M7222 – 2. CVIČENÍ : **GLM02b** (*Beetle Mortality*)

Protože jde o velmi malý soubor provedeme jeho načtení pomocí příkazu `matrix()` a postupně vytvoříme datový rámec.

```
> mdata <- matrix(c(1.6907, 6, 59, 1.7242, 13, 60, 1.7552,
     18, 62, 1.7842, 28, 56, 1.8113, 52, 63, 1.8369, 53,
     59, 1.861, 61, 62, 1.8839, 60, 60), ncol = 3, byrow = T)
> colnames(mdata) <- c("dose", "killed", "population")
> (data <- data.frame(mdata))
```

```
    dose killed population
1 1.6907      6         59
2 1.7242     13         60
3 1.7552     18         62
4 1.7842     28         56
5 1.8113     52         63
6 1.8369     53         59
7 1.8610     61         62
8 1.8839     60         60
```

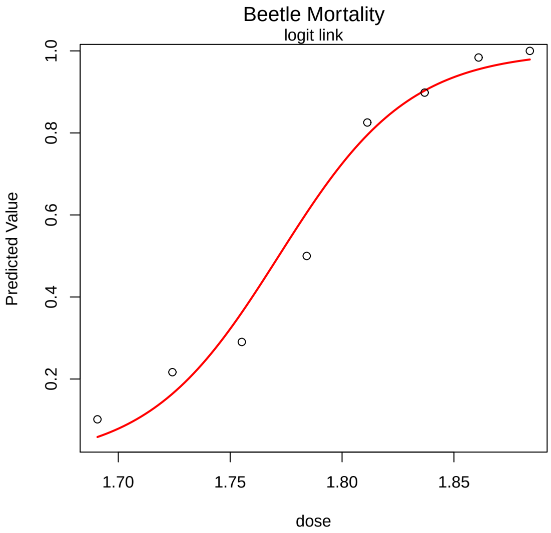
Protože nemáme k dispozici nula–jedničková data, budeme muset proceduru glm použít trochu jiným způsobem. Závisle proměnnou budou tvořit dva sloupce, v prvním bude počet zemřelých jedinců, ve druhém bude počet jedinců, kteří přežili. Jako první zvolíme logit linkovací funkci.

```
> m1.logit <- glm(cbind(killed, population - killed) ~
     dose, data = data, family = binomial(logit))
> summary(m1.logit)
```

```
Call:
glm(formula = cbind(killed, population - killed) ~ dose, family = binomial(logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

Tentokrát použijeme pro vykreslení výsledku příkaz `Predict.Plot()` z knihovy `TeachingDemos`. Do grafu navíc vykreslíme relativní četnosti brouků, kteří danou dávku nepřežili.

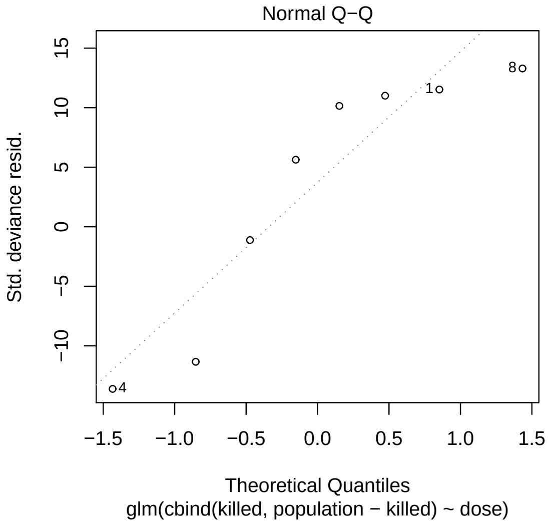
```
> library(TeachingDemos)
> Predict.Plot(m1.logit, pred.var = "dose", type = "response",
     plot.args = list(col = "red", lwd = 2))
> points(data$dose, data$killed/data$population)
> mtext("Beetle Mortality", side = 3, line = 1, cex = 1.25)
> mtext("logit link", side = 3, line = 0, cex = 1)
```



Obrázek 1: Logistická křivka spolu s relativními četnostmi.

Pro tento model opět provedeme grafickou analýzu reziduí a vykreslíme ROC křivku.

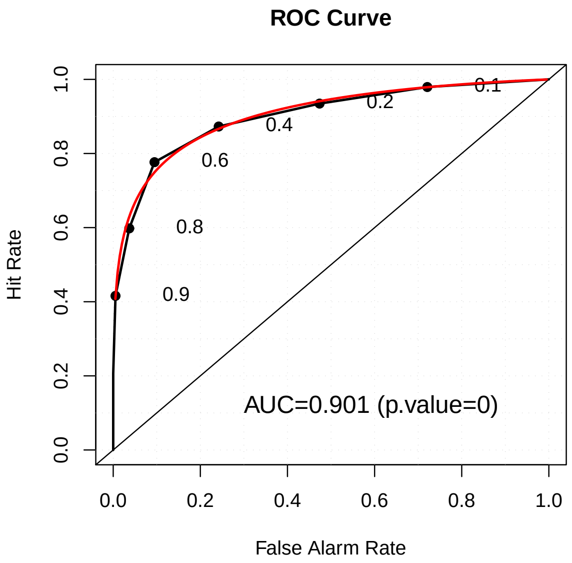
```
> plot(m1.logit, which = 2, cex = 0.75)
```



Obrázek 2: Q-Q graf reziduí s logit linkovací funkcí.

Jestliže budeme chtít vykreslit ROC křivku, budeme muset přejít z kumulovaných dat na nula-jedničková data.

