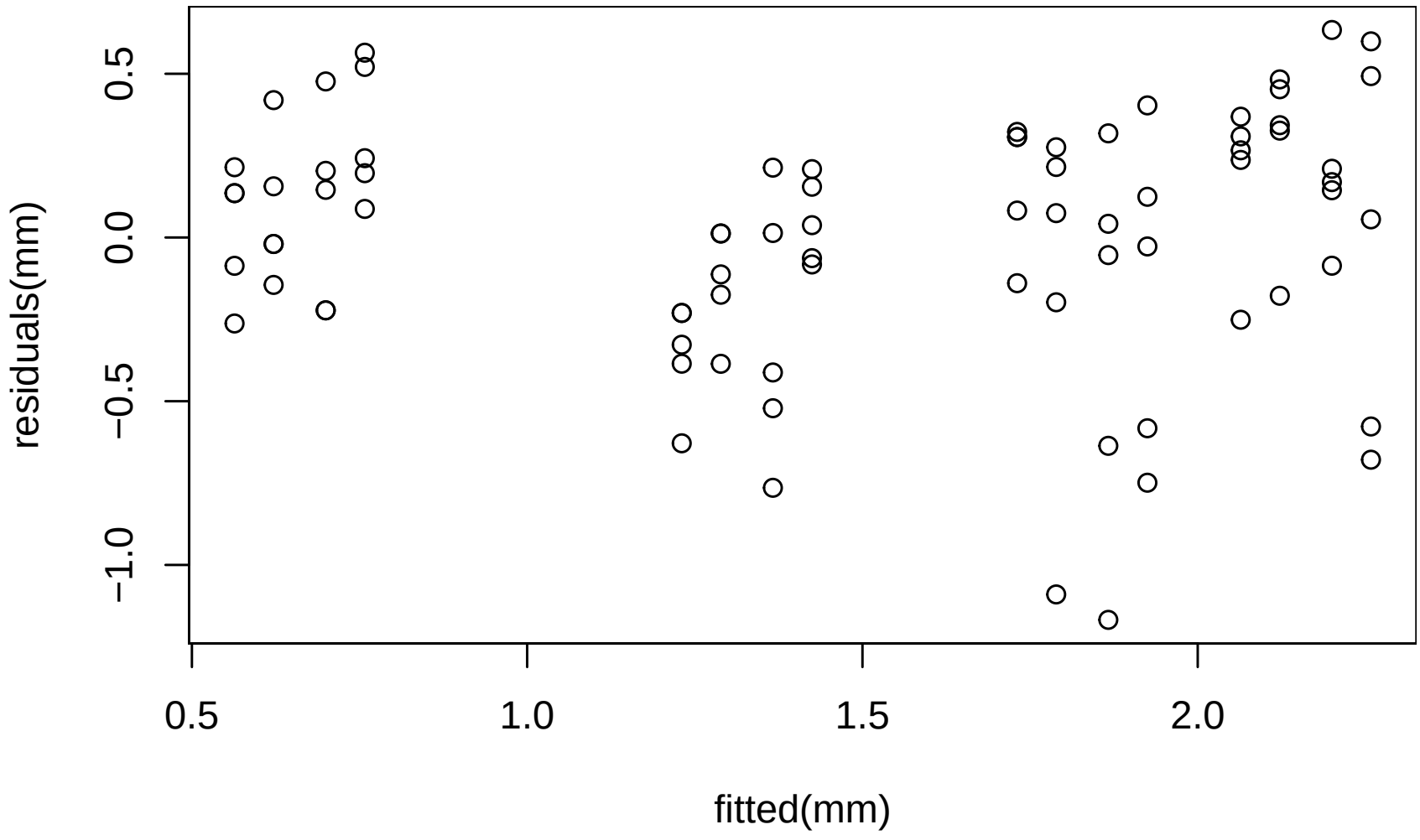
No more possible to simply look at the fit of the model

It is one of the difficulties encountered in the process of check of the adequation of the model to the data !

Check of use conditions

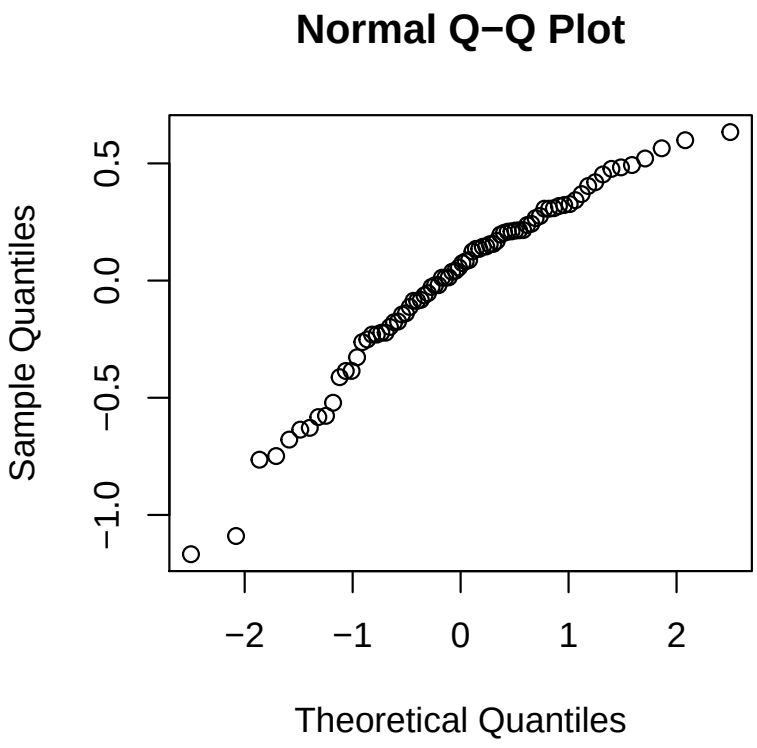
Plot of residuals against fitted values as in simple linear regression

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



Quantile-quantile plot of residuals as in simple linear regression

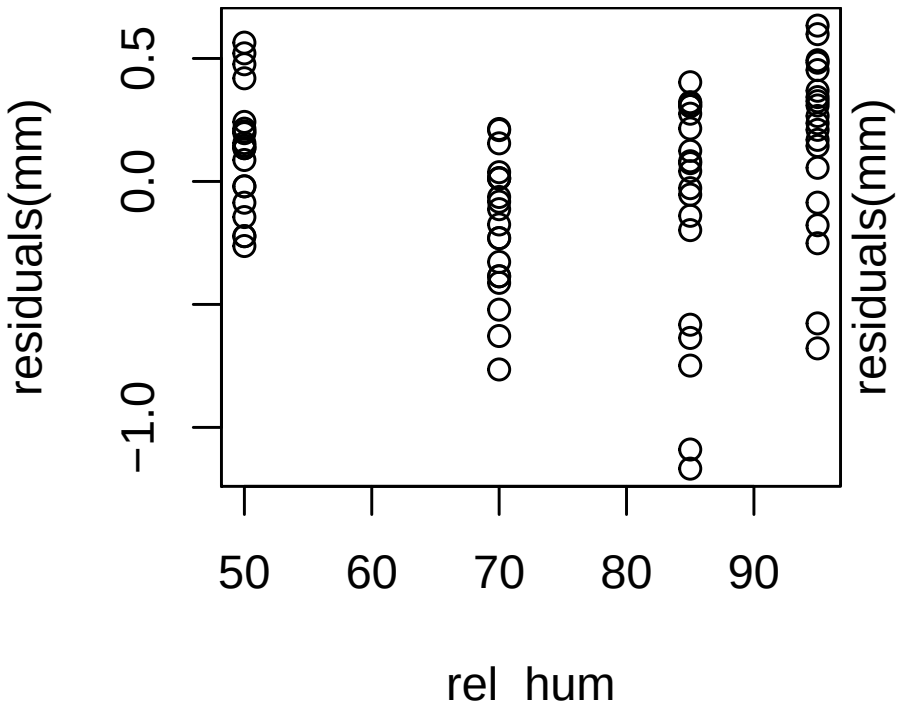
```
qqnorm(residuals(mm))
```

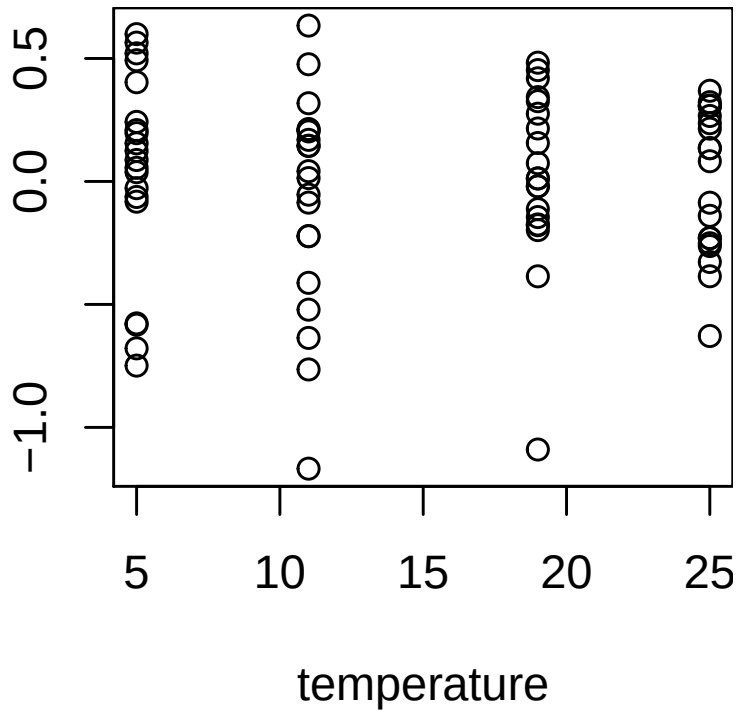


Add a plot of residuals against each independent variable

Especially useful to detect violation of the hypothesis of linear relationships (the case for the relative humidity in this example !)

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

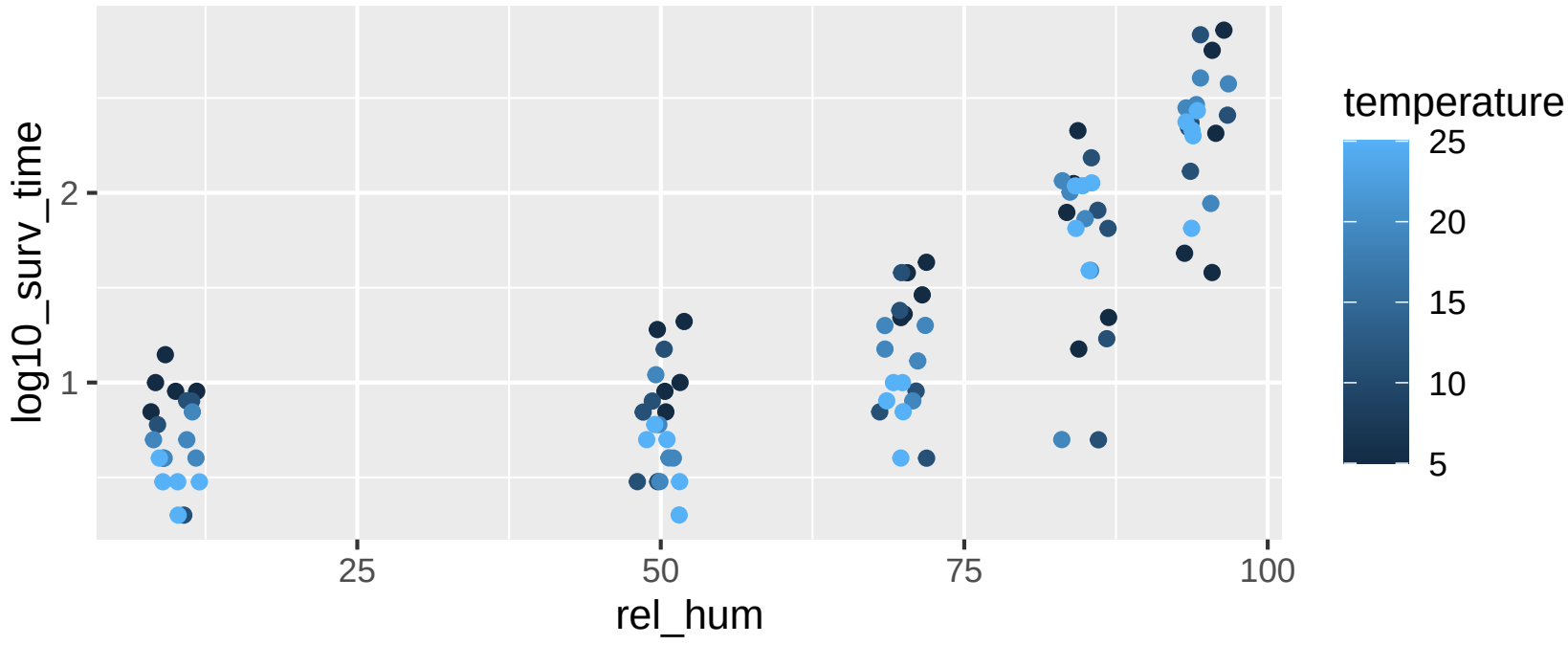




Go back to the data

We could have anticipated this problem, and it would have been worse if the data at 10% have been kept in the data set.

