# Lösungsskizze Serie 6

**1.** Es ist

$$\bar{\theta}^{(-k)} = \arg \min_{(\theta_j; j \neq k)} S(\theta)$$

Im linearen Fall hat man:

$$\bar{\theta}^{(-k)} = \arg \min_{\theta; \theta_k = \theta_k^0} \| y - X\theta \|^2$$

D.h. man minimiert $\| y - X\theta \|^2$ unter der Nebenbedingung $\theta_k - \theta_k^0 = 0$. Also führt man den Lagrangemultiplikator $\lambda$ ein und bekommt:

$$\bar{\theta}^{(-k)} = \arg \min_{\theta, \lambda} \{ \| y - X\theta \|^2 + \lambda(\theta_k - \theta_k^0) \}$$

Durch ableiten bekommt man:

$$X^T y - (X^T X)\bar{\theta}^{(-k)} = \lambda e_k$$

und somit

$$\bar{\theta}^{(-k)} = \underbrace{(X^T X)^{-1} X^T y}_{= \hat{\theta}} - (X^T X)^{-1} \lambda e_k$$

also

$$\lambda = -\frac{(\hat{\theta}_k - \theta_k^0)}{((X^T X)^{-1})_{kk}}$$

und komponentenweise geschrieben, folgt

$$\bar{\theta}_j^{(-k)}(\theta_k) = \hat{\theta}_j - \frac{((X^T X)^{-1})_{jk}}{((X^T X)^{-1})_{kk}} (\hat{\theta}_k - \theta_k)$$

Und somit bekommen wir:

$$\begin{aligned}
\bar{S}_k(\theta_k) &= \min_{\theta_j; j \neq k} S(\theta) \\
&= \min_{\theta; \theta_k \text{ fest}} \| y - X\theta \|^2 \\
&= \| y - X\bar{\theta}^{(-k)}(\theta_k) \|^2 \\
&= \| y - X\hat{\theta} + X(\hat{\theta} - \bar{\theta}^{(-k)}(\theta_k) \|^2 \\
&= \| y - X\hat{\theta} \|^2 + \| X(\hat{\theta} - \bar{\theta}^{(-k)}(\theta_k)) \|^2 \qquad (*) \\
&= S(\hat{\theta}) + \| X(X^T X)^{-1} \lambda e_k \|^2 \\
&= S(\hat{\theta}) + e_k^T \lambda (X^T X)^{-1} \lambda e_k \\
&= S(\hat{\theta}) + \frac{\hat{\theta}_k - \theta_k}{(X^T X)_{kk}^{-1}}
\end{aligned}$$

Gleichheit in $(*)$ gilt, da $y - X\hat{\theta}$ orthogonal zu $X$ ist.



**2.** **a)** Dieses Modell geht davon aus, dass der Achsenabschnitt $\theta_1$ und die horizontale Asymptote $\theta_1 + \theta_2$ für beide Methionin-Nahrungsmittelzusätze gleich sind und dass sie sich nur in der Zunahmerate $\theta_3$, resp. $\theta_3\theta_4$ unterscheiden.
Wollen wir die Nullhypothese beantworten, ob sich die beiden Zunahmeraten unterscheiden, müssen wir testen ob $\theta_4$ signifikant von 1 verschieden ist.
*Hinweis:* Aus der Form der Kurve können wir erwarten, dass $\theta_3 < 0$ und $\theta_4 > 0$ ist.

**b)** Mit den gegebenen Startwerten bekommen wir:

```
> truthennen <- read.table.url("http://stat.ethz.ch/Teaching/
                                        Datasets/body.dat",header=TRUE)
> library(nls)
> truthennen.nls<-nls(weight~T1+T2*(1-exp(T3*(T4*sourceA + sourceB))),
                  data=truthennen,start=list(T1=640,T2=160,T3=-7.2,T4=0.9))
> summary(truthennen.nls)
Formula: weight ~ T1 + T2 * (1 - exp(T3 * (T4 * sourceA + sourceB)))

Parameters:
    Estimate Std. Error t value Pr(>|t|)
T1 638.83908    6.58766  96.975 8.10e-11 ***
T2 175.90406    6.21084  28.322 1.28e-07 ***
T3  -6.38720    0.80191  -7.965 0.000208 ***
T4   0.79107    0.04889  16.182 3.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.942 on 6 degrees of freedom

Correlation of Parameter Estimates:
           T1      T2     T3
T2 -0.4073477
T3  0.7679316 0.17643
T4 -0.0009674 0.07078 0.2910
```

Die Schätzungen für die Koeffizienten lauten also

$$\widehat{\theta_1} = 638.8\,, \quad \widehat{\theta_2} = 175.9\,, \quad \widehat{\theta_3} = -6.39\,, \quad \widehat{\theta_4} = 0.79\,.$$

