Solution to Series 4

1. **a)** For a graphical display see solution to **d)**.

b) In order to be able to linearize properly, we choose $\alpha < \min(\texttt{chlorine})$. Taking $\alpha_0 = \min(\texttt{chlorine}) - 0.01 = 0.38 - 0.01 = 0.37$ as the starting value, we can linearize the nonlinear model (1) as follows:

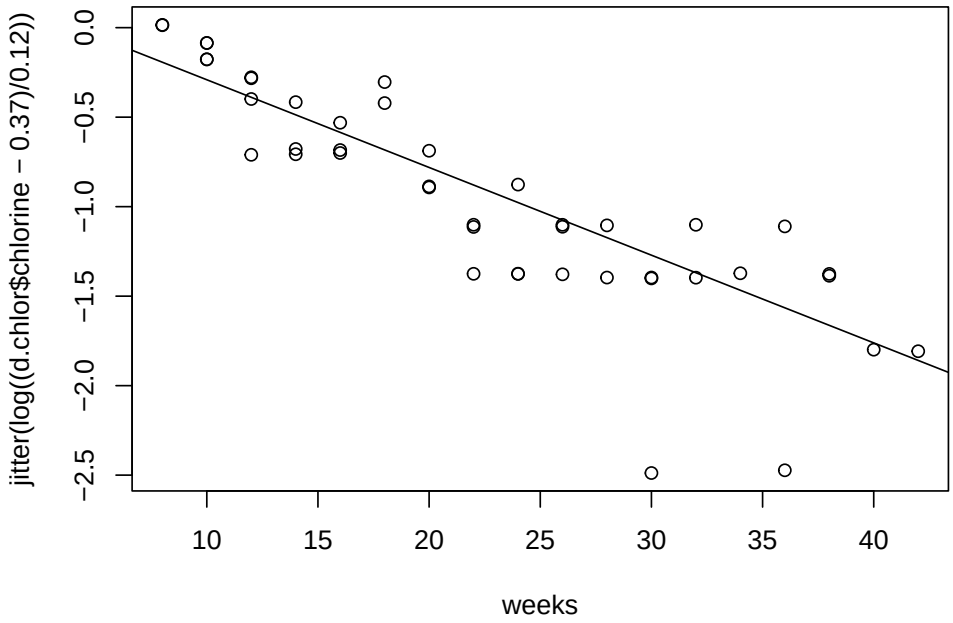
$$\begin{aligned} \texttt{chlorine} &= 0.37 + (0.49 - 0.37) \cdot \exp(\beta \cdot \texttt{weeks} + \gamma) \\ \frac{\texttt{chlorine} - 0.37}{(0.49 - 0.37)} &= \exp(\beta \cdot \texttt{weeks} + \gamma) \\ \log\left(\frac{\texttt{chlorine} - 0.37}{0.12}\right) &= \beta \cdot \texttt{weeks} + \gamma \end{aligned}$$

This equation looks like a linear regression model, however the structures of the residuals are wrong. Nonetheless we fit it using least-squares or an MM-algorithm, in order to get starting values. In this case, we use the MM method which yields: $\beta_0 = -0.05$, $\gamma_0 = 0.20$:

```
> d.chlor <- read.table("http://stat.ethz.ch/Teaching/Datasets/cas-das/chlor.dat",
                        header = TRUE)
> library(robustbase)
> r.rlm <- lmrob(log((chlorine - 0.37) / 0.12) ~ weeks, data = d.chlor)
> summary(r.rlm)

Call:
lmrob(formula = log((chlorine - 0.37)/0.12) ~ weeks, data = d.chlor)
 \--> method = "MM"
Residuals:
    Min      1Q  Median      3Q     Max
-1.2138 -0.2060 -0.0168  0.1088  0.4665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19890    0.07633    2.61    0.013 *
weeks       -0.04900    0.00348  -14.08   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.24
Multiple R-squared:  0.803,        Adjusted R-squared:  0.799
Convergence in 11 IRWLS iterations

Robustness weights:
 observation 35 is an outlier with |weight| = 0 ( < 0.0023);
 4 weights are ~= 1. The remaining 39 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.111   0.883   0.942   0.900   0.984   0.999
Algorithmic parameters:
       tuning.chi                bb        tuning.psi
         1.55e+00          5.00e-01          4.69e+00
       refine.tol           rel.tol         scale.tol
         1.00e-07          1.00e-07          1.00e-10
        solve.tol          zero.tol       eps.outlier
         1.00e-07          1.00e-10          2.27e-03
            eps.x warn.limit.reject warn.limit.meanrw
         7.64e-11          5.00e-01          5.00e-01
     nResample         max.it       best.r.s       k.fast.s
```

```
          500             50              2              1
      k.max    maxit.scale      trace.lev            mts
          200             200              0           1000
compute.rd fast.s.large.n
           0          2000
               psi           subsampling
         "bisquare"         "nonsingular"
               cov compute.outlier.stats
      ".vcov.avar1"                 "SM"
seed : int(0)
```

```
> plot(d.chlor$weeks, jitter(log((d.chlor$chlorine - 0.37) / 0.12)), xlab = "weeks")
> abline(r.rlm, lty = 1)
```