```
> library(verification)
> N0 <- sum(data$population - data$killed)
> N1 <- sum(data$killed)
> par(mar = c(5, 5, 3, 0) + 0.1)
> binvar <- c(rep(0, N0), rep(1, N1))
> T <- c(rep(fitted(m1.logit), data$population - data$killed),
     rep(fitted(m1.logit), data$killed))
> AUC <- roc.area(binvar, T)
> auc.txt <- paste("AUC=", round(AUC$A, 3), " (p.value=",
     round(AUC$p.value, 10), ")", sep = "")
> roc.plot(binvar, T, binormal = T, plot = "both")
> text(0.3, 0.1, auc.txt, adj = c(0, 0), cex = 1.25)
```



Obrázek 3: ROC křivka a hodnota AUC pro replikovaná data (pomocí příkazů `roc.area` a `roc.plot` z knihovny `verification`).

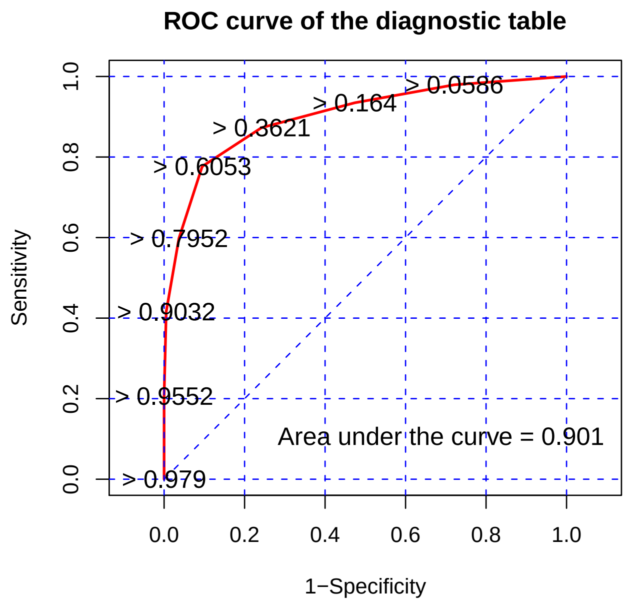
Jinou možností je použít příkaz `roc.from.table()` z knihovy `epicalc`, který dokáže pracovat se vstupní tabulkou

```
> library(epicalc)
> tableROC <- as.table(cbind(data$population - data$killed,
     data$killed))
> colnames(tableROC) <- c("non-killed", "killed")
> rownames(tableROC) <- as.character(round(fitted(m1.logit),
     digits = 4))
> par(mar = c(5, 5, 3, 0) + 0.1)
> roc.from.table(tableROC, title = TRUE, auc.coords = c(0.25,
     0.1), cex = 1.2, lwd = 2)
```

```
$auc
[1] 0.9010852
```

```
$original.table
       Non-diseased Diseased
0.0586           53        6
0.164            47       13
0.3621           44       18
0.6053           28       28
0.7952           11       52
0.9032            6       53
0.9552            1       61
0.979             0       60

$diagnostic.table
         1-Specificity Sensitivity
           1.000000000   1.0000000
> 0.0586   0.721052632   0.9793814
> 0.164    0.473684211   0.9347079
> 0.3621   0.242105263   0.8728522
> 0.6053   0.094736842   0.7766323
> 0.7952   0.036842105   0.5979381
> 0.9032   0.005263158   0.4158076
> 0.9552   0.000000000   0.2061856
> 0.979    0.000000000   0.0000000
```