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



Polynomial models: a solution ?

One simple way to **take into account a non linearity in the relationship** between the dependent variable and one (or more) regressor is to use polynomial models.

```
(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
```

This polynomial model can be fitted by least squares: it is a linear model with just one more independent variable (the square of the relative humidity).

Your turn to handle multiple regression

Using the same data set `dhum` that excludes the driest condition:

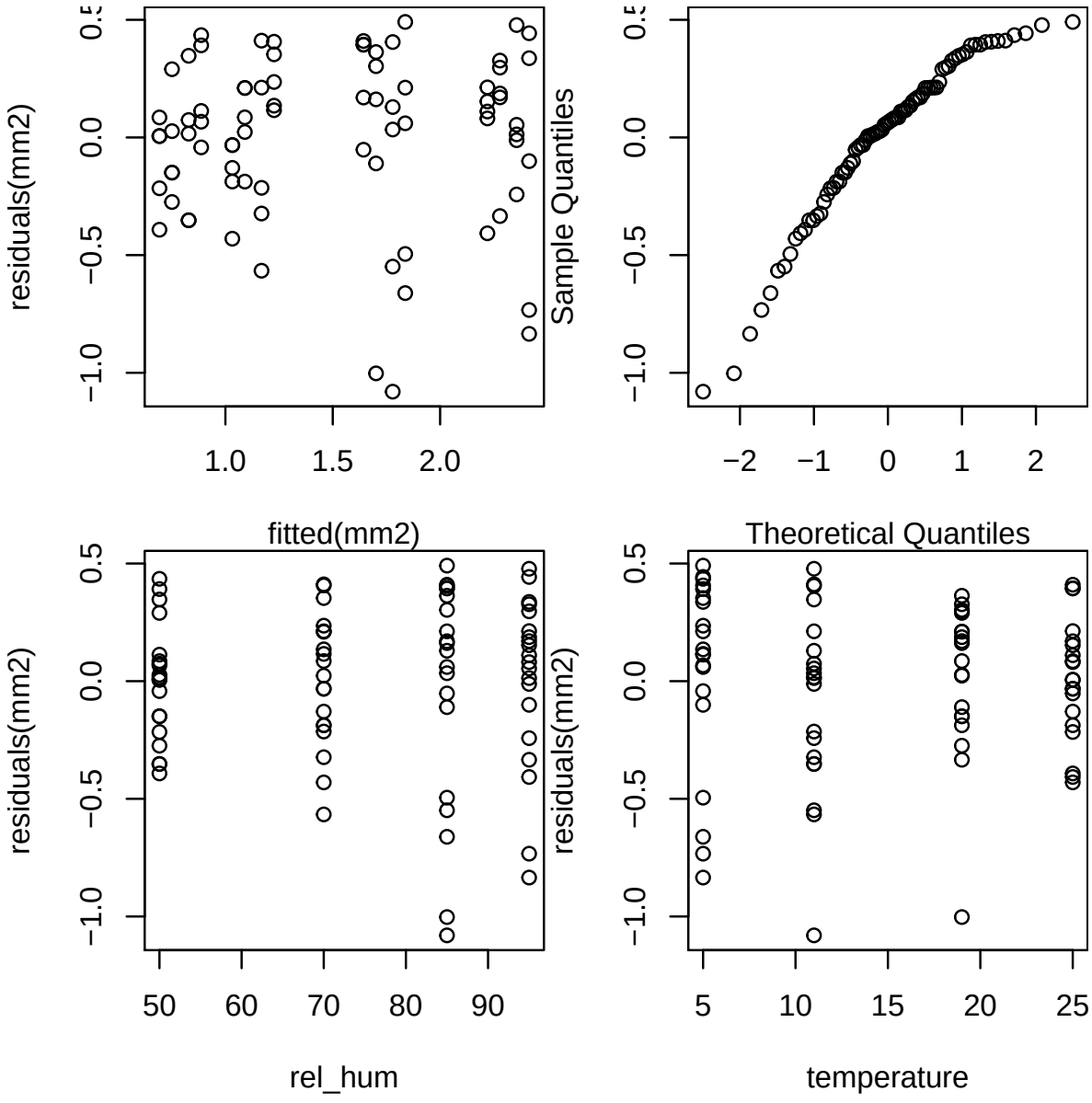
1. Look at the **residual plots** obtained with this new model. Is the **problem due to non linearity solved** ?
2. Look at the summary of the new model `mm2rel_hum` and try to **interpret the p-values**.

Using the whole data set `dtot` (with all conditions):

3. Fit a polynomial model of second order **without taking into account the impact of the temperature**.
4. Look at the summary and at the residuals.
5. Represent the **fitted model on the data** and **question the biological relevance** of this model. For that question you should define a new data frame with evenly spaced values within the range of tested conditions of relative humidity and use the `predict()` function directly on this new data set.

Answer to question 1

No more trend on the residual plot against the relative humidity.



Answer to question 2

Significant improvement of the fit adding the square relative humidity, but no significant improvement adding the temperature.

```
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 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Answer to question 3

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

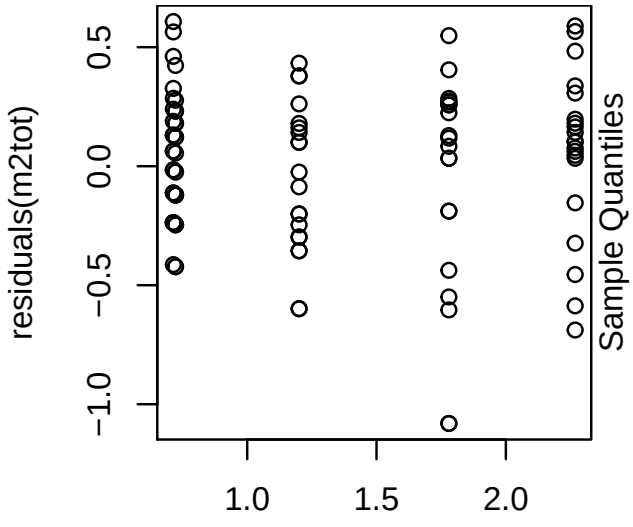
```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + I(rel_hum^2), data = dtot)
##
## Coefficients:
##  (Intercept)       rel_hum  I(rel_hum^2)
##     0.929364     -0.024727      0.000409
```

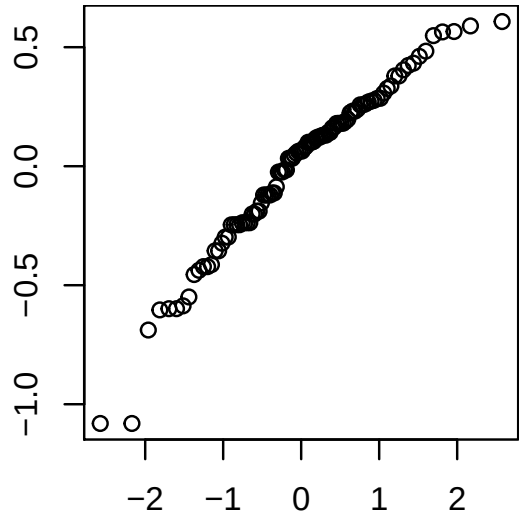
Answer to question 4 - summary

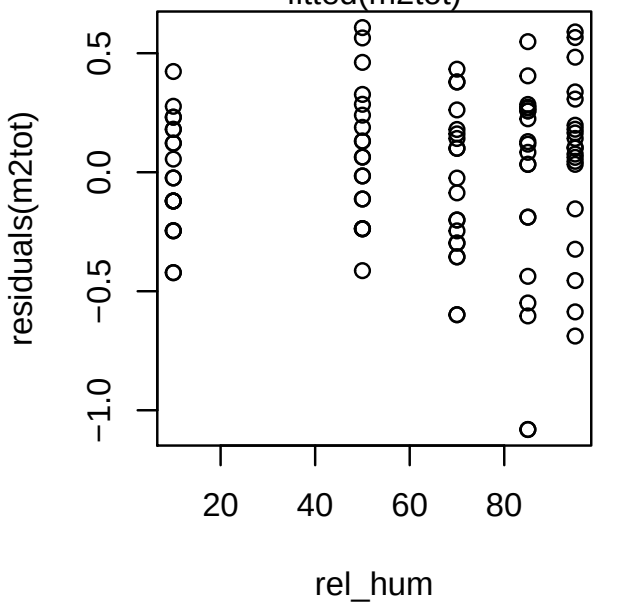
```
summary(m2tot)
```

