Zurich University
of Applied Sciences

zh
aw

School of
Engineering

IDP Institute of
Data Analysis and
Process Design

WBL Statistik 2024 — Nonlinear Regression

A Powerful Tool With Considerable Complexity

Half-Day 2: Improved Inference and Visualisation

Andreas Ruckstuhl
Institut für Datenanalyse und Prozessdesign
Zürcher Hochschule für Angewandte Wissenschaften

Outline:

Half-Day 1 Estimation and Standard Inference

- The Nonlinear Regression Model
- Iterative Estimation - Model Fitting
- Inference Based on Linear Approximations

Half-Day 2 Improved Inference and Visualisation

- Likelihood Based Inference
- Profile t Plot and Profile Traces
- Parameter Transformations

Half-Day 3 Bootstrap, Prediction and Calibration

- Bootstrap
- Prediction
- Calibration

Outlook

2.1 Likelihood Based Inference

- F-Test for the whole parameter vector $\underline{\theta}^*$:

$$T = \frac{(n-p)}{p} \cdot \frac{S\left\langle\underline{\theta}^*\right\rangle - S\left\langle\widehat{\underline{\theta}}\right\rangle}{S\left\langle\widehat{\underline{\theta}}\right\rangle} \overset{a}{\sim} F_{p,n-p} \, .$$

  It is like in linear regression, where the result is exactly correct for every n.

- And the resulting confidence region is

$$\left\{ \underline{\theta} \;\middle|\; S\left\langle\underline{\theta}\right\rangle \leq S\left\langle\widehat{\underline{\theta}}\right\rangle \left(1 + \tfrac{p}{n-p}\, q_{1-\alpha}^{F_{p,n-p}}\right) \right\} \, .$$

- In case of the linear regression, this confidence region is identical to the confidence region based on multivariate normal distribution of $\widehat{\underline{\beta}}$.

In case of the nonlinear regression, this confidence region is more accurate than that one based on multivariate normal distribution of $\widehat{\underline{\beta}}$.

However, it is very difficult to calculate this more accurate confidence region!

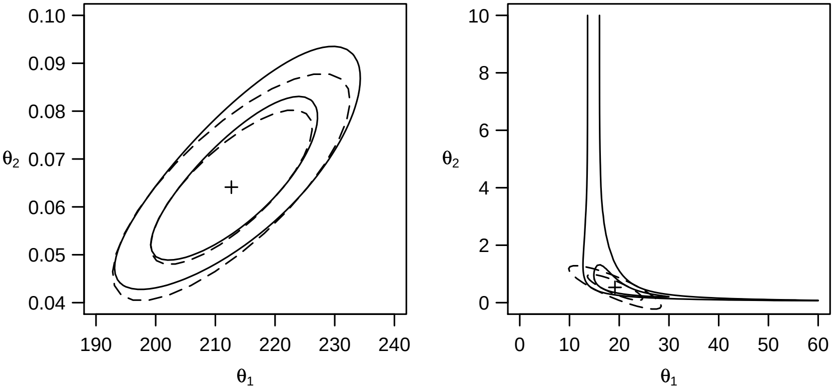
$p = 2$: We can determine the more accurate confidence region by standard contouring methods, that is, by evaluating $S\left\langle\underline{\theta}\right\rangle$ for a grid of $\underline{\theta}$ values and approximating the contours by straight line segments in the grid.

example, see next slide

$p \geq 3$: There are no contour plots.

Likelihood Contour Lines

Nominal 80 and 95% likelihood contours lines (——) and confidence ellipsoids based on Wald-type asymptotic approximations (– – – –). + indicates the least-squares estimation. These solutions do agree satisfactorily in the example Puromycin (left), but do disagree in the example 'Biochemical Oxygen Demand' (right) clearly.



F-Test for a single Parameter: „$\theta_k = \theta_k^*$“

- - Such a null hypothesis ignores the other parameters.
  - The other parameters, $\underline{\theta}_{\text{-}k}$, are fitted to the data by least-squares ☞ $\underline{\widetilde{\theta}}_{\text{-}k}$.
  - The minimum is called $\widetilde{S}_k$. It depends on θ_k^*, hence $\widetilde{S}_k := \widetilde{S}_k \langle \theta_k^* \rangle$.

