Series 4

1. At a certain time the proportion of chlorine in a product is 50% and it decreases with time. The dataset `http://stat.ethz.ch/Teaching/Datasets/cas-das/chlor.dat` contains the proportion of chlorine depending on time. We assume the nonlinear regression model:

$$\texttt{chlorine} = \alpha + (0.49 - \alpha) \cdot \exp(\beta \cdot \texttt{weeks} + \gamma) + \varepsilon \qquad (1)$$

a) Make a scatter plot of the data.

b) In order to fit the nonlinear regression model we first have to specify the starting values α_0, β_0 and γ_0 for α, β and γ. For the determination of the starting value α_0, proceed as follows:

- Since the proportion of chlorine decreases with time, β will have a negative sign. Therefore: $\lim_{\texttt{weeks}\to\infty} \texttt{chlorine} = \alpha$
  As a starting value for α we can thus choose the smallest (last measured?) value of chlorine.
- However to avoid problems with linearization (logarithm of negative numbers), α has to be smaller than all `chlorine` values. Therefore we may choose for instance $\alpha_0 = \min(\texttt{chlorine}) - 0.01$ as a starting value for α.

To determine β_0 and γ_0 you can then proceed in the following manner:

- Replace α by the specific value α_0 in the model formula (1) and linearize the model for β and γ. What linearized model do you get?
- Determine the starting values β_0 and γ_0 for your linearized model.
  R-Hint: Since there are two extreme observations, use the robust estimation method `lmrob(...)` from the R package `robustbase` to calculate β_0 and γ_0.

c) Fit the nonlinear regression model (1) and determine the estimates of α, β, γ.
(If you cannot solve **b)**, use $\alpha_0 = 0.37$, $\beta_0 = -0.05$, and $\gamma_0 = 0.25$ as starting values.)
R-Hint:

```
> r.nls <- nls(chlorine ~ alpha + (0.49 - alpha) * exp(beta * weeks + gamma),
               data = d.chlor,
               start = list(alpha =  ... , beta =  ... , gamma =  ... ),
               trace = TRUE)
```

d) Display the data and the fitted curve in a scatter plot. Also add the curve based on the starting values and compare the two curves.
R-Hint: Plug in your estimated values for $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\gamma}$:
`> f.chlor <- function(x)  {` $\widehat{\alpha}$ `+ (0.49 -` $\widehat{\alpha}$`) * exp(`$\widehat{\beta}$ `* x +` $\widehat{\gamma}$`)}`   # or use `predict()`
`> t.x <- seq(8, 42, 0.05); lines(t.x, f.chlor(t.x))`

e) Calculate 95% confidence intervals for α, β and γ by hand using the formula from 1.3.e. (Script). Compare your result with the output of `confint()`.

f) Run a residual analysis (TA-plot and QQ-plot). Does the model fit the data well?
R-Hint:

```
> plot(fitted(...), resid(...))
> qqnorm(resid(...))
> qqline(resid(...))
> identify(...)
```

Source: The dataset is discussed as an exercise in *Applied Regression Analysis, Wiley & Sons, 1966, p. 276*, written by Draper and Smith and has its origin in H. Smith and S. D. Dubey, *"Some reliability problems in the chemical industry", Industrial Quality Control, 21, 1964, no.2, p. 64-70.*

2. In a study of the Gubrist tunnel the following variables have been measured:

`pDiesel`	percentage of diesel vehicles in the traffic
`vFz`	average velocity of the vehicles (in km/h)
`vLuft`	air velocity in the tunnel (in m/s)
`Emiss.NOx`	average emission factor NO$_x$ of the vehicles (in mg/km)

Load the data via
`read.table("http://stat.ethz.ch/Teaching/Datasets/cas-das/gubrist.dat",`
`header = TRUE, sep = ";").`

We would like to estimate how much each vehicle type "benzin" and "diesel" contributes to the average emission factor, i.e. we would like to estimate the different emission factors for the two types "benzin" and "diesel". If we define β_B (= emission factor "benzin") and β_D (= emission factor "diesel") for these measures, we can use the model

$$\texttt{Emiss.NOx} = \beta_B \cdot \texttt{pBenzin} + \beta_D \cdot \texttt{pDiesel} + \varepsilon$$

to analyse the data. With $\texttt{pBenzin} = 1 - \texttt{pDiesel}$ we get

$$\texttt{Emiss.NOx} = \alpha + \beta \cdot \texttt{pDiesel} + \varepsilon \,, \tag{2}$$

where $\alpha = \beta_B$ and $\beta = -\beta_B + \beta_D$. The estimated coefficient $\widehat{\alpha} = \widehat{\beta_B}$ is the emission factor for the "benzin" vehicles and $\widehat{\alpha} + \widehat{\beta} = \widehat{\beta_D}$ the emission factor for "diesel" vehicles.

a) Estimate the coefficients α and β in model (2) with the least squares method.
R-Hint: Use the argument "`na.action = na.omit"` in `lm()` to omit the missing values in the dataset.

b) Make a residual analysis of the model in **a)**. In particular look at the TA-plot and the QQ-plot. Are the assumptions of the model fulfilled?

c) We should log-transform the response because we can see that the distribution of the residuals in the linear model (2) is skewed. Only transforming the target variable is not very recommended in this situation, since the estimated coefficients in the transformed model loose their meaning of being emission factors. Therefore we transform both sides of the model (2):

$$\log(\texttt{Emiss.NOx}) = \log(\alpha + \beta \cdot \texttt{pDiesel}) + \widetilde{\varepsilon}\,. \tag{3}$$

We assume a symmetric distribution for the error $\widetilde{\varepsilon}$. (Note that this model is equivalent to $\texttt{Emiss.NOx} = (\alpha + \beta \cdot \texttt{pDiesel}) \cdot \varepsilon^*$ with ε^* being lognormal distributed.)
Estimate the coefficients α and β of the transformed model (3) with the nonlinear least squares method. Which starting values do you use? Are the estimated coefficients different from the results in **a)**?
R-Hint: With `na.omit(d.gubrist)` you may delete the missing values in `d.gubrist`.

```
> r.nonlin <- nls(... ~ ..., data = na.omit(d.gubrist), start = list(..., ...))
```

d) Make a residual analysis. Was it worth doing the transformation?

Exercise hour: Monday, June 17, afternoon.