```
##
## 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
```

Answer to question 4 - residuals





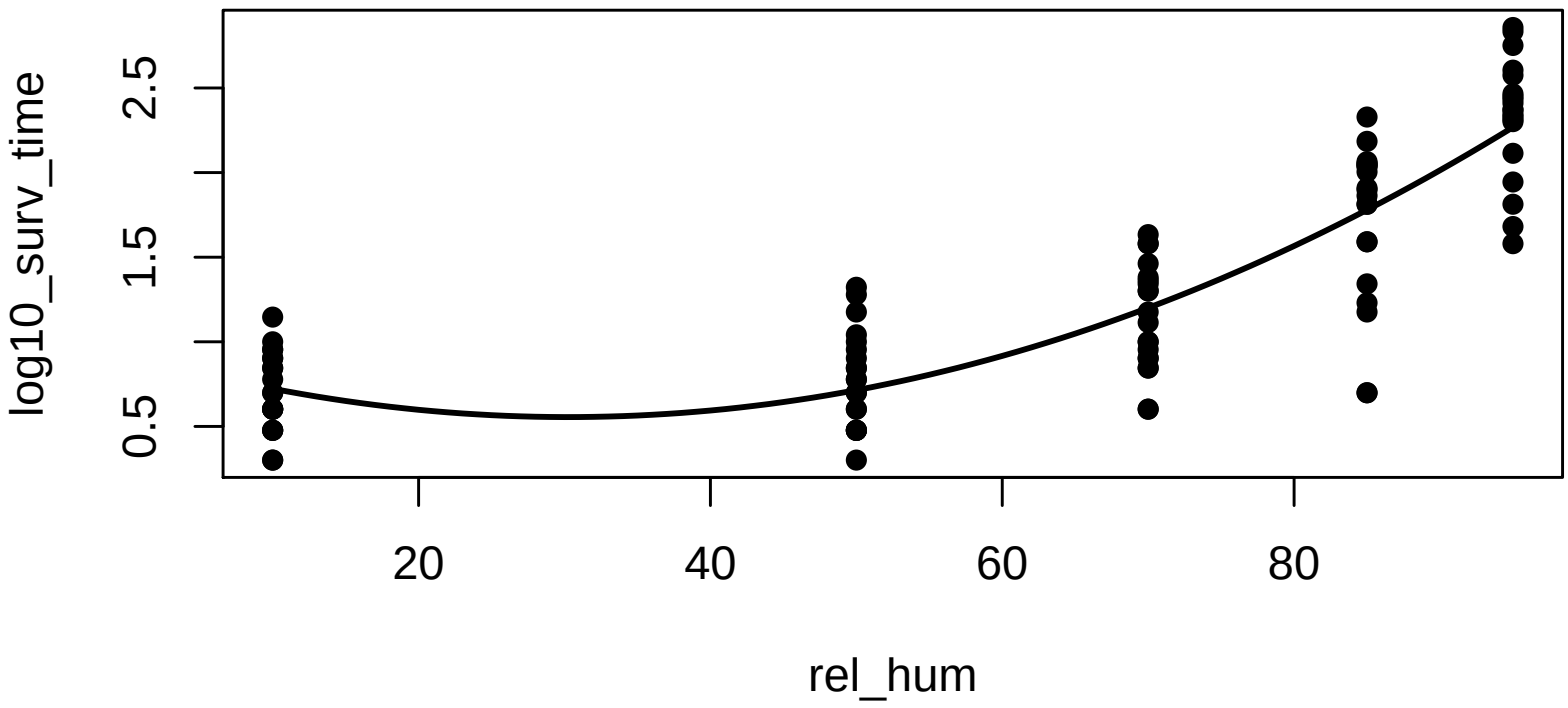


Answer to question 5 - the R code

```r
plot(log10_surv_time ~ rel_hum, data = dtot, pch = 16)
data4pred <- data.frame(rel_hum = seq(10, 95, length.out = 50))
pred <- predict(mm2tot, newdata = data4pred)
lines(pred ~ data4pred$rel_hum, lwd = 2)
```

Answer to question 5 - the plot

Is a minimum of the survival time between 20 and 40 % of humidity expected ?



The case of a qualitative independent variables

The ANOVA 1 linear model

The ANOVA 1 linear model

Some of the **independent variables may be qualitative** (e.g. the sex) or may be **transformed in a qualitative variable (often named a factor)** to cope with violation of the linearity condition.

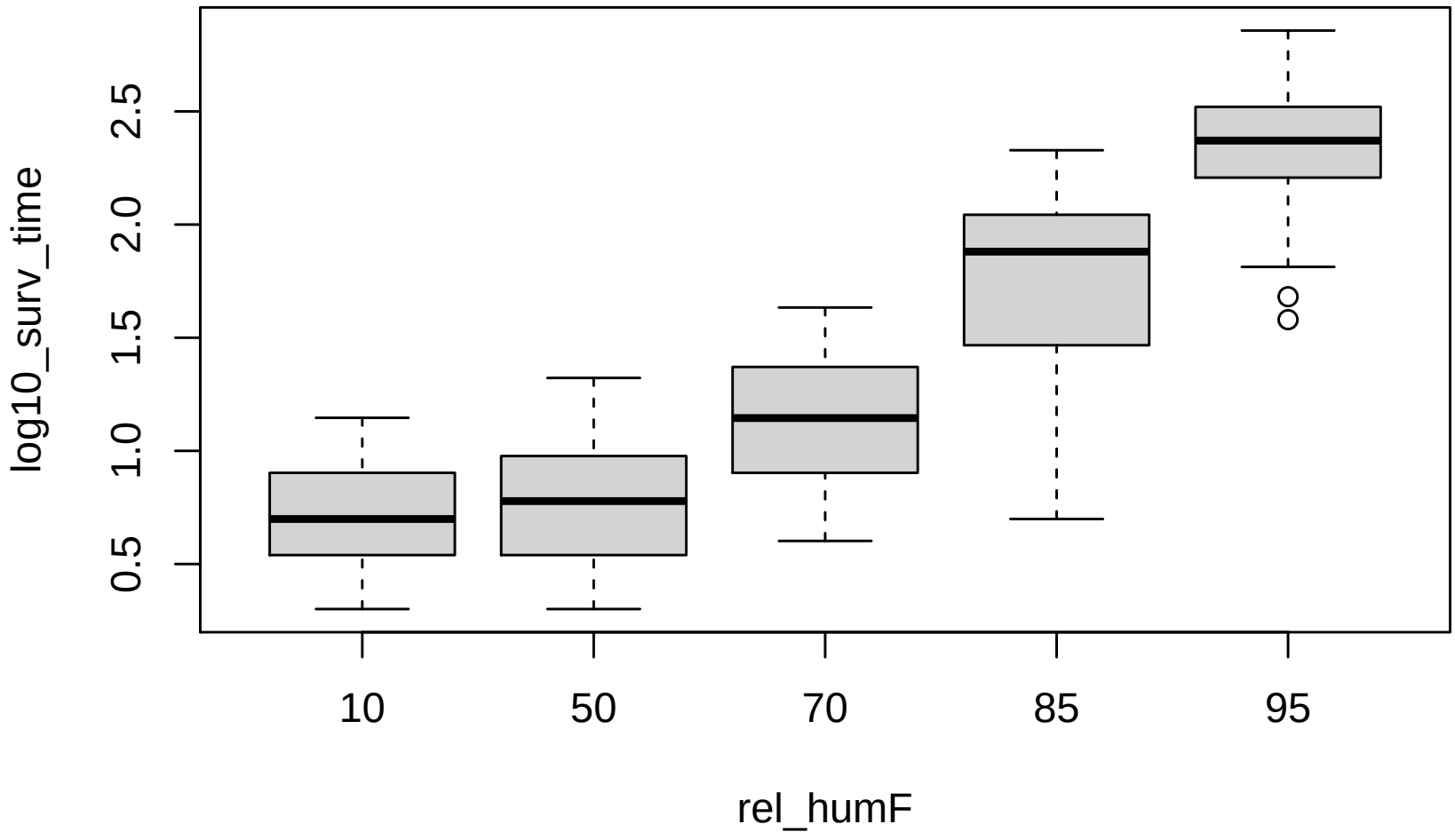
Imagine we want to model the **impact of the relative humidity** on the survival time (in log_{10}) by considering it **as a qualitative variable with 5 conditions**, neglecting the potential impact of the temperature (for the moment).

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

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

Plot of data using boxplots

```
par(mar = c(4, 4, 1, 1))
plot(log10_surv_time ~ rel_humF, data = dtot)
```



Formalization of the ANOVA 1 linear model

- **Classical formalization**: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ with $\epsilon_{ij} \sim N(0, \sigma)$ with $\sum \alpha_i = 0$
- Formalization using $p - 1$ **dummy variables** X_1 to X_{p-1} coding for the **membership of each observation to the** p **groups except the reference one**:
  $Y_k = \beta_0 + \beta_1 X_{1,k} + \cdots + \beta_{p-1} X_{p-1,k} + \epsilon_k$ with $\epsilon_k \sim N(0, \sigma)$
- Link between both formalizations:
  - mean of group 1 $= \mu + \alpha_1 = \beta_0$
  - mean of group 2 $= \mu + \alpha_2 = \beta_0 + \beta_1$
  - mean of group i $= \mu + \alpha_i = \beta_0 + \beta_{i-1}$

So each coefficient of the linear model corresponds to the **differences of the mean in each class to the mean in the reference class.**

Fit a linear model

```
(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
```

Performs the classical one-way analysis of variance

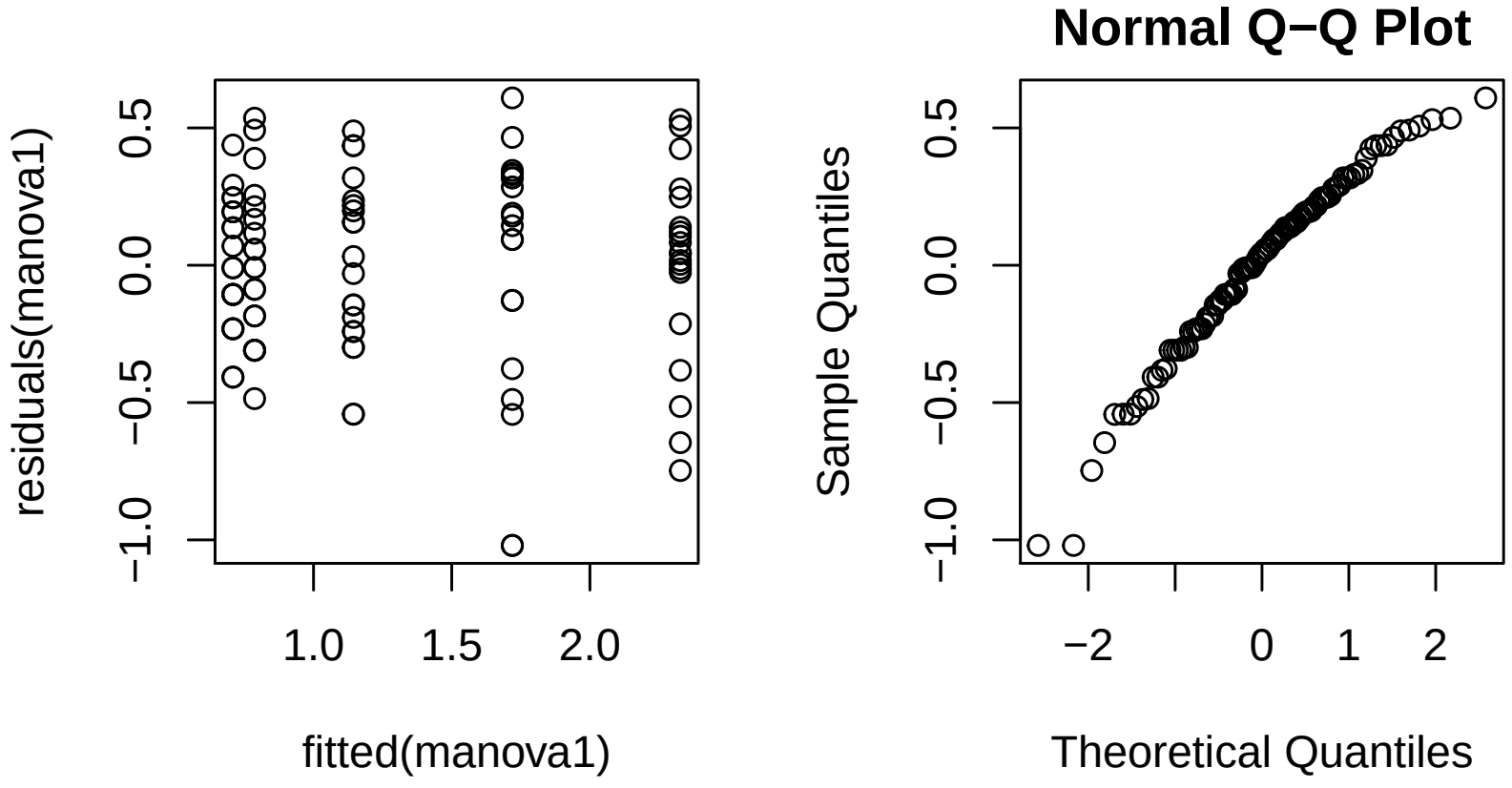
```
anova(manova1)
```