- The F-test statistic for the test "$\theta_k = \theta_k^*$" is

$$\widetilde{T}_k = (n-p)\,\frac{\widetilde{S}_k \langle \theta_k^* \rangle - S\left\langle \underline{\widehat{\theta}} \right\rangle}{S\left\langle \underline{\widehat{\theta}} \right\rangle,} \;.$$

It is approximately $F_{1,n-p}$ distributed.

- In **linear** regression, this F-test is equivalent to the t-test,
  since the test statistic of the F-test is proportional to the squared of the test statistic of the t-test.

- In **nonlinear** regression, this F-test is **not** equivalent to the t-test of the asymptotic Wald-type test.

A more accurate 't-Test'

Based on the previous result, we can construct a t-type test which is more accurate than that introduced initially:

Take the square-root from the F-test statistic and multiply it with the sign of $\widehat{\theta}_k - \theta_k^*$,

$$T_k \left\langle \theta_k^* \right\rangle := \operatorname{sign} \left\langle \widehat{\theta}_k - \theta_k^* \right\rangle \frac{\sqrt{\widetilde{S}_k \left\langle \theta_k^* \right\rangle - S \left\langle \underline{\widehat{\theta}} \right\rangle}}{\widehat{\sigma}} .$$

This test statistic is t_{n-p} distributed approximately.

(In linear regression, this test statistic is equivalent to the usual t-test.)

2.2 Profile t Plot and Profile Traces

Based on the just introduced test statistic, a graphical tool called **profile t plot** can be designed for assessing the quality of the linear approximation:

We plot the test statistic $T_k \left\langle \theta_k^* \right\rangle$ as a function of θ_k^* – the **profile t function**

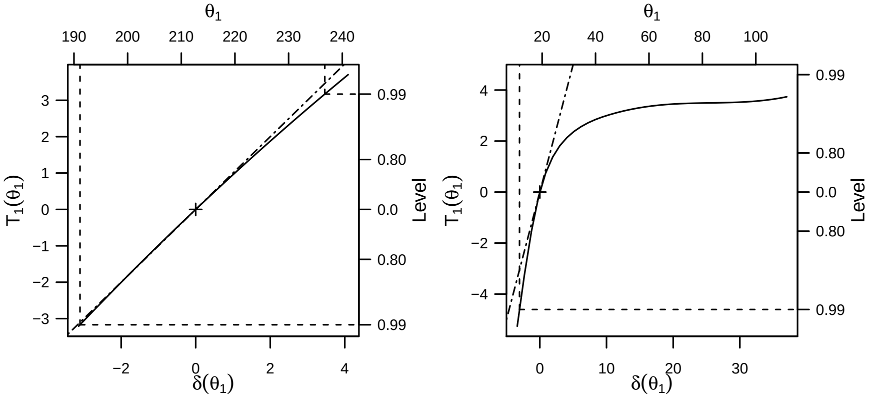
- In **linear regression**, the profile t function is a **straight line**.
- In **nonlinear regression**, the profile t function can be **any monotone increasing function**.

Profile t Plot:

$$\text{Plot } T_k \left\langle \theta_k^* \right\rangle \text{ versus } \quad \delta_k \left\langle \theta_k^* \right\rangle := \frac{\theta_k^* - \widehat{\theta}_k}{se\left\langle (\widehat{\theta}_k) \right\rangle}$$

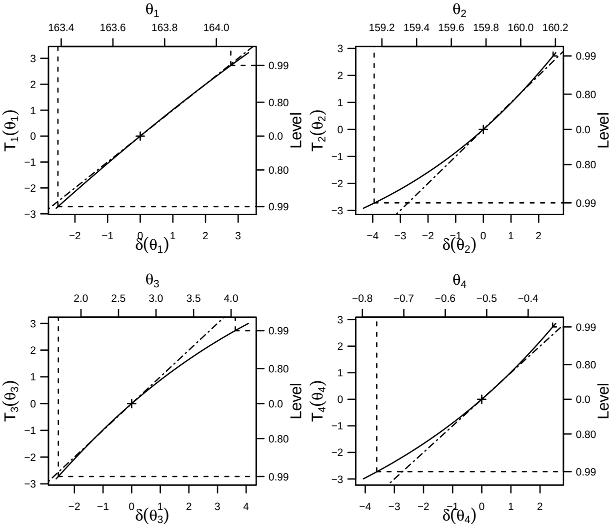
- The more curved the profile t function is the stronger the nonlinearity in a neighbourhood of $\widehat{\theta}_k$!
- Hence, the profile t plot shows how accurate the linear approximation of the standard test and standard confidence interval is.
- The neighbourhood important for statistics is given by $|\delta_k \left\langle \theta_k^* \right\rangle| \leq 2.5$. Why?

