Angewandte Regression — Musterlösungen zur Serie 8

1. **a)** Siehe Teilaufgabe d).

b) Wir betrachten das Modell

$$\frac{exp(\tilde{\eta_k})}{1+exp(\tilde{\eta_k})} = \pi_k = \mathcal{E}[\frac{\tilde{Y_k}}{m_k}]$$

wobei $\tilde{Y_k} = \sum_{i:x_i=\tilde{x_k}} Y_i$ die im Dataframe unter `y` angegebenen gruppierten Beobachtungen sind. D.h. $\tilde{Y_k} \sim \mathcal{B}(m_k, \pi_k)$, wobei die in der Variablen `m` angegebenen Werte den m_k's entsprechen. Die Altersgruppe (`age`) entspricht $\tilde{x_k}$. Es ist also

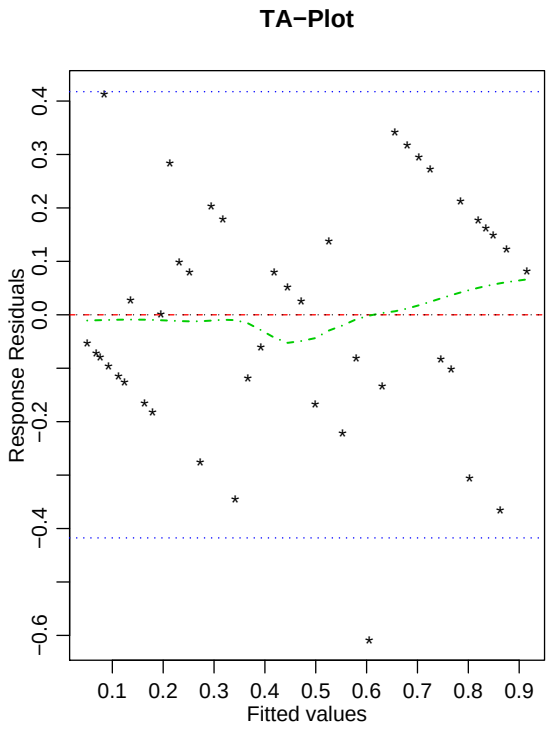
$$\tilde{\eta_k} = logit(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \beta_0 + \beta_1\tilde{x_k}$$

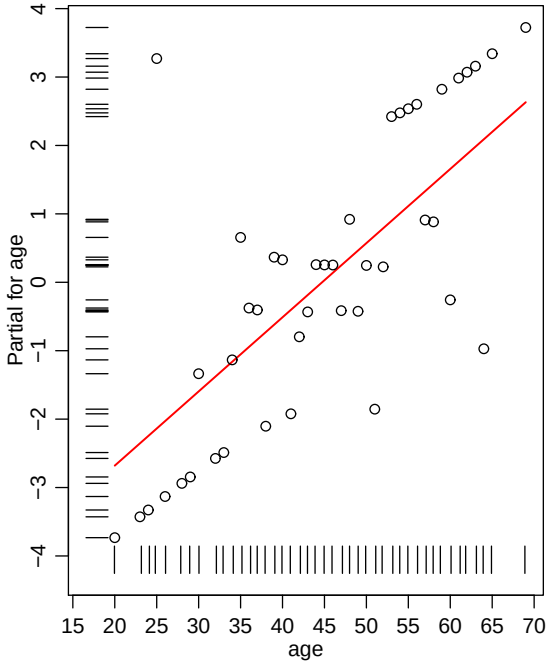
und wir schätzen die Koeffizienten β_0 und β_1.
R-Output von `summary(r.glm.regr)`:

```
Call:
regr(formula = cbind(y, m - y) ~ age, data = d.heart, method = "glm",
    family = "binomial")

Terms:
                  coef   stcoef    signif R2.x df p.value
(Intercept) -5.0992974 0.000000 -2.276854   NA  1       0
age          0.1083923 1.433178  2.255391    0  1       0

         deviance df      p.value
Model    28.31267  1 1.032183e-07
Residual 25.15345 41           NA
Null     53.46612 42           NA

Family is binomial.  Dispersion parameter taken to be 1.
AIC: 63.888
```