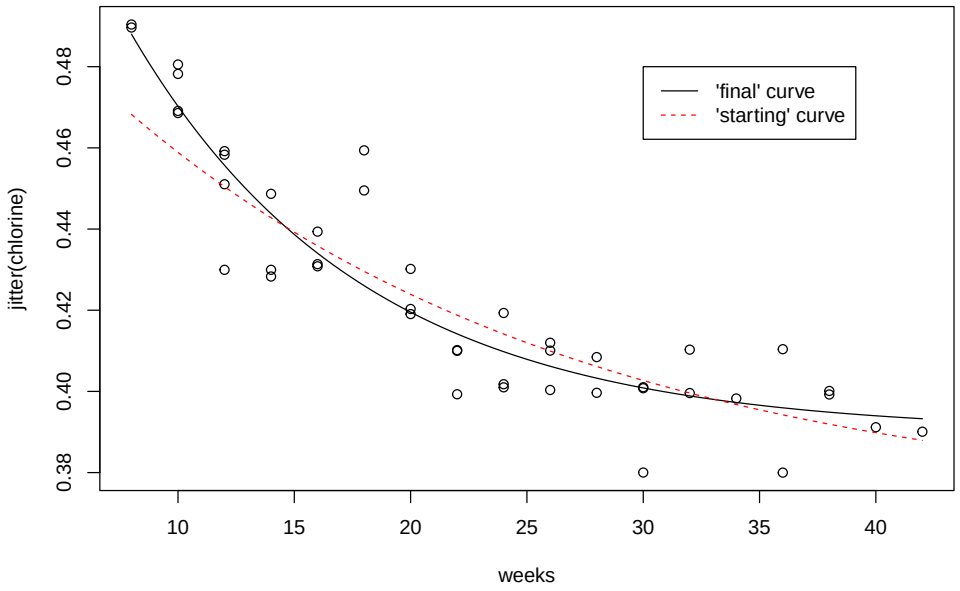
c) Using non linear regression yields the following estimation of the parameters:
$\hat{\alpha} = 0.39,\ \hat{\beta} = -0.1$ and $\hat{\gamma} = 0.78$:

```
> r.nls <- nls(chlorine ~ alpha + (0.49 - alpha) * exp(beta * weeks + gamma),
               data = d.chlor, start = list(alpha = 0.37, beta = -0.05, gamma = 0.20))
> summary(r.nls)

Formula: chlorine ~ alpha + (0.49 - alpha) * exp(beta * weeks + gamma)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
alpha  0.38963    0.00584   66.70  < 2e-16 ***
beta  -0.09916    0.01810   -5.48  2.4e-06 ***
gamma  0.78200    0.18652    4.19  0.00014 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.011 on 41 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 3.43e-06
```

d)

```
> f.chlor <- function(x){ 0.39 + (0.49-0.39) * exp(-0.1 * x + 0.78) }
> plot(jitter(chlorine) ~ weeks, d.chlor)
> t.x <- seq(8, 42, 0.05)
> lines(t.x, f.chlor(t.x))
> f.chlor2 <- function(x){0.37 + (0.49-0.37) * exp(-0.05 * x + 0.2)}
```

```
> lines(t.x, f.chlor2(t.x), col = "red", lty = 2)
> legend(30, 0.48, legend = c("'final' curve", "'starting' curve"),
         col = c("black", "red"), lty = c(1, 2))
```



e) We can construct approximate $(1-\alpha)\cdot 100\%$ confidence intervals using the formula

$$\hat{\theta}_k \pm q_{1-\alpha/2}^{t_{n-p}} \cdot \sqrt{\hat{V}_{kk}} = \hat{\theta}_k \pm q_{1-\alpha/2}^{t_{n-p}} \cdot s.e.(\hat{\theta}_k).$$