```
## 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 ' '
```

As expected in this example, it shows a significant effect of the relative humidity on the survival time (in log_{10})

Check the use conditions

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



Interpret the 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
```

```
# Coefficients with their 95% confidence intervals
coef(manova1)
```

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

Each one corresponds to the **difference of the mean in each class to the mean in the reference class** (first level of the factor, by default with alphabetic ordering of the levels)

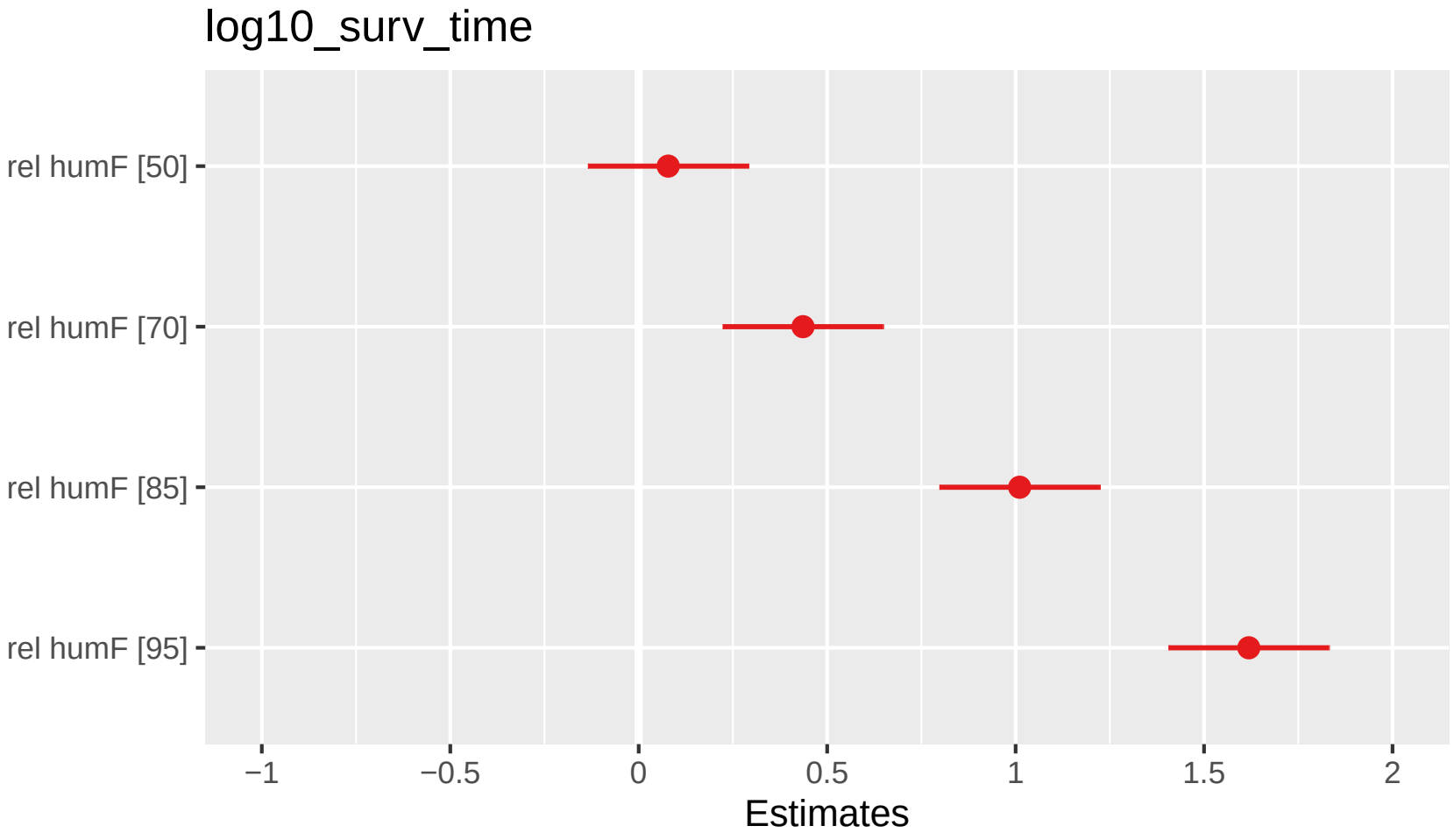
Interpret the confidence intervals of the coefficients

```
confint(manova1)
```

```
##                2.5 % 97.5 %
## (Intercept)    0.558  0.859
## rel_humF50    -0.135  0.291
## rel_humF70     0.223  0.649
## rel_humF85     0.798  1.224
## rel_humF95     1.406  1.832
```

Forest plot to vizualise the coefficients with their confidence intervals

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



The ANOVA 2 linear model

The ANOVA 2 linear model

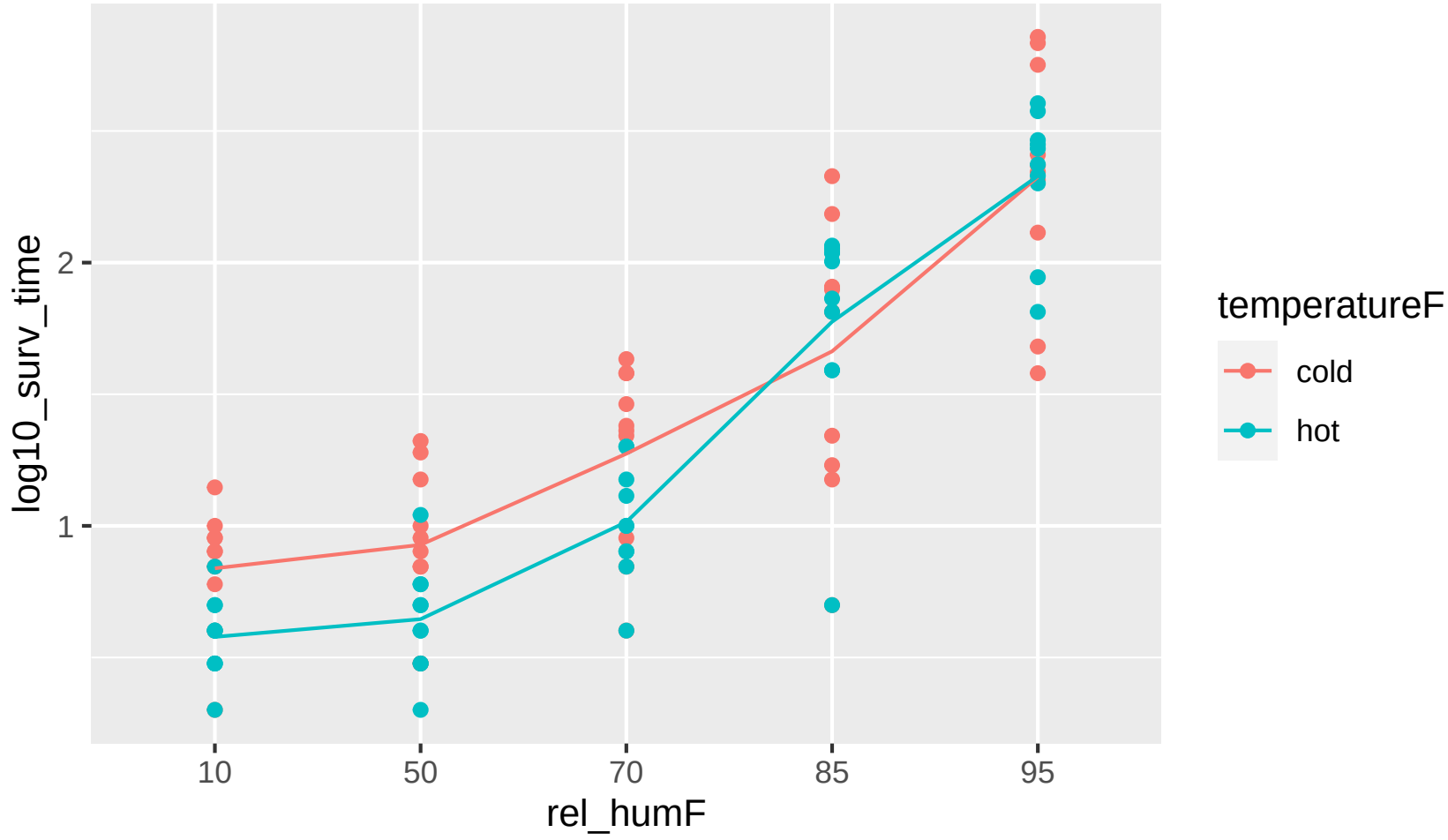
Now imagine we want to **add** in this model the impact of the **temperature**, also **transformed in a qualitative variable** with two modalities, cold ($< 15°C$) or hot.

```
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
```