**c)** Für den Parameter $\theta_4$ erhalten wir ein approximatives 95%-Konfidenzintervall, gegeben durch

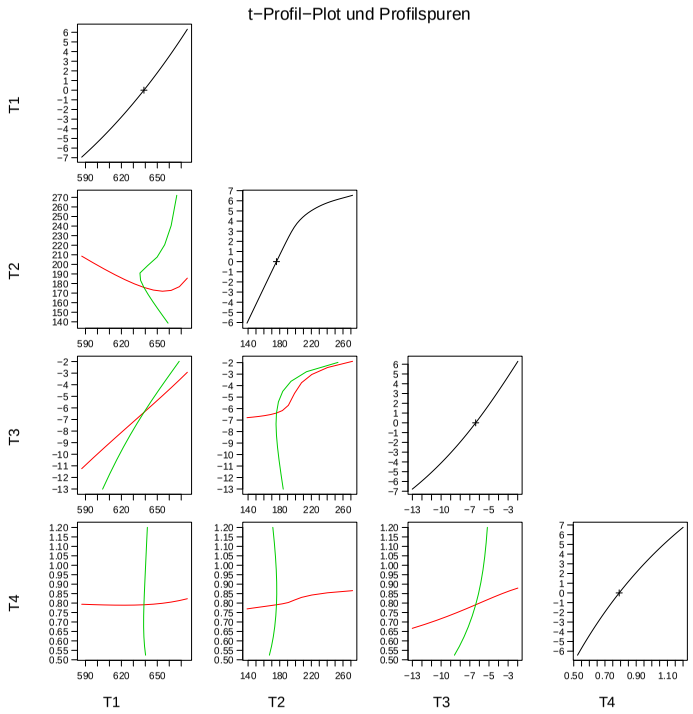
$$\widehat{\theta_4} \pm \hat{\sigma}(\hat{\theta}_4) \cdot t_{0.975,6} = 0.79 \pm 0.0489 \cdot 2.45 = [0.67, 0.91]\,.$$

Da 1 nicht in diesem Konfidenzintervall liegt, ist $\theta_4$ signifikant von 1 verschieden, d.h., die Zunahmeraten unterscheiden sich signifikant.

**d)** `> library(SfS)` oder
`> source(''/'Pfadname des Files'/p.profileTraces.R'')`
`> truthennen.profile <- profile(truthennen.nls)`
`> p.profileTraces(truthennen.profile)`
Die Likelihood-Profilspuren zeigen, dass die Regressionsfunktion zu einer starken Nichtlinearität neigt. Zudem sieht man, dass die Parameter $\theta_1$ und $\theta_3$ ziemlich stark korreliert sind, während dem die andern nur schwach oder gar nicht. Aus den Profil-t-Funktionen sieht man, dass die lineare Approximation in der Nähe der KQ-Lösung (mit + eingezeichnet) gut passt.

**3.** **a)**

```
> heart<-read.table.url("http://stat.ethz.ch/Teaching/Datasets/heart.dat",
                                                        header=TRUE)
> heart.glm <-  glm(cbind(y,n-y)~age, family=binomial,data=heart)
> summary(heart.glm)
 Call:
 glm(formula = cbind(y, n - y) ~ age, family = binomial, data = heart)
```




```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.36404  -0.54657   0.02464   0.55254   1.53530

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.09918    1.10500  -4.615 3.94e-06 ***
age          0.10839    0.02372   4.571 4.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.466  on 42  degrees of freedom
Residual deviance: 25.153  on 41  degrees of freedom
AIC: 63.888

Number of Fisher Scoring iterations: 3
```

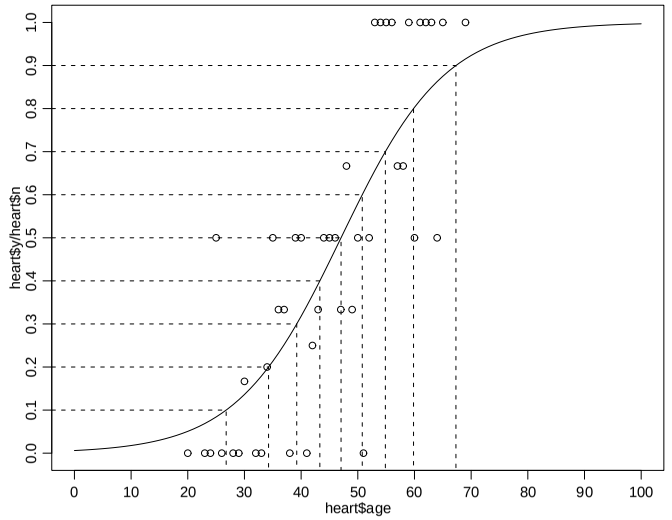
Wir bekommen somit: $\widehat{\alpha} = -5.10$ und $\widehat{\beta} = 0.11$. Der Einfluss des Alters ist signifikant. Das positive Vorzeichen von $\widehat{\beta}$ bedeutet, dass die Wahrscheinlichkeit, Symptome zu zeigen, mit dem Alter zunimmt. Die Grösse des Koeffizienten kann wegen der logit-Skala nicht so einfach interpretiert werden.