Here, we use $n = 44$ and $p = 3$ since we have 44 observations and we estimate the three parameters α, β and γ.

```
> ## by hand
> h <- qt(0.975, 41) * summary(r.nls)$coefficients[, 2]
> coef(r.nls) + cbind(-h, h)
```

```
                    h
alpha  0.378  0.4014
beta  -0.136 -0.0626
gamma  0.405  1.1587
```

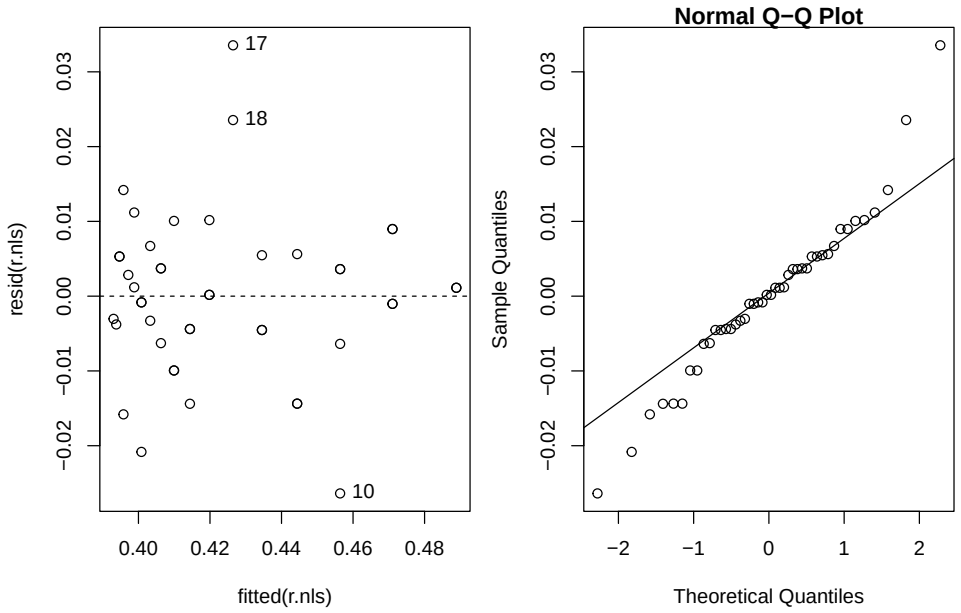
```
> ##  with confint
> confint(r.nls)
```

```
        2.5%   97.5%
alpha  0.374  0.3995
beta  -0.140 -0.0647
gamma  0.429  1.1952
```

The two calculations yield almost the same results. But why are they not identical? (Answer in the lecture next week)

f) The TA and the QQ-plot show three peculiar observations (10, 17 and 18) with corresponding residuals having absolute values larger than 0.02. One should check what is wrong with these observations:

```
> par(mfrow = c(1, 2))
> plot(fitted(r.nls), resid(r.nls))
> abline(h = 0, lty = 2)
> identify(fitted(r.nls), resid(r.nls))
> h <- qqnorm(resid(r.nls))
> qqline(resid(r.nls))
> identify(h)
```