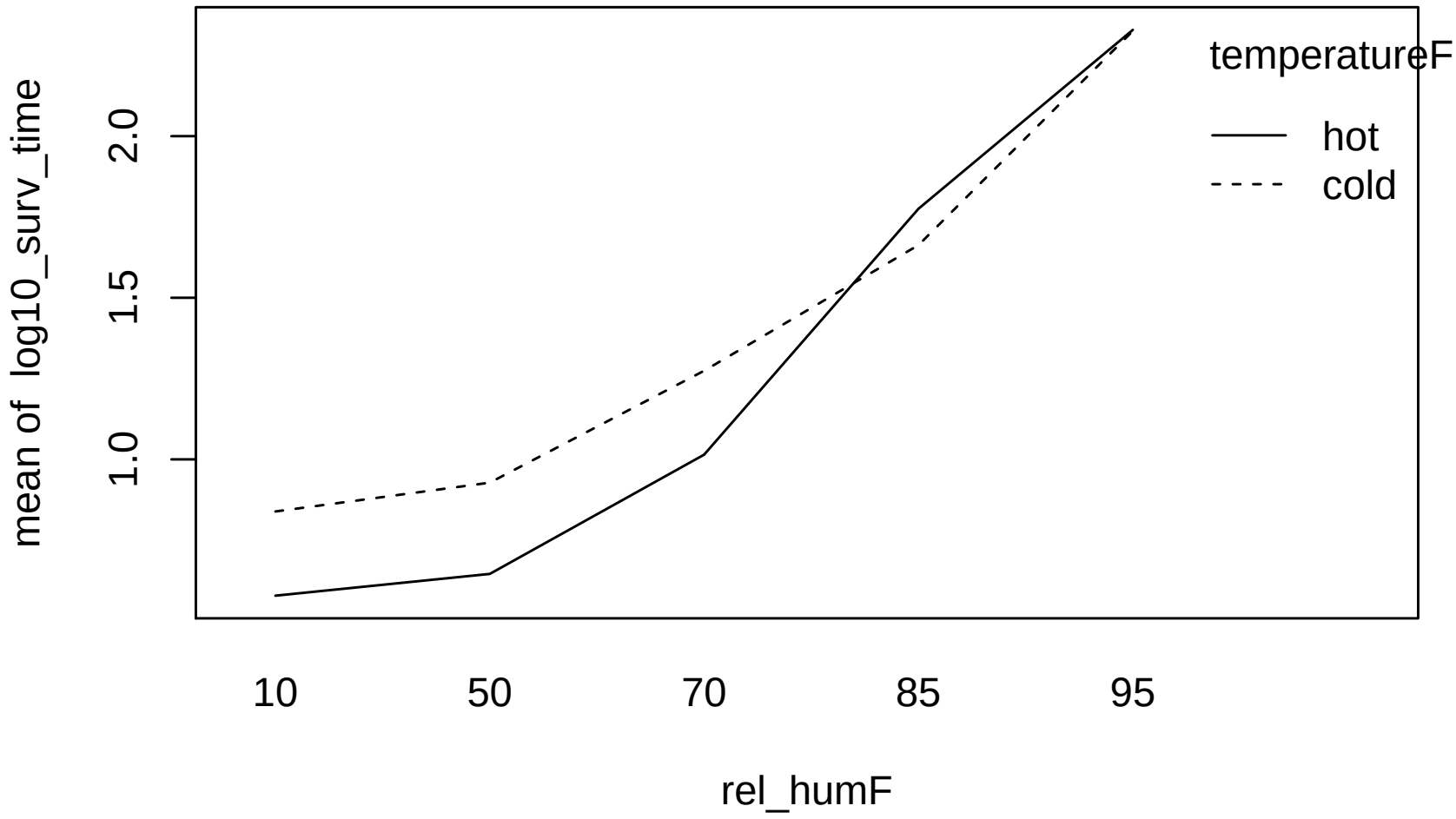
An interaction plot - using ggplot2

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



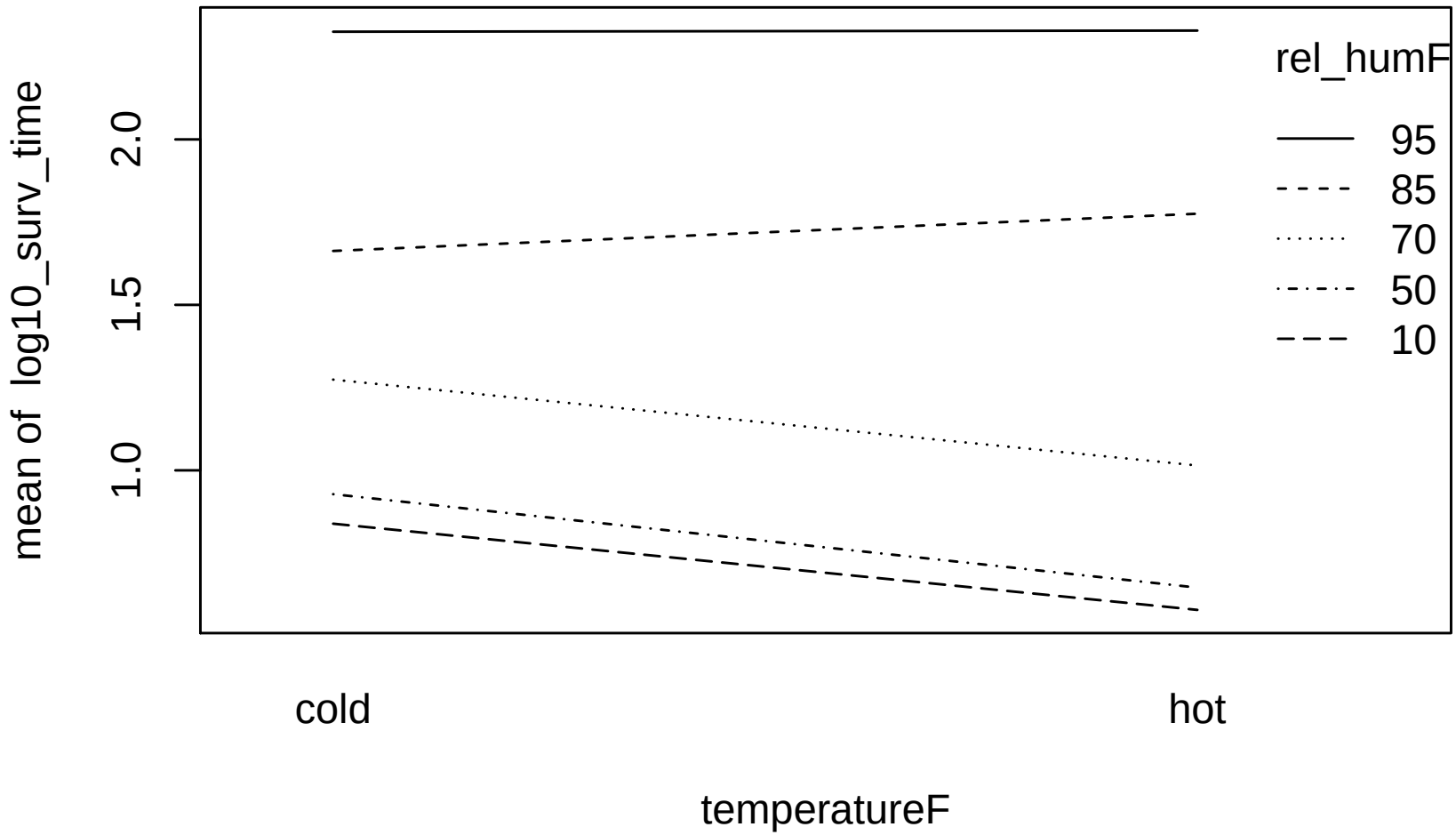
An interaction plot - using graphics

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



An second version of the interaction plot - using graphics

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



ANOVA 2 models without or with interaction between two factors A and B

Two factors A and B may each have an effect on the dependent variable **without interacting**. The **effects** are then said **additive**.

The effect of A does not depend on the modality of B, and vice versa.

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

Two factors may interact. The **model is no more additive**. It incorporates **interaction** terms γ_{ij}.

The effect of A depends on the modality of B, and vice versa.

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

The choice between both models must be guided by the **prior biological knowledge** and the **data observation**.

Fit the ANOVA 2 model without interaction

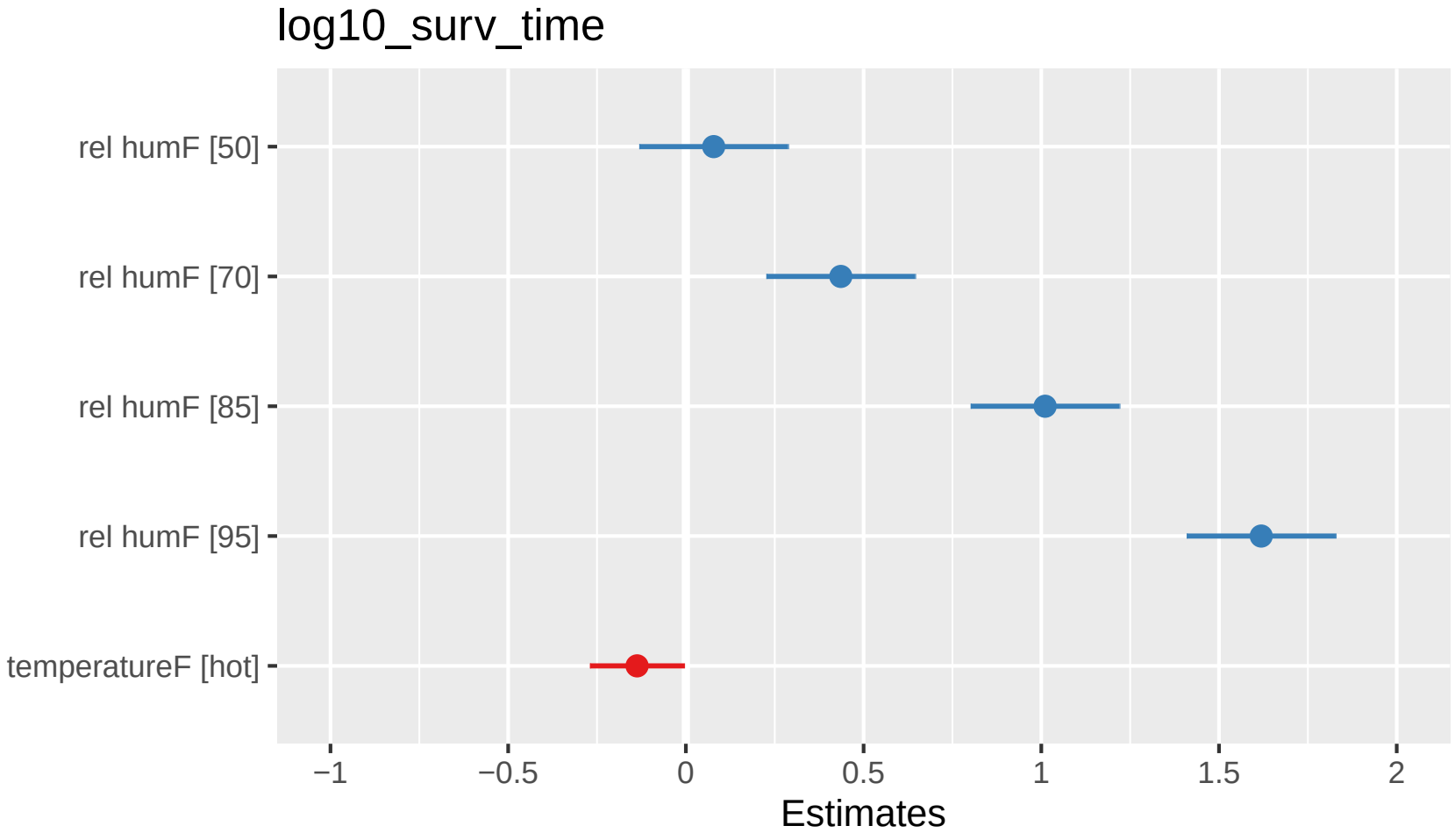
```
(manova2 <- lm(log10_surv_time ~ rel_humF + temperatureF,
               data = dtot))
```

```
##
## 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. of predicted mean at 70% humidity and hot temperature: 0.777 + 0.436 - 0.137

Interpretation of the coefficients

```
plot_model(manova2)
```



Fit the ANOVA 2 model with interaction

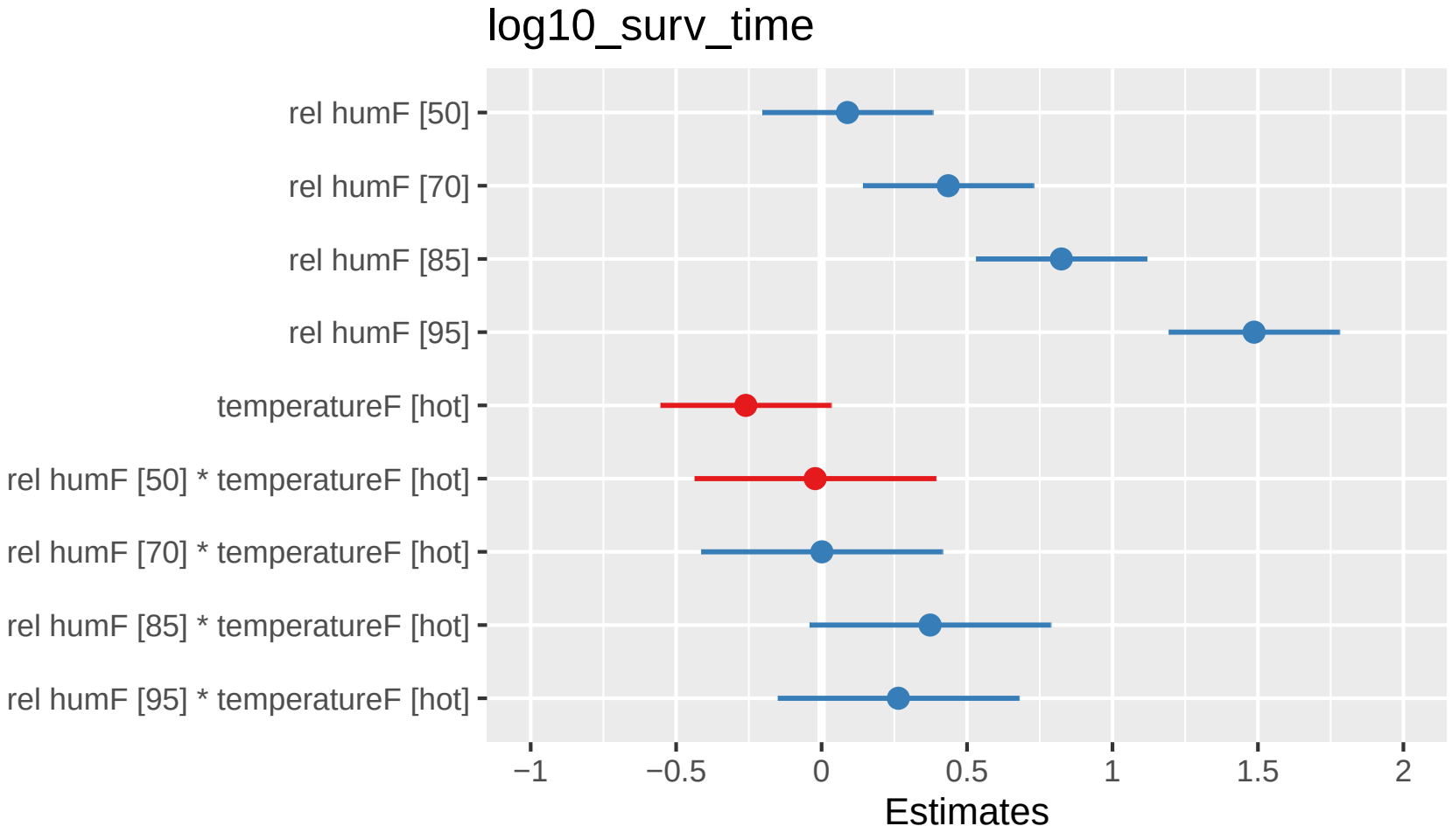
```
(manova2int <- lm(log10_surv_time ~ rel_humF + temperatureF +
                    rel_humF:temperatureF, data = dtot))
```