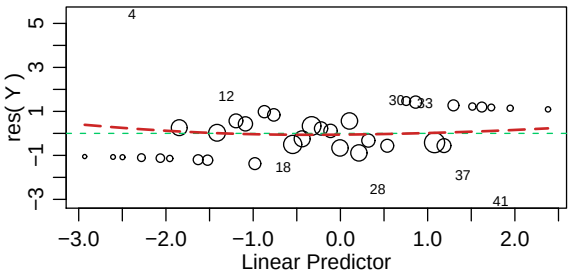
Der Einfluss des Alters ist signifikant. Das positive Vorzeichen von $\widehat{\beta_1}$ bedeutet, dass die Wahrscheinlichkeit, Symptome zu zeigen mit dem Alter zunimmt.
Bemerkung: Falls Sie mit `glm` arbeiten, wird für die Koeffizienten ein Wald-Test durchgeführt. Bei `regr` wird ein Likelihood-Quotienten-Test durchgeführt (siehe R-Skript zum Block "Verallgemeinerte lineare Modelle"). Der Wald-Test hat eine geringe Macht, der Likelihood-Quotienten-Test ist daher vorzuziehen.

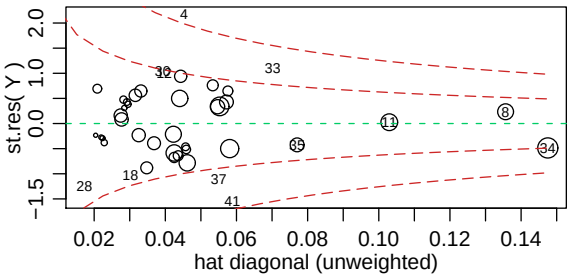
c) Der Tukey-Anscombe-Plot und der Partial-Residuals-Plot zeigen keine Abweichung von der Linearitätsannahme im Modell. Es ist aber ein Ausreisser (links, oben) erkennbar. Im von `plot.regr` gezeichneten Termplot (Bild unten) können wir zusätzlich noch die Grösse des Koeffizienten, die Gewichtung der Punkte und die Beobachtungsnummern der extremen Werte herauslesen.