**b)**
```
> age.neu <- 0:100
```



```
> heart.pred <- predict(heart.glm,newdata=data.frame(age=age.neu),
                                                    type="response")
> plot(heart$age, heart$y/heart$n, xlim=c(0,100), ylim=c(0,1))
> lines(age.neu, heart.pred)
> perc <- (1:9)/10
> x.age <- (log(perc/(1-perc)) - coef(heart.glm)[1])/coef(heart.glm)[2]
> names(x.age) <- perc
> round(x.age)
 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
  27  34  39  43  47  51  55  60  67

> for(n in 1:9)
    lines(c(-4,x.age[n],x.age[n]), c(perc[n],perc[n],-0.04), lty=2)
```




Aus dem Modell folgt nämlich

$$x_i = \frac{\log\left(\frac{\pi_i}{1-\pi_i}\right) - \alpha}{\beta}$$

Wenn wir $\pi_i = 0.1, 0.2, \ldots, 0.9$ wählen und anstelle von $\alpha$ und $\beta$ die geschätzten Koeffizienten einsetzen, erhalten wir gerade das Alter, bei dem wir erwarten würden, dass 10%, 20%, ..., 90% der Personen Symptome zeigen.

| Symptome | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|---|
| Alter | 27 | 34 | 39 | 43 | 47 | 51 | 55 | 60 | 67 |

Zwischen 39 und 55 Jahren nimmt die Wahrscheinlichkeit, Symptome zu zeigen, alle 4 Jahre um etwa 10% zu. Vorher und nachher nimmt die Wahrscheinlichkeit weniger schnell zu. Ab 67 Jahren kann man erwarten, dass über 90% Symptome zeigen.

**4. a)**
```
> diabetes<-read.table.url("http://stat.ethz.ch/Teaching/Datasets/
                                                    diabetes.dat",header=TRUE)
```

```
> library(survival5)
> diabetes.cox <- coxph(Surv(lzeit,cens)~sex + diab + alter, data=diabetes)
> summary(diabetes.cox)
Call:
coxph(formula = Surv(lzeit, cens) ~ sex + diab + alter, data = diabetes)

  n= 66

        coef exp(coef) se(coef)     z     p
sex   0.1402      1.15   0.3400 0.412 0.680
diab  0.5302      1.70   0.3507 1.512 0.130
alter 0.0293      1.03   0.0133 2.200 0.028

      exp(coef) exp(-coef) lower .95 upper .95
sex        1.15      0.869     0.591      2.24
diab       1.70      0.588     0.855      3.38
alter      1.03      0.971     1.003      1.06

Rsquare= 0.151   (max possible= 0.99 )
Likelihood ratio test= 10.8  on 3 df,   p=0.013
Wald test            = 9.81  on 3 df,   p=0.0202
Score (logrank) test = 10.2  on 3 df,   p=0.0171
```

Nur das Alter scheint signifikant zu sein. Die Tatsache Diabetes zu haben, scheint keinen Einfluss auf die Überlebenszeit zu haben. Da Koeffizient von `alter`$= 0.029 > 0$, nimmt also die "Ausfallrate" zu, d.h. die Wahrscheinlichkeit eines Patienten nach der Operation zu sterben, nimmt mit zunehmendem Alter zu.

**b)**
```
> diabetes.cox2 <- coxph(Surv(lzeit,cens)~sex + diab + alter + alter*diab,
                                                              data=diabetes)
> summary(diabetes.cox2)
Call:
coxph(formula = Surv(lzeit, cens) ~ sex + diab + alter + alter *
    diab, data = diabetes)

  n= 66

               coef exp(coef) se(coef)      z     p
sex         0.1968     1.217   0.3428  0.574 0.570
diab        3.3108    27.407   2.0400  1.623 0.100
alter       0.0558     1.057   0.0244  2.281 0.023
diab.alter -0.0421     0.959   0.0300 -1.401 0.160

            exp(coef) exp(-coef) lower .95 upper .95
sex             1.217     0.8214     0.622      2.38
diab           27.407     0.0365     0.503   1493.93
alter           1.057     0.9458     1.008      1.11
diab.alter      0.959     1.0430     0.904      1.02

Rsquare= 0.177   (max possible= 0.99 )
Likelihood ratio test= 12.9  on 4 df,   p=0.0119
Wald test            = 8.74  on 4 df,   p=0.068
Score (logrank) test = 10.4  on 4 df,   p=0.0338
```

Die zusätzlich eingeführte Variable `alter·diab` scheint nichts zu bringen. Sieht man



sich aber die Koeffizienten genauer an, sieht man, dass der Koeffizient von `alter`$\approx -$`alter·diab` ist, sich also im Modell aufheben. Man vermutet somit, dass bei einem Patienten, der Diabetes hat, nur die Tatsache dass er Diabetes hat, Einfluss darauf hat, ob er stirbt oder nicht, nicht aber sein Alter. Die Wahrscheinlichkeit das ein Patient stirbt ist für solche mit Diabetes gleich konstant, während bei denen ohne, die Wahrscheinlichkeit mit dem Alter zunimmt.