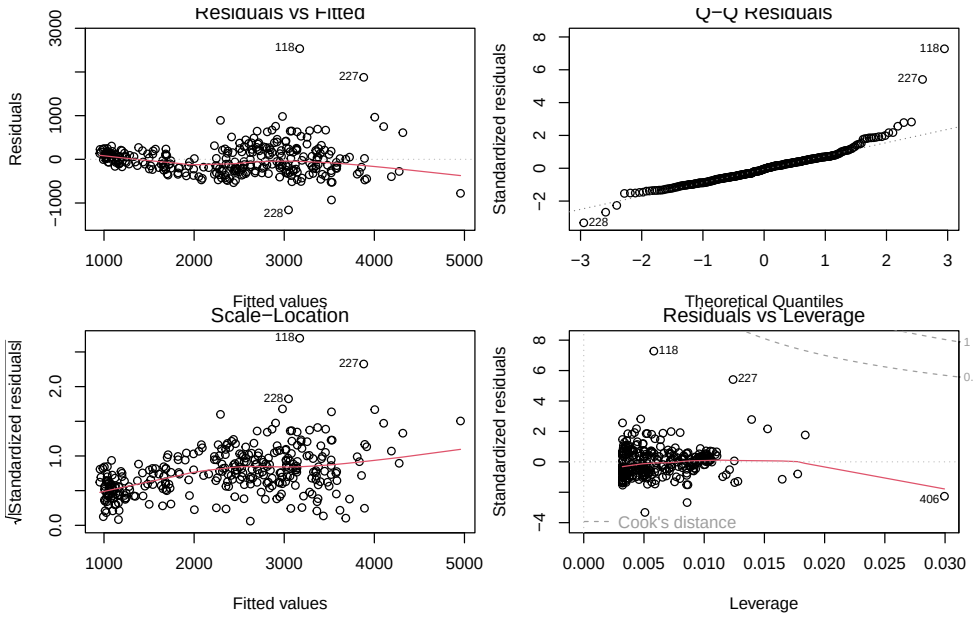
2. **a)** **Linear model:** $\texttt{Emiss.NOx} = \alpha + \beta \cdot \texttt{pDiesel} + \varepsilon$

```
> d.gubrist <- read.table("http://stat.ethz.ch/Teaching/Datasets/cas-das/gubrist.dat",
                          header = TRUE, sep = ";")
> r.lin <- lm(Emiss.NOx ~ pDiesel, data = d.gubrist, na.action = na.omit)
> summary(r.lin)

Call:
lm(formula = Emiss.NOx ~ pDiesel, data = d.gubrist, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-1158.4  -214.0   -15.5   167.4  2534.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    605.2       43.4    13.9   <2e-16 ***
pDiesel      12712.9      279.7    45.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349 on 310 degrees of freedom
  (360 observations deleted due to missingness)
Multiple R-squared:  0.87,        Adjusted R-squared:  0.869
F-statistic: 2.07e+03 on 1 and 310 DF,  p-value: <2e-16
```

The estimated emission factors are:

category	emission factor (in mg/km) per vehicle
Benzin	$\widehat{\alpha} = 605$
Diesel	$\widehat{\alpha} + \widehat{\beta} = 605 + 12713 = 13318$

b)
```
> par(mfrow = c(2, 2))
> plot(r.lin)
```



The normal plot shows that the model assumptions are not met. The distribution of the residuals is skewed and there are some outliers. If there were only outliers (but residuals that are not skewed) model (2) should be estimated using a robust method. Unfortunately this does not help with the skewness of the error distribution. The increasing variance suggests a transformation of the response variable.

c) **Nonlinear model:** $\log(\texttt{Emiss.NOx}) = \log(\alpha + \beta \cdot \texttt{pDiesel}) + \widetilde{\varepsilon}$

As starting values for the nonlinear regression, we can use the values $\widehat{\alpha} = 605$ and $\widehat{\beta} = 12713$ estimated in **a)**.

```
> r.nonlin <- nls(log(Emiss.NOx) ~ log(alpha + beta * pDiesel),
                  data = na.omit(d.gubrist), start = list(alpha = 605, beta = 12713))
> summary(r.nonlin)

Formula: log(Emiss.NOx) ~ log(alpha + beta * pDiesel)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
alpha    668.7       21.2    31.5   <2e-16 ***
beta   12046.8      209.0    57.6   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.125 on 310 degrees of freedom

Number of iterations to convergence: 3
Achieved convergence tolerance: 2.83e-06
```

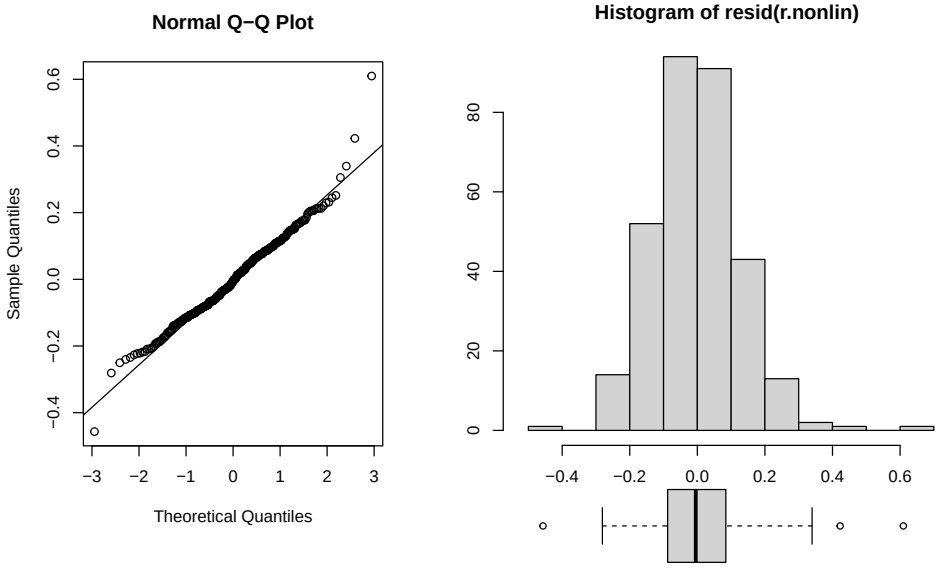
We get the following estimated emission factors:

category	emission factor (in mg/km) per vehicle
Benzin	$\widehat{\alpha} = 669$
Diesel	$\widehat{\alpha} + \widehat{\beta} = 669 + 12047 = 12716$

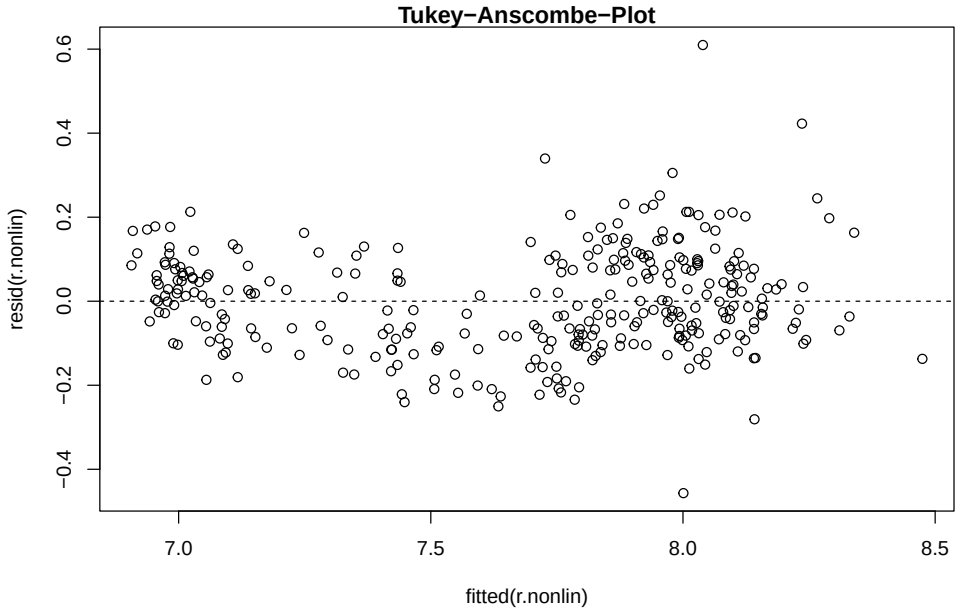
d)

```
> ### QQ plot and histogram
> #  define layout
> t.mat <- matrix(c(1, 2, 1, 3), 2, byrow = TRUE)
> t.nf <- layout(t.mat, widths = c(5, 6), heights = c(4, 1))
> # layout.show(t.nf)
> # make plots
> t.range <- range(resid(r.nonlin)) * 1.15
> par(mar = c(7, 4, 4, 3))
> qqnorm(resid(r.nonlin)); qqline(resid(r.nonlin))
```

```
> par(mar = c(1, 3, 3, 1))
> hist(resid(r.nonlin), breaks = 15, xlim = t.range, xlab = "")
> par(mar = c(0, 3, 0, 1))
> boxplot(resid(r.nonlin), horizontal = TRUE, ylim = t.range, boxwex = 1.3,
          axes = FALSE)
```




```
> ### Tukey-Anscombe-Plot
> plot(fitted(r.nonlin), resid(r.nonlin)); abline(h = 0, lty = 2)
> title("Tukey-Anscombe-Plot")
```



For the nonlinear model, the QQ-plot shows a less skewed distribution of the errors. There still seem to be some outliers. One should check those and ideally fit model (3) using a robust method.
The estimated regression lines do not differ much. But the estimated standard error of α (this is the emission factor for "benzin") is reduced by 50% due to a better structure of the residuals.