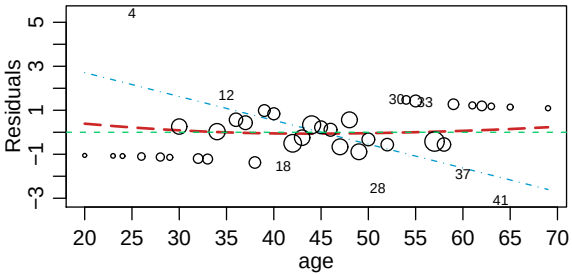




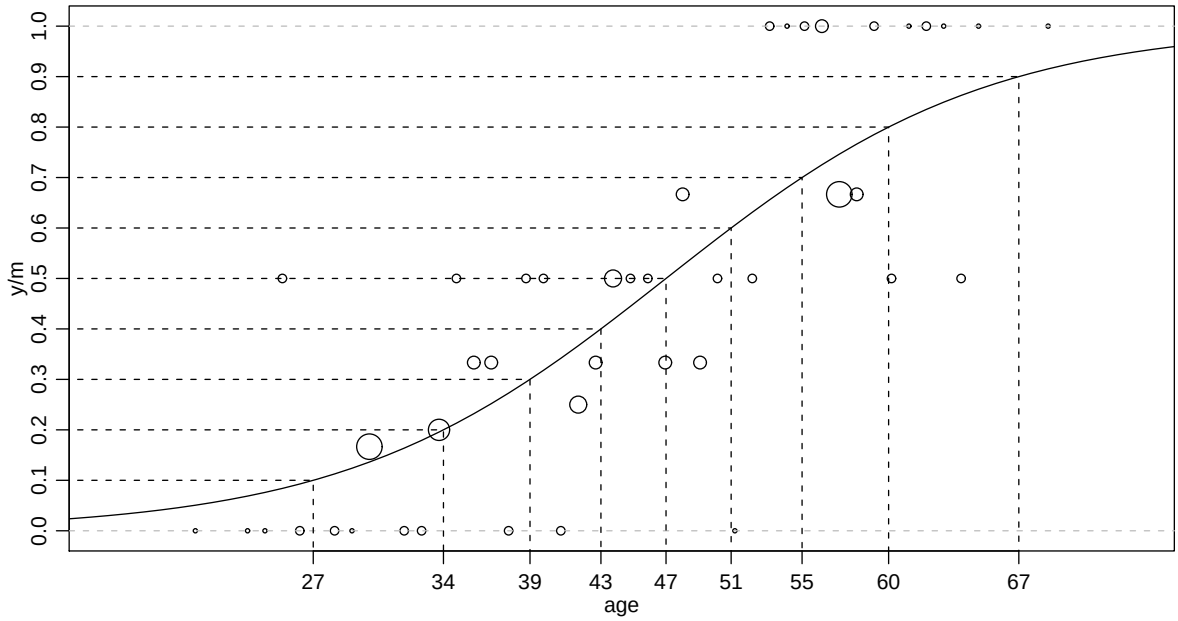
Mit `plot(r.glm.regr)` erhalten wir:







d) Die Antwort kann direkt aus der Graphik abgelesen oder berechnet werden:



Aus dem Modell folgt nämlich

$$x_k = \frac{\log\left(\frac{\pi_k}{1-\pi_k}\right) - \beta_0}{\beta_1}$$

Wenn wir $\pi_k = 0.1, 0.2, \ldots, 0.9$ wählen und anstelle von β_0 und β_1 die geschätzten Koeffizienten einsetzen, erhalten wir gerade das Alter, bei dem wir erwarten würden, dass 10%, 20%, ..., 90% der Personen Symptome zeigen.

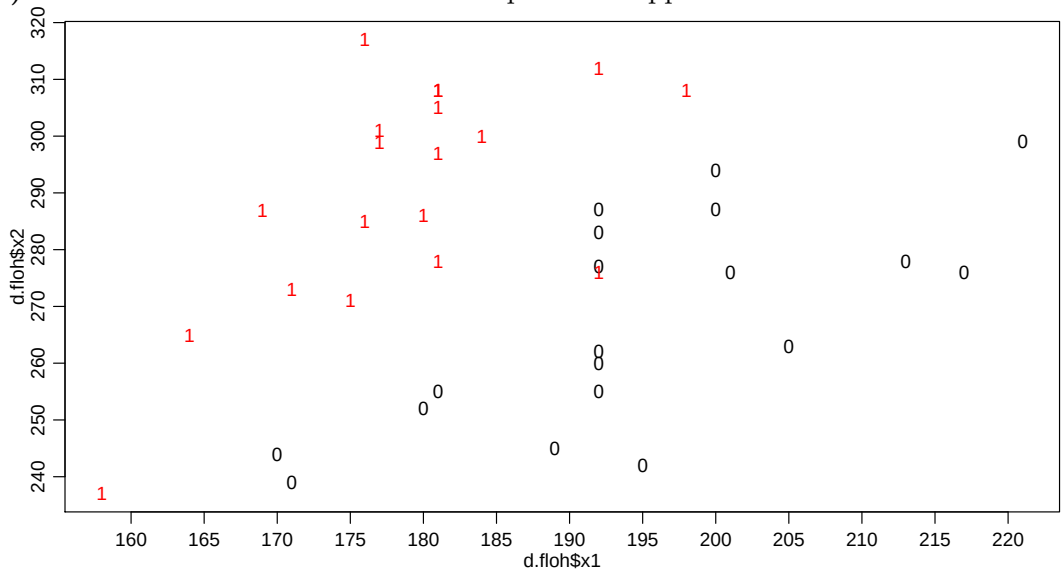
Symptome	10%	20%	30%	40%	50%	60%	70%	80%	90%
Alter	27	34	39	43	47	51	55	60	67

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.