Example: Profile t Plots



Profile *t* Plot (——) for θ_1 for the examples Puromycin data (left) and Biochemical Oxygen Demand data (right).

Example: Cellulose membrane (5) - Profile t plots



Example: Cellulose membrane (6)

Wald-type CI

R Output:

Parameters:

	Value	Std. Error	t value
θ_1	163.706	0.1262	1297.21
θ_2	159.784	0.1595	1002.03
θ_3	2.675	0.3813	7.02
θ_4	-0.512	0.0703	-7.28

Residual standard error: 0.293 on 35 df

Approximate 95% confidence intervals

$(\widehat{\theta}_k \pm se\left\langle\widehat{\theta}_k\right\rangle \cdot q_{0.975}^{t_{35}})$

θ_1:	[163.45, 163.96]
θ_2:	[159.46, 160.11]
θ_3:	[1.90, 3.45]
θ_4:	[-0.65, -0.37]

“profile”-type CI

R Output:

> confint(Mem.fit)
Waiting for profiling to be done...

	2.5%	97.5%
θ_1	163.4661097	163.9623994
θ_2	159.3562993	160.0952200
θ_3	1.9262575	3.6407940
θ_4	-0.6882365	-0.3797975

θ_1:	[163.47, 163.96]
θ_2:	[159.36, 160.10]
θ_3:	[1.93, 3.64]
θ_4:	[-0.69, -0.38]

Likelihood Profile Traces

Likelihood profile traces are another useful tool.

The Parameter $\widetilde{\theta}_j$, estimated at $\theta_k = \theta_k^*$ $(k \neq j)$, is evaluated as a function; hence the notation $\widetilde{\theta}_j^{(k)} \left\langle \theta_k^* \right\rangle$.

Remember:

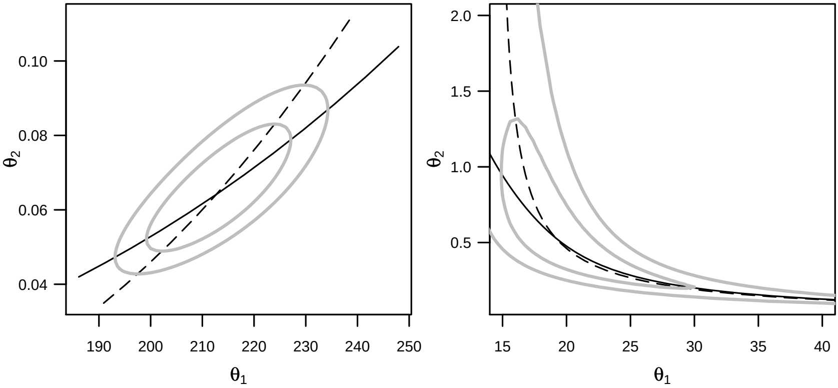
$$\min_{\{\theta_h, h \neq k\}} S\left\langle \theta_1, \ldots, \theta_k^*, \ldots, \theta_p \right\rangle = S\left\langle \widetilde{\theta}_1, \ldots, \widetilde{\theta}_{k-1}, \theta_k^*, \widetilde{\theta}_{k+1}, \ldots, \widetilde{\theta}_p \right\rangle \overset{short}{=} \widetilde{S}_k \left\langle \theta_k^* \right\rangle$$

Plot the profile trace $\widetilde{\theta}_j^{(k)}$ versus θ_k^* overlaid by the profile trace $\widetilde{\theta}_k^{(j)}$ versus θ_j^* but reflected at the 45° line; that is

	y-coordinate	vs	x-coordinate	line type
	$\widetilde{\theta}_j^{(k)}$	vs	θ_k^*	solid
overlaid by	θ_j^*	vs	$\widetilde{\theta}_k^{(j)}$	dashed

Examples of Likelihood Profile Traces

Likelihood Profile Traces for the example Puromycin (left) and the example Biochemical Oxygen Demand (right), complemented by the 80%- and 95% confidence region (gray curve)



Properties of Likelihood Profile Traces

With linear regression:

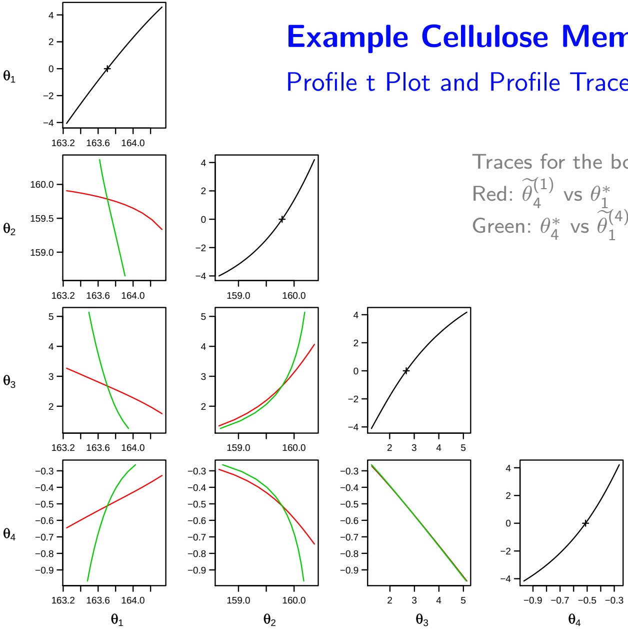
- The profile traces are two straight lines.
- The angle between these two lines represents the correlation between the estimated parameters corresponding to the lines
- If the correlation between the parameters is 0, then the lines are orthogonal to each other.
- If the correlation between the parameters is either 1 or -1, then the lines overlay.

With nonlinear regression:

- Both traces may be curved.
- The heavier the traces deviated from a straight line, the more insufficient is the linear approximation and the inference based on it.
- The angle between these two traces at the intersection still represents the correlation between the two estimated parameters $\widehat{\theta}_j$ and $\widehat{\theta}_k$.

Example Cellulose Membrane (7)

Profile t Plot and Profile Traces.



Traces for the bottom left corner:

Red: $\widetilde{\theta}_4^{(1)}$ vs θ_1^*

Green: θ_4^* vs $\widetilde{\theta}_1^{(4)}$

2.3 Parameter Transformations

- In this section we study the effects of transforming the parameters.
- This topic is based on the fact that the mean regression function can usually be written down by mathematically equivalent expressions.

- For example
  - The two expression for the Michaelis-Menten function are equivalent

$$\frac{\theta_1 x}{\theta_2 + x} = \frac{x}{\varphi_1 + \varphi_2 x} \, .$$

Hence

$$\varphi_1 := \frac{\theta_2}{\theta_1} \quad \text{and} \quad \varphi_2 := \frac{1}{\theta_1} \, .$$

  - Or, we have the two equivalent expressions

$$\theta_1 e^{\theta_2 x} = \varphi_1 \varphi_2^x$$

hence,

$$\varphi_1 := \theta_1 \quad \text{and} \quad \varphi_2 := e^{\theta_2} \, .$$

Motivation

The parameters of the regression function are transformed to

- get rid of **collinearities**
- improve the **convergence** of the algorithm
- improve the linear approximation (e.g., the Wald-type asymptotic) which results in ("nicer profile traces")
- and hence to obtain a **better quality of the Wald-type confidence intervals**

Parameter transformation does not chance either the deterministic nor the stochastic part of the regression model!

– in contrast to variable transformations.

Constraints of the Parameter Domain

Subject matter theory: Parameter domain is subject to constraints

e.g., $\theta_1 > 0$, $a < \theta_2 \leq b$

What to do?

Ignore the constraints and observe

- whether the algorithm converge and
- where to.

If this fails:

Most of the constraints are such that they can be imposed by a suitable transformation of the concerned parameter

Examples of Constraints

- $\theta > 0$: Trsf. $\theta \to \varphi = \log\langle\theta\rangle$ ☞ $\theta = \exp\langle\varphi\rangle > 0$ for all φ
  $h\langle x;\,\theta\rangle \to h\langle x;\,e^{\varphi}\rangle$

- $a < \theta < b$: Trsf. $\theta \to \varphi = \log\left\langle\frac{b-\theta}{\theta-a}\right\rangle$ ☞ $\theta = a + \frac{b-a}{1+\exp\langle\varphi\rangle}$

- Let $h\langle x;\,\underline{\theta}\rangle = \theta_1 e^{-\theta_2 x} + \theta_3 e^{-\theta_4 x}$ with $\theta_2,\,\theta_4 > 0$

  The two pairs of parameters $(\theta_1,\,\theta_2)$ and $(\theta_3,\,\theta_4)$ are exchangeable and may thus cause convergence problems

  Workaround: Impose the constraint $\theta_2 < \theta_4$!