```
> roc1 <- roc.from.table(tableROC, graph = FALSE)
> cut.points <- rownames(roc1$diagnostic.table)
> text(x = roc1$diagnostic.table[, 1], y = roc1$diagnostic.table[,
     2], labels = cut.points, cex = 1.2)
```



Obrázek 4: ROC křivka a hodnota AUC pro binomická data (pomocí příkazů `roc.from.table` z knihovny `epicalc`).

Ukážeme si, že stejnou ROC křivku dostaneme jak pro hodnoty `fitted(m1.logit)`, což jsou odhady

$$\widehat{\pi}_1 = \widehat{\pi}(x_1) = \widehat{Y}_1/n_1, \ldots, \widehat{\pi}_N = \widehat{\pi}(x_N) = \widehat{Y}_N/n_N,$$

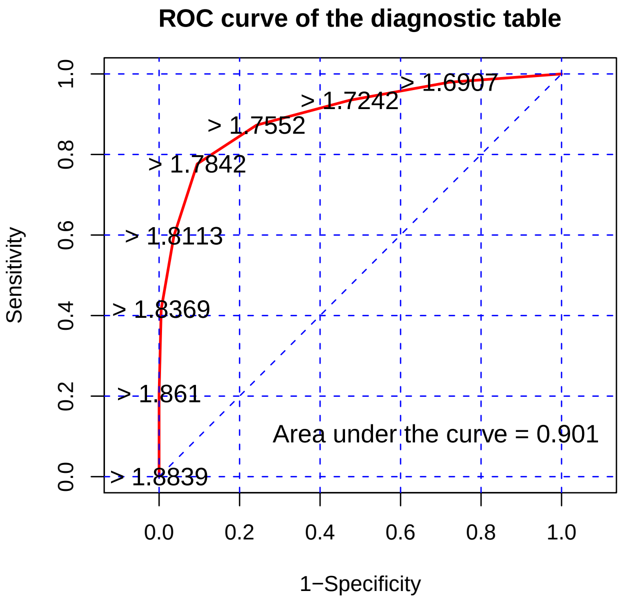
tak pro hodnoty původní proměnné `dose` $x_1, \ldots, x_N$, neboť hodnoty $\widehat{\pi_k}$ mají stejné pořadí jako x_k. ROC křivka $ROC(x)$ má totiž tu důležitou vlastnost, že se nezmění při monotonní transformaci dat, která je v našem případě

$$\widehat{\pi}_k = \widehat{\pi}(x_k) = \frac{1}{1 + \exp(\widehat{\eta}(x_k))} = \frac{1}{1 + \exp(\hat{a} + \hat{b}x_k)}.$$

```
> tableROC <- as.table(cbind(data$population - data$killed,
     data$killed))
> colnames(tableROC) <- c("non-killed", "killed")
> rownames(tableROC) <- as.character(data$dose)
> par(mar = c(5, 5, 3, 0) + 0.1)
> roc.from.table(tableROC, title = TRUE, auc.coords = c(0.25,
     0.1), cex = 1.2, lwd = 2)
```

```
$auc
[1] 0.9010852

$original.table
       Non-diseased Diseased
1.6907           53        6
1.7242           47       13
1.7552           44       18
1.7842           28       28
1.8113           11       52
1.8369            6       53
1.861             1       61
1.8839            0       60

$diagnostic.table
         1-Specificity Sensitivity
           1.000000000   1.0000000
> 1.6907   0.721052632   0.9793814
> 1.7242   0.473684211   0.9347079
> 1.7552   0.242105263   0.8728522
> 1.7842   0.094736842   0.7766323
> 1.8113   0.036842105   0.5979381
> 1.8369   0.005263158   0.4158076
> 1.861    0.000000000   0.2061856
> 1.8839   0.000000000   0.0000000
```

```
> roc1 <- roc.from.table(tableROC, graph = FALSE)
> cut.points <- rownames(roc1$diagnostic.table)
> text(x = roc1$diagnostic.table[, 1], y = roc1$diagnostic.table[,
     2], labels = cut.points, cex = 1.2)
```



Obrázek 5: ROC křivka a hodnota AUC pro binomická data (pomocí příkazů `roc.from.table` z knihovny `epicalc`).

ÚKOLY

1. Proveďte obdobnou analýzu i pro další dvě linkovací funkce, tj. pro probit a komplementární log-log linkovací funkci.