Erstellen der Graphik mit R:

```
age.neu <- 0:100
r.pred <- predict(r.glm.regr, newdata=data.frame(age=age.neu),
                  type="response")
symbols(d.heart$age, d.heart$y/d.heart$m, circles=d.heart$m,
        inches=0.1, xlab="age", ylab="y/m", ylim=c(0,1), xaxt="n")
abline(h=0, lty=2, col="grey")
abline(h=1, lty=2, col="grey")
lines(age.neu, r.pred)
t.p <- (1:9)/10
r.age <- (log(t.p/(1-t.p))-coef(r.glm.regr)[1])/coef(r.glm.regr)[2]
names(r.age) <- t.p
for (n in 1:9)
     lines(c(-4,r.age[n],r.age[n]),c(t.p[n],t.p[n],-0.4), lty=2)
axis(1, at=r.age, labels=round(r.age))
```

2. **a)** Die beiden Arten sind deutlich als separate Gruppen erkennbar.



b) Das logistische Regressionsmodell lautet

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot \texttt{x1}_i + \beta_2 \cdot \texttt{x2}_i + \beta_3 \cdot \texttt{unt}_i\,, \tag{1}$$

wobei $\pi_i = E[\texttt{art}_i] = P[\texttt{art}_i = 1|\texttt{x1}_i, \texttt{x2}_i, \texttt{unt}_i]$ die Wahrscheinlichkeit ist (gegeben die Beobachtungen), dass der Käfer zur Art *Carduorum* gehört.

c) Die Variable `unt` ist auf dem 5%-Niveau nicht signifikant. Berechnung mit R:

```
> r.glm <- glm(art ~ x1+x2+unt, family=binomial, data=d.floh)
> summary(r.glm)
Call:
glm(formula = art ~ x1 + x2 + unt, family = binomial, data = d.floh)
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.543402  -0.059469  -0.000277   0.074273   2.060227

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.7971    14.9056   0.322   0.7476
x1           -0.4131     0.1688  -2.447   0.0144 *
x2            0.2567     0.1069   2.401   0.0163 *
unt           1.6864     2.2616   0.746   0.4559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.6792  on 37  degrees of freedom
Residual deviance:  8.8958  on 34  degrees of freedom
AIC: 16.896

Number of Fisher Scoring iterations: 7
```

Die Untersucher-Variable `unt` kann somit weggelassen und das Modell (**??**) reduziert werden zu

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \cdot \texttt{x1}_i + \beta_2 \cdot \texttt{x2}_i\,. \tag{2}$$

In diesem Modell sind die Variablen `x1` und `x2` immer noch signifikant:

```
> r.glm2 <- glm(art ~ x1+x2, family=binomial, data=d.floh)
> summary(r.glm2)
 Call:
 glm(formula = art ~ x1 + x2, family = binomial, data = d.floh)

 Deviance Residuals:
        Min          1Q      Median          3Q         Max
 -1.1076756  -0.0666525  -0.0002102   0.0656085   2.3673292

 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  11.2148    12.8560   0.872   0.3830
 x1           -0.4090     0.1714  -2.386   0.0170 *
 x2            0.2339     0.1006   2.325   0.0201 *
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 52.6792  on 37  degrees of freedom
 Residual deviance:  9.5138  on 35  degrees of freedom
 AIC: 15.514
 Number of Fisher Scoring iterations: 7
```