  Trsf. $\underline{\theta} \to \underline{\varphi}$ with $\theta_1 = \varphi_1$, $\theta_2 = e^{\varphi_2}$, $\theta_3 = \varphi_3$, and $\theta_4 = e^{\varphi_2}\cdot(1 + e^{\varphi_4})$

  ☞ $h\left\langle x;\,(\theta_1,\varphi_2,\theta_3,\varphi_4)^T\right\rangle = \theta_1 \exp\langle -e^{\varphi_2}x\rangle + \theta_3 \exp\langle -e^{\varphi_2}\cdot(1+e^{\varphi_4})\cdot x\rangle$

Collinearity in Matrix $\boldsymbol{A}$

Example to show the problem: Let $h\langle x;\, \underline{\theta}\rangle = \theta_1 \cdot e^{-\theta_2 x}$ $(*)$

The partial derivatives (☞ matrix $\boldsymbol{A}$) are

$$\frac{\partial}{\partial\theta_1} h\langle x;\, \underline{\theta}\rangle = e^{-\theta_2 x} \qquad \frac{\partial}{\partial\theta_2} h\langle x;\, \underline{\theta}\rangle = -\theta_1 \cdot x \cdot e^{-\theta_2 x}$$

Hence

$$\underline{a}_1^T := (e^{-\theta_2 x_1}, \ldots, e^{-\theta_2 x_n})$$

$$\underline{a}_2^T := (-\theta_1 \cdot x_1 \cdot e^{-\theta_2 x_1}, \ldots, -\theta_1 \cdot x_n \cdot e^{-\theta_2 x_n})$$

The vectors $\underline{a}_1$ and $\underline{a}_2$ incline to collinearity if all $x_i > 0$.

Reformulate $(*)$: $h\langle x;\, \underline{\theta}\rangle = \theta_1 \cdot \exp\langle -\theta_2(x - x^* + x^*)\rangle$

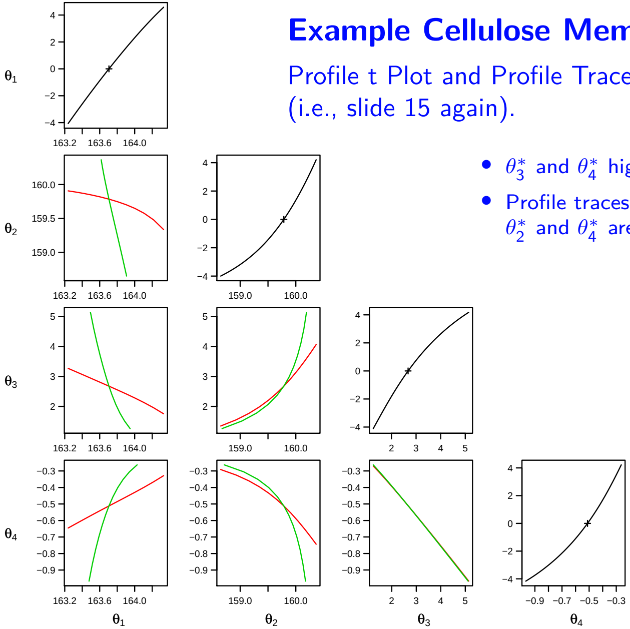
Applying the reparametrization $\varphi_1 := \theta_1 \cdot e^{-\theta_2 x^*}$ und $\varphi_2 := \theta_2$ we obtain

$$h\left\langle x, \underline{\varphi}\right\rangle = \varphi_1 \cdot \exp\left\langle -\varphi_2(x - x^*)\right\rangle .$$

This functions results in (approximately) optimal matrix $\boldsymbol{A}$ if $x^* = \bar{x}$ is chosen.

Example Cellulose Membrane (7)

Profile t Plot and Profile Traces (i.e., slide 15 again).