```
##
## 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. of predicted mean at 70% humidity and hot temperature: 0.839 + 0.435 - 0.261 + 0.001

Interpretation of the coefficients

```
plot_model(manova2int)
```



Comparison of models with and without interaction

Those two nested models can be compared using an F test.

```
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
```

On this example the interaction is not significant (but it should not be the only reason to make you choose the simpler model).

A linear model with both qualitative and quantitative independent variables

A linear model with both qualitative and quantitative independent variables

Imagine we want to model the impact on the survival time (in log_{10})

of the **relative humidity considered as a quantitative variable**

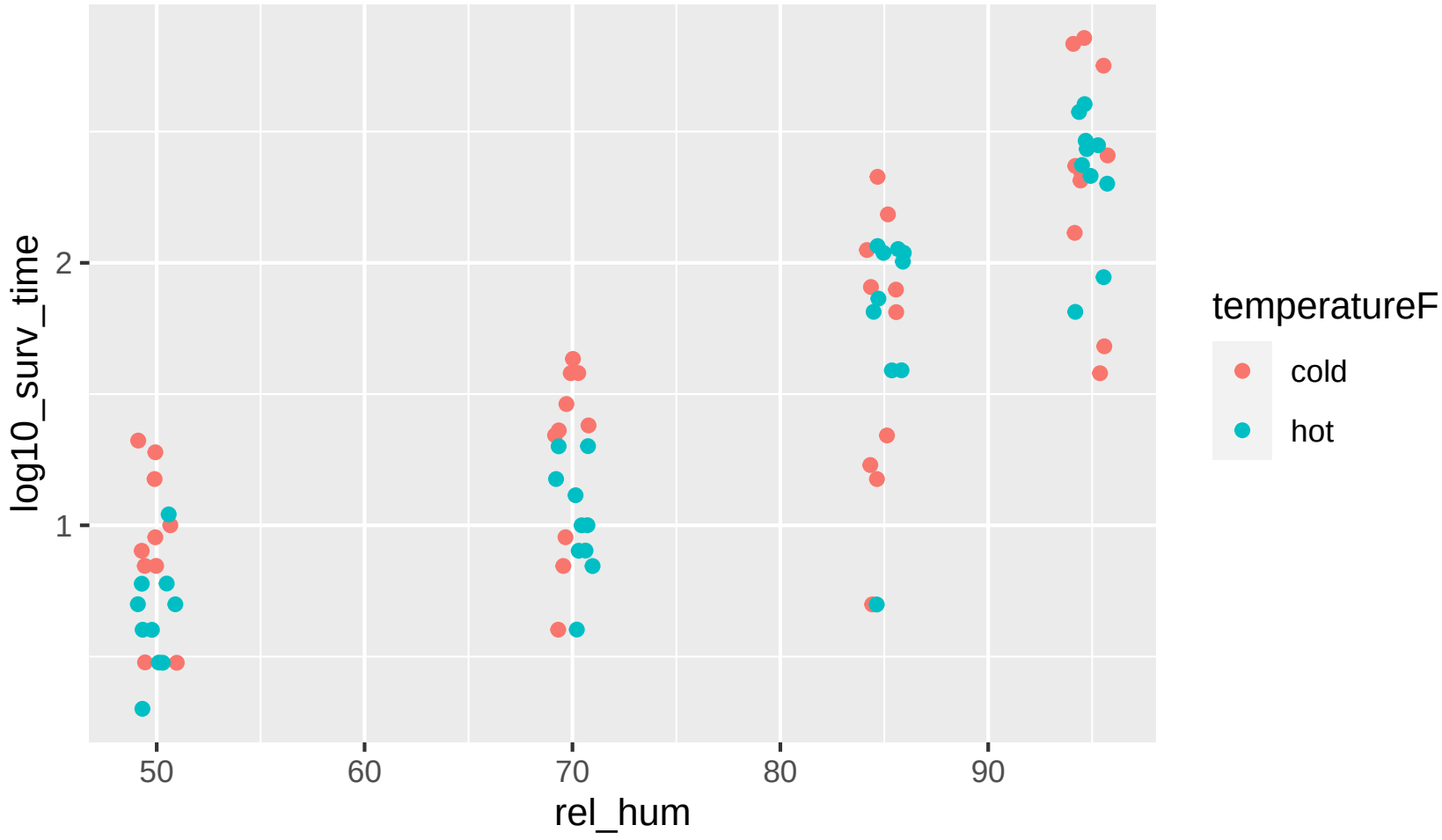
and the temperature **as a qualitative variable with 2 conditions**, cold ($< 15°C$) or hot,

excluding data at the driest condition.

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

Look at the data

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



ANCOVA models without or with interaction between a factor A and a covariate X

The model without interaction

The slope β does not depend on the modality of A.

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

The model with interaction

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

The slopes β_i are different

The choice between both models must be guided by the **prior biological knowledge** and the **data observation**.

Fit the ANCOVA model without interaction

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

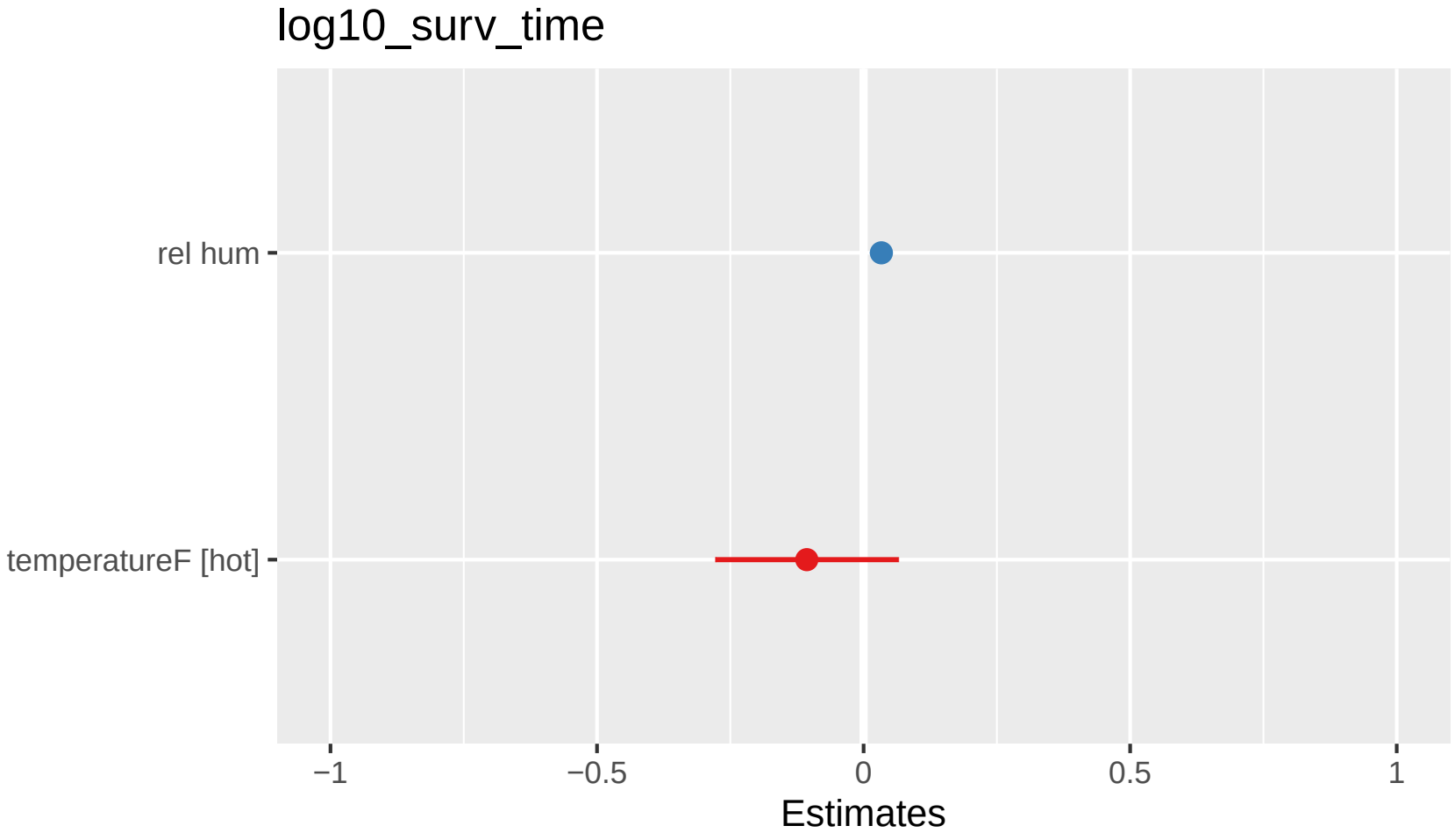
```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperatureF, data = dhum)
##
## Coefficients:
##     (Intercept)          rel_hum  temperatureFhot
##         -0.9533           0.0333          -0.1065
```

Note that the intercept has no biological in this case (0 not within the range of studied X values).

Ex. of predicted mean at 65% *humidity and hot temperature:*
$-0.9533 + 0.0333 \times 65 - 0.1065$

Interpretation of the coefficients

```
plot_model(mancova)
```



Standardization of the regression coefficients to help their interpretation

The **regression coefficients β (or β_i) are dependent of the order of magnitude of the covariate X.** To **make them comparable** to each other and to the coefficients corresponding to factors, it is recommended to divide them by $2 \times SD_X$ with SD_X the standard deviation of the X values.