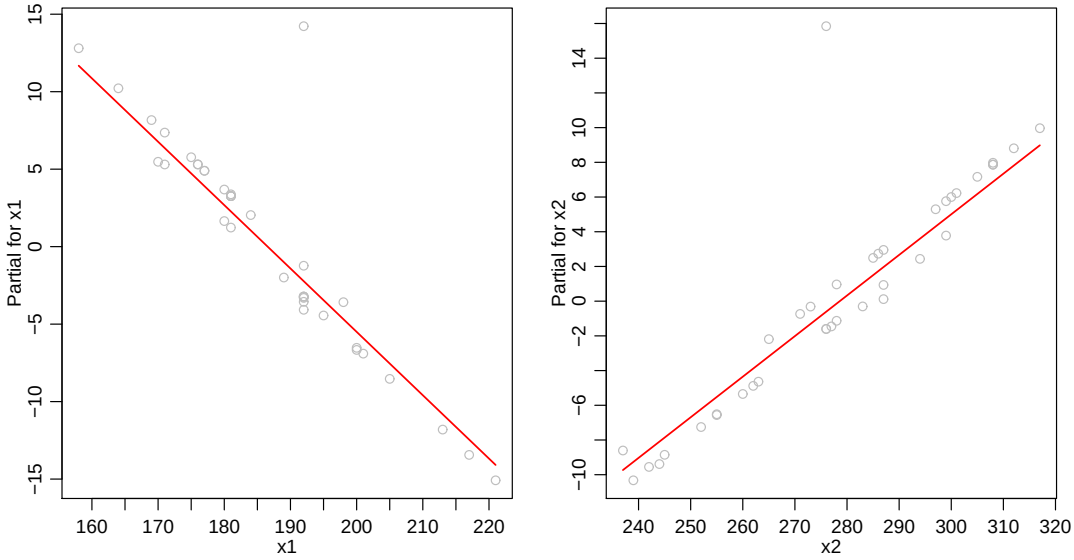
Der Modellvergleich liefert (Vgl. Skript, Kap. 1.3) Folgendes: Die Devianz-Differenz beträgt $9.51 - 8.90 = 0.61$. Diese Realisierung der Testgrösse ist unter H_0 : (*Das kleinere Modell ist richtig*) ${\chi^2}_{4-3}$-verteilt. Man erhält mit dem R-Befehl `1-pchisq(9.51 - 8.90,4 - 3)` den p-Wert 0.43 und wird also die Nullhypothese beibehalten und davon ausgehen, dass das kleinere Modell `r.glm2` genügt. Diesen Test kann man übrigens auch direkt mit R durchführen:

```
> anova(r.glm, r.glm2, test="Chi")
```

R-Output:

```
 Analysis of Deviance Table
Model 1: art ~ x1 + x2 + unt
Model 2: art ~ x1 + x2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        34     8.8958
2        35     9.5138 -1  -0.6180    0.4318
```

d) Die Zielgrosse im Modell (**??**) hängt linear von den erklärenden Variablen `x1` und `x2` ab. Ausserdem gibt es einen deutlich erkennbaren Ausreisser.



Bemerkung: Da hier ungruppierte Daten vorliegen, ist der Tukey-Anscombe Plot schwierig zu interpretieren. Wir verzichten deshalb darauf, ihn zu zeichnen.

e) Die Wahrscheinlichkeit, dass der Käfer mit $\texttt{x1} = 197$ und $\texttt{x2} = 303$ zur Art *Carduorum* gehört, lässt sich mit folgender Formel berechnen:

$$\begin{aligned}\widehat{\pi}_i &= \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot \texttt{x1}_i + \widehat{\beta}_2 \cdot \texttt{x2}_i)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot \texttt{x1}_i + \widehat{\beta}_2 \cdot \texttt{x2}_i)} \\ &= \frac{\exp(11.215 - 0.409 \cdot 197 + 0.234 \cdot 303)}{1 + \exp(11.215 - 0.409 \cdot 197 + 0.234 \cdot 303)} = 0.82\,.\end{aligned}$$

Da $\widehat{\pi}_i > 0.5$ ist, wird der Käfer also der Art *Carduorum* zugeordnet.
Mit dem R-Befehl `predict(r.glm2, newdata=data.frame(x1=197,x2=303), type="response")` erhalten wir selbstverständlich das gleiche.