- θ_3^* and θ_4^* highly correlated
- Profile traces of θ_2^* and θ_3^* as well as θ_2^* and θ_4^* are twisted clearly

Example Cellulose Membrane (8)

Regression function

$$h\langle x,\, \underline{\theta}\rangle = \frac{\theta_1 + \theta_2 \cdot 10^{\theta_3+\theta_4((x_i-x^*)+x^*)}}{1+10^{\theta_3+\theta_4((x_i-x^*)+x^*)}}$$

Remove collinearity by introducing $\varphi_3 := \theta_3 + \theta_4 \cdot x^*$, where $x^* = \text{median}\langle x_i\rangle$:

$$h\langle x,\, \underline{\theta}\rangle = \frac{\theta_1 + \theta_2 \cdot 10^{\varphi_3+\theta_4\cdot(x_i-x^*)}}{1+10^{\varphi_3+\theta_4\cdot(x_i-x^*)}}$$

Improve linear approximation:
Step 1: Introduce $\varphi_4 := 10^{\theta_4}$:

$$h\langle x,\, \underline{\theta}\rangle = \frac{\theta_1 + \theta_2 \cdot 10^{\varphi_3} \cdot {\varphi_4}^{(x_i-x^*)}}{1+10^{\varphi_3} \cdot {\varphi_4}^{(x_i-x^*)}}$$

Step 2:

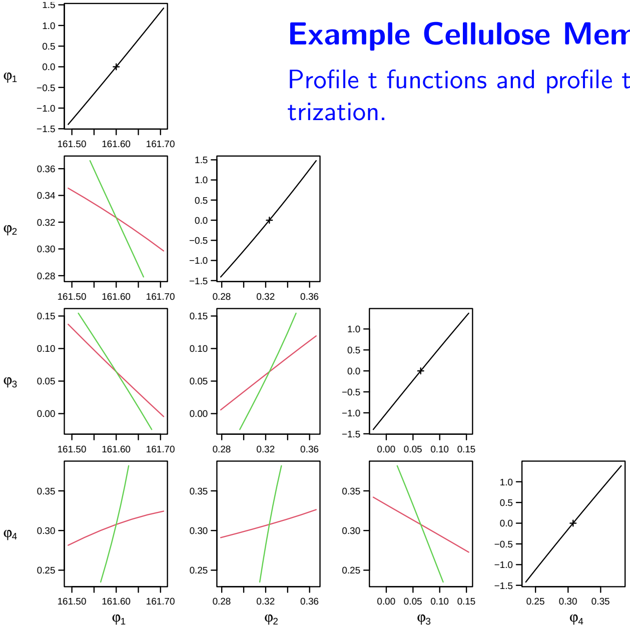
$$\varphi_1 := \frac{\theta_1 + \theta_2\, 10^{\varphi_3}}{10^{\varphi_3}+1}\,,$$

$$\varphi_2 := \log_{10}\left(\frac{\theta_1 - \theta_2}{10^{\varphi_3}+1}\, 10^{\varphi_3}\right)$$

$$h\langle x,\, \underline{\theta}\rangle = \varphi_1 + 10^{\varphi_2}\, \frac{1-{\varphi_4}^{(x_i-x^*)}}{1+10^{\varphi_3}\, {\varphi_4}^{(x_i-x^*)}}$$

Example Cellulose Membrane (9)

Profile t functions and profile traces after reparametrization.



Example Cellulose Membrane (10)

Original parametrization

Parameters:

	Value	Std. Error	t value
θ_1	163.706	0.1262	1297.21
θ_2	159.785	0.1594	1002.03
θ_3	2.675	0.3813	7.02
θ_4	-0.512	0.0703	-7.28

Residual standard error: 0.293137 on 35 df

Correlation of Parameter Estimates:

	θ_1	θ_2	θ_3
θ_2	-0.256		
θ_3	-0.434	0.771	
θ_4	0.515	-0.708	-0.989

Reparametrized

Parameters:

	Value	Std. Error	t value
φ_1	161.6001	0.0739	2187.12
φ_2	0.3234	0.0313	10.32
φ_3	0.0644	0.0595	1.08
φ_4	0.3077	0.0498	6.18

Residual standard error: 0.2931 on 35 df

Correlation of Parameter Estimates:

	φ_1	φ_2	φ_3
φ_2	-0.561		
φ_3	-0.766	0.641	
φ_4	0.151	0.354	-0.312

Successful Reparametrization

A **successful reparametrization** depends both

- on the **regression function** and
- on the **dataset**

☞ There are no general guidelines

which results in a tedious search for successful reparameterisations.