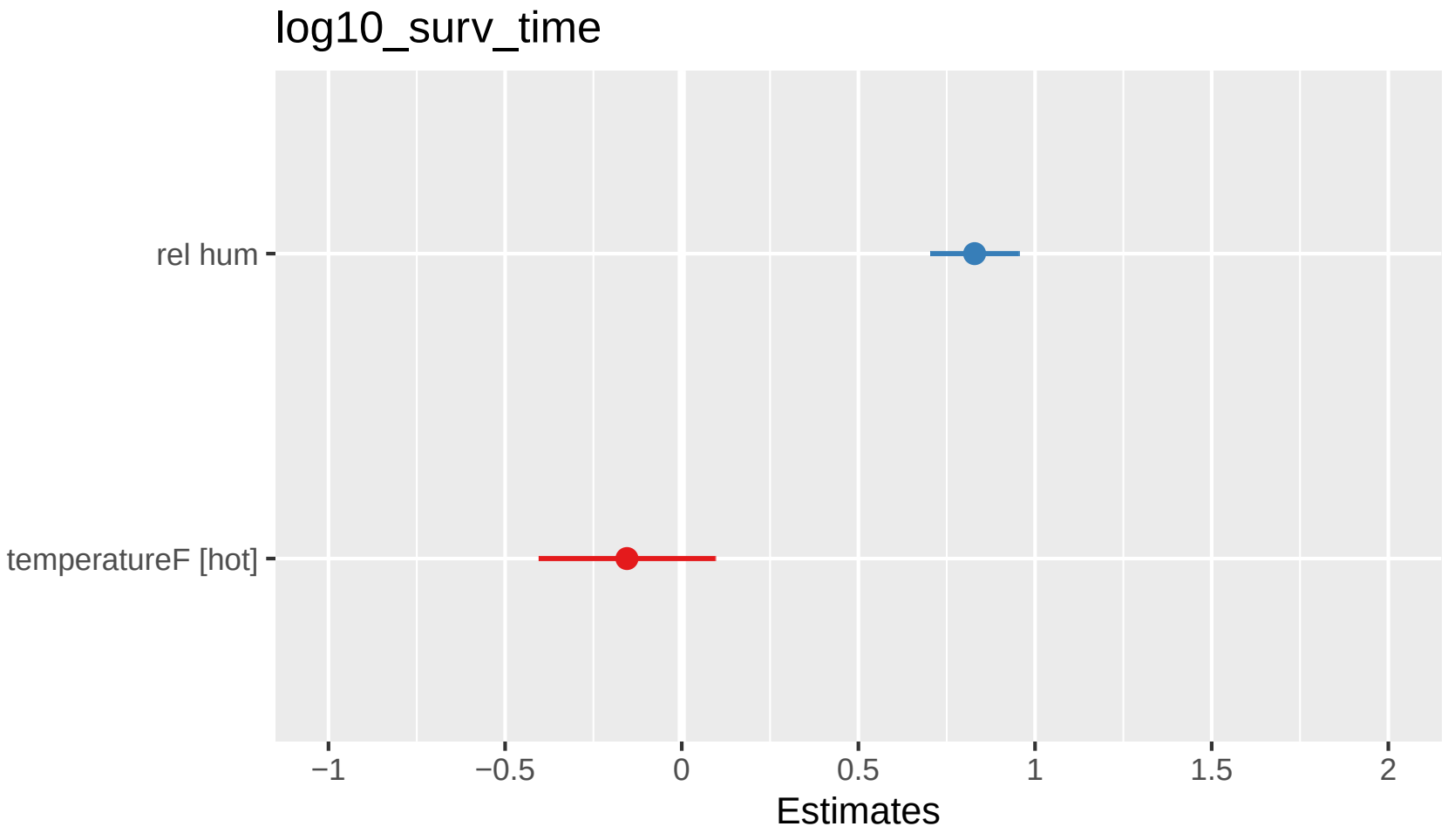
See Gelman A (2008) "Scaling regression inputs by dividing by two standard deviations." Statistics in Medicine 27: 2865-2873. for theoretical justification.

It can be easily done using the argument `type` of the `plot_model()`function as below.

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

Interpretation of standardized coefficients of the ANCOVA model without interaction

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



Fit the ANCOVA model with interaction

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

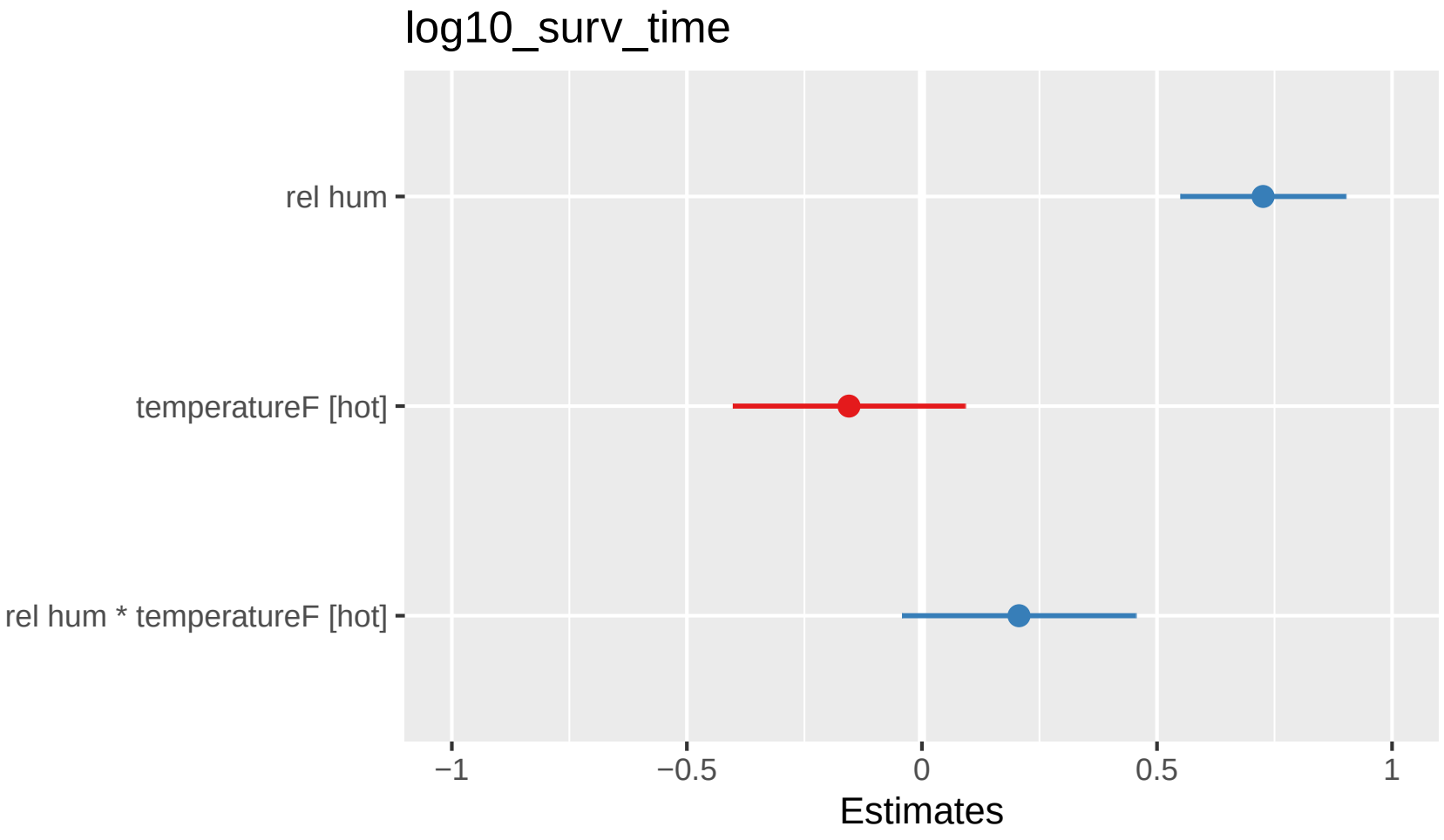
```
##
## Call:
## lm(formula = log10_surv_time ~ rel_hum + temperatureF + rel_hum:temperatureF
##     data = dhum)
##
## Coefficients:
##              (Intercept)                  rel_hum          temperatureFhot
##                  -0.64182                  0.02919                  -0.72941
## rel_hum:temperatureFhot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

Ex. of predicted mean at 65% *humidity and cold temperature:* $-0.6418 + 0.0292 \times 65$

And for predicted mean at 65% *humidity and cold temperature ?*

Interpretation of standardized coefficients of the ANCOVA model with interaction

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

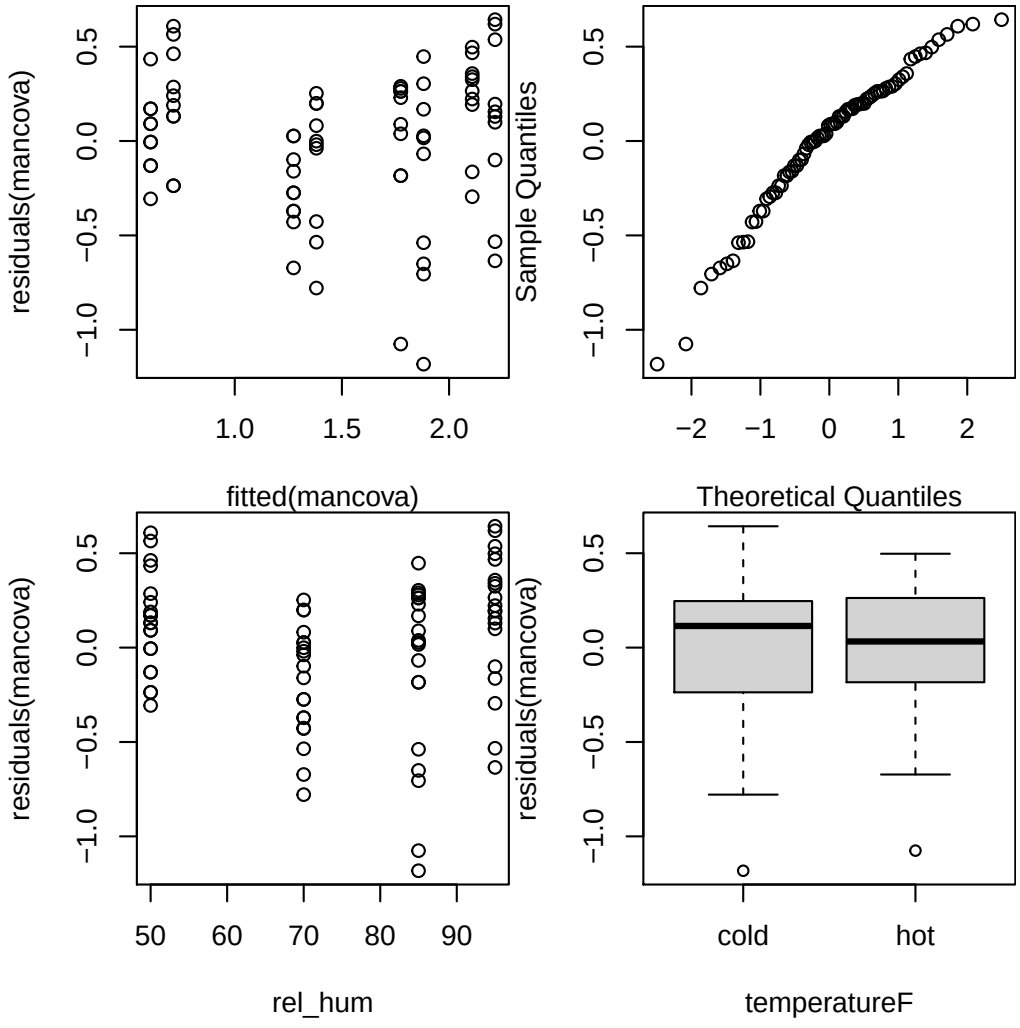


Your turn to handle qualitative and quantitative independent variables

Using those two fitted models (`mancova` and `mancovaint`),

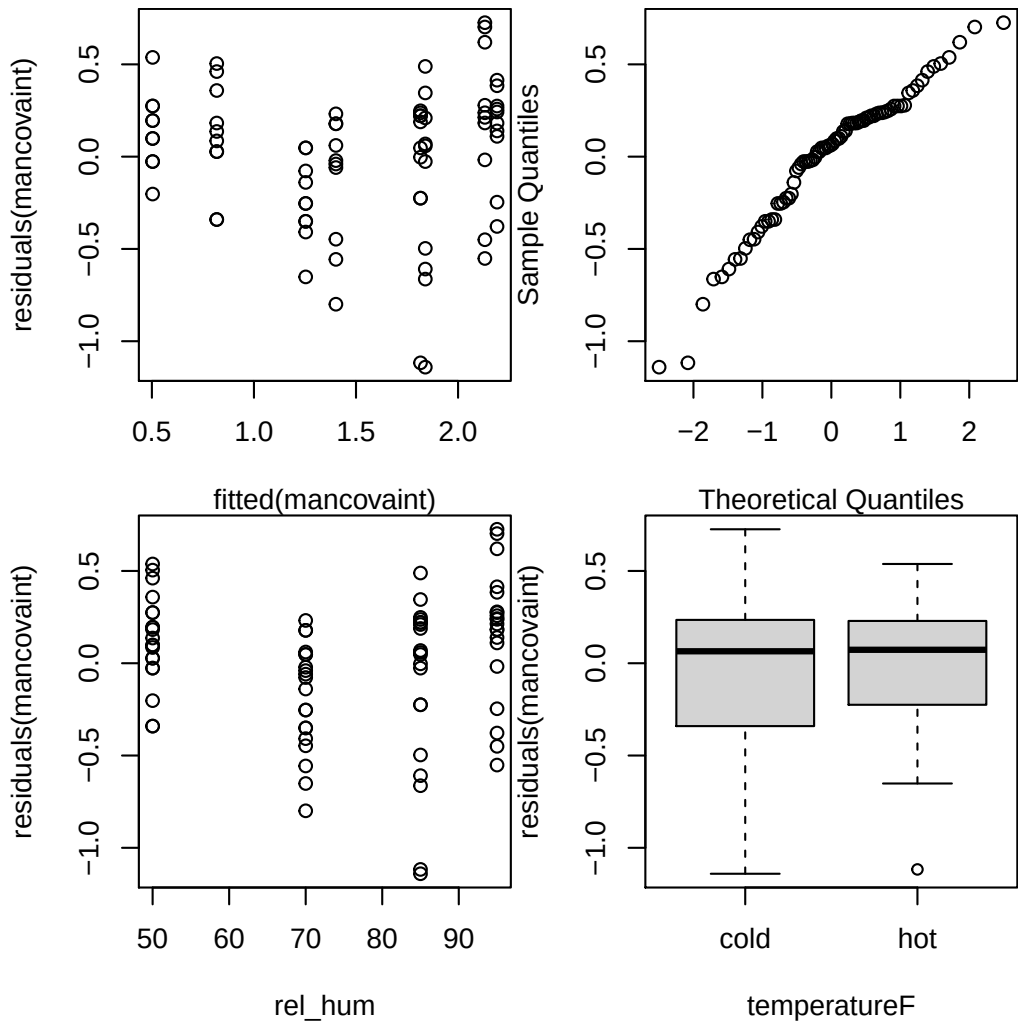
1. Look at the residuals for each model.
2. Predict the survival time of one tick exposed at 65% humidity and at a temperature between $15°C$ and $25°C$ with its 95% confidence interval, using the `predict()` function with each model.
3. Plot each model on the data the `abline()` function.

Answer to question 1 - model without interaction



The main observed trend is the one due to non linearity of the relation between the outcome and the relative humidity.

Answer to question 1 - model with interaction



As previously the main observed trend is the one due to non linearity of the relation between the outcome and the relative humidity.

Answer to question 2