Another Example:

$$h\langle \underline{x}, \underline{\theta}\rangle = \frac{\theta_1\theta_3(x^{(2)} - x^{(3)})}{1 + \theta_2 x^{(1)} + \theta_3 x^{(2)} + \theta_4 x^{(3)}} \qquad (*)$$

$$= \frac{x^{(2)} - x^{(3)}}{\frac{1}{\theta_1\theta_3} + \frac{\theta_2}{\theta_1\theta_3}x^{(1)} + \frac{\theta_3}{\theta_1\theta_3}x^{(2)} + \frac{\theta_4}{\theta_1\theta_3}x^{(3)}}$$

$$= \frac{x^{(2)} - x^{(3)}}{\phi_1 + \phi_2 x^{(1)} + \phi_3 x^{(2)} + \phi_4 x^{(3)}} \qquad (**)$$

The parametrization $(**)$ is preferd to $(*)$ in most cases (cf. exercises).

Interpretation?

In most cases, the original parameters have a physical interpretation

☞ **parameter must be back-transformed**

Standard approach for back-transformation:

Example: Used parameter transformation: $\theta \longrightarrow \phi = \log\langle\theta\rangle$

Let $\widehat{\phi}$ and $\widehat{\sigma}_{\widehat{\phi}}$ the estimated parameters.

Estimate θ by $\widehat{\theta} = \exp\left\langle\widehat{\phi}\right\rangle$. Its standard error is obtained commonly by **Gaussian error propagation rule** (cf. Stahel, Sec 6.10):

$$\widehat{\sigma}^2_{\widehat{\theta}} \approx \left(\left. \frac{\partial \exp\langle\phi\rangle}{\partial\phi} \right|_{\phi=\widehat{\phi}} \right)^2 \widehat{\sigma}^2_{\widehat{\phi}} = \left(\exp\left\langle\widehat{\phi}\right\rangle\right)^2 \widehat{\sigma}^2_{\widehat{\phi}} \qquad ☞ \quad \widehat{\sigma}_{\widehat{\theta}} \approx \exp\left\langle\widehat{\phi}\right\rangle \widehat{\sigma}_{\widehat{\phi}} .$$

Hence, an approximate 95% confidence interval for θ is:

$$g\langle\widehat{\phi}\rangle \pm \widehat{\sigma}_{\widehat{\theta}}\, q_{0.975}^{t_{n-p}} = \exp\langle\widehat{\phi}\rangle \left(1 \pm \widehat{\sigma}_{\widehat{\phi}}\, q_{0.975}^{t_{n-p}}\right) . \quad (*)$$

But this approach is not recommended because ... see next slide

Why Parameter Transformation?

1. **so that the parameter falls within a predefined domain.**
   Confidence intervals according to (*) may violate this requirement!

2. **due to the insufficient quality of the confidence interval**
   Gaussian error propagation rule will nullify the achievements by the reparametrization since it uses the **same linear approximation** as the Wald-type asymptotic!

Alternative to the standard approach:

- Back-transformation of the complete confidence interval;
  Example:

$$\left\{\theta \,:\, \log\left\langle\theta\right\rangle \in \widehat{\phi} \pm \widehat{\sigma}_{\widehat{\phi}}\, q_{0.975}^{t_{df}}\right\}$$

  forms a better, but still approximate 95% confidence interval for θ. It is identical to

$$= \left[\exp\left\langle\widehat{\phi} - \widehat{\sigma}_{\widehat{\phi}}\, q_{0.975}^{t_{df}}\right\rangle ,\, \exp\left\langle\widehat{\phi} + \widehat{\sigma}_{\widehat{\phi}}\, q_{0.975}^{t_{df}}\right\rangle\right] ,$$

  since $\log\left\langle\right\rangle$ and $\exp\left\langle\right\rangle$ are strictly increasing.

- In case of bullet point 2, the most convenient approach is to form the confidence interval based on the **profile t function.**

Take Home Message Half-Day 2

- The commonly used **confidence intervals** are based on a (crude) linear **approximation**.
- Use **graphical tools like profile t plots and profile traces** to assess the quality of the approximated confidence intervals (and hence the linear approximation).
- If insufficient:
  More accurate confidence intervals can be calculated for single parameters θ_k by using **profile t functions** (as in confint() implemented anyway).
- Convergence properties of the estimating algorithm and the quality of the Wald-type conference intervals can be improved by applying **suitable reparametrizations** (parameter transformations).

  If the interpretation of the original parameters is crucial, then the confidence interval should also be backtransformed

  and not be determined by Gaussian error propagation rule.