```r
data4pred <- data.frame(rel_hum = 65, temperatureF = "hot")

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

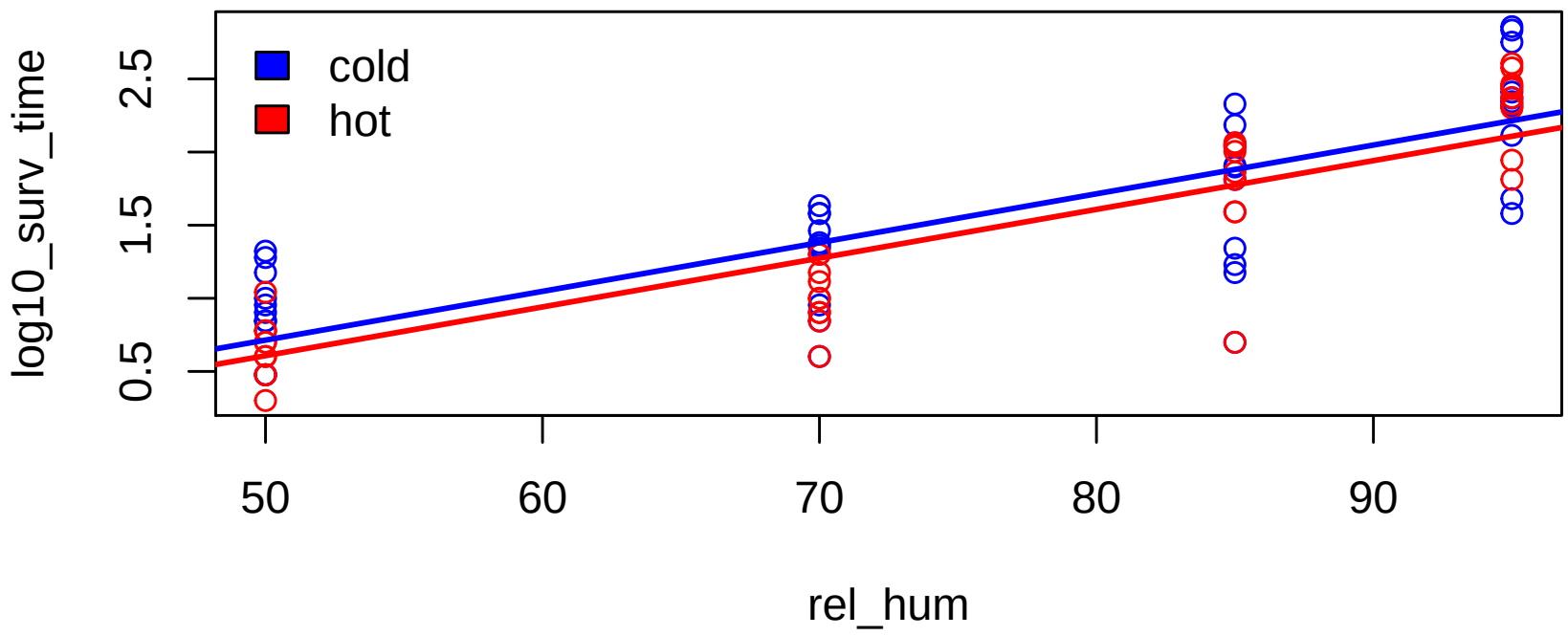
```
##    fit  lwr  upr
## 1 1.11 0.33 1.89
```

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

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

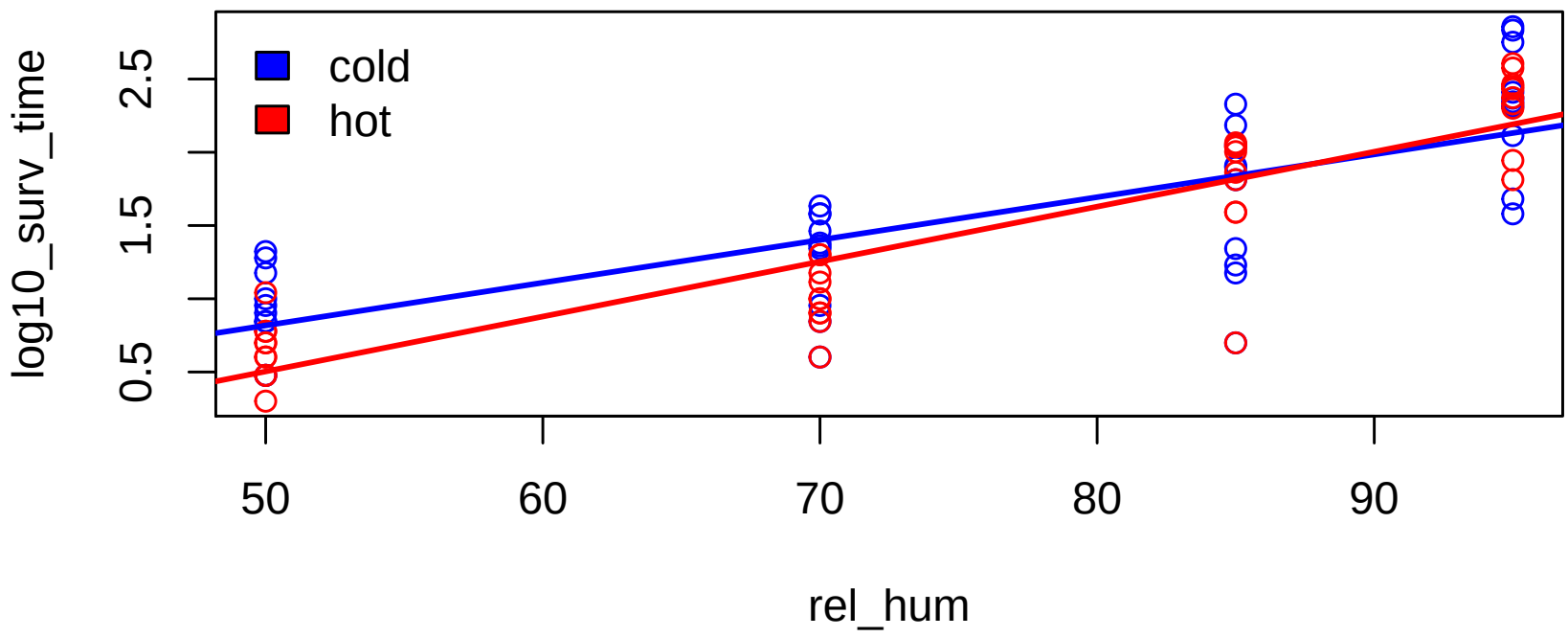
Answer to question 3 - model without interaction

```
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")
```



Answer to question 3 - model with 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 of the Gaussian linear model

More complex linear models

Now you have all the lego pieces to **build more complex linear models**.

And you have the basis to **understand the various extensions of the linear model**.

Non linear regression

If the **model is a non linear function of the parameters**.

In our example, if you want to model the non linear relationship between the survival time (in log) and the relative humidity using a more realistic model than the second order polynomial one.

Logistic regression

Logistic regression (special case of the Generalized Linear Model - GLM) is used when the outcome is no more continuous, but is a **binary outome**.

In our example, if you want the outcome is: survival or not at a specific time.

Logistic regression is very often used to **identify risk factors**, for example with the outcome being the presence or not of a disease in farms.

Mixed models

Mixed models are models taking into account **random effects** of random factors ($\neq$ deterministic effects of fixed factors),

such as farm effects, when there is more than one observation per farm,

or animal effect, when there is more than one observation per animal, ...

Survival models

Our example was a very specific one as the survival time was known for all the individuals.

More classically survival data include censored data (e.g. individuals that are not dead at the end of the study: their **survival time is right censored**, known to be above a value).

Moreover, the distribution of the survival times is not necessarily lognormal.

Those problems can be taken into account using a **semi-parametric approach** (Cox model) or a **parametric approach** (see Wongnak *et al.* 2022 on our example).

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.

Conclusion

Conclusion

Next time we will continue our discovery of linear models and their extensions by addressing

- the **strategy to build models** (which independent variables to incorporate),
- the **understanding of their coefficients**,
- and **some limits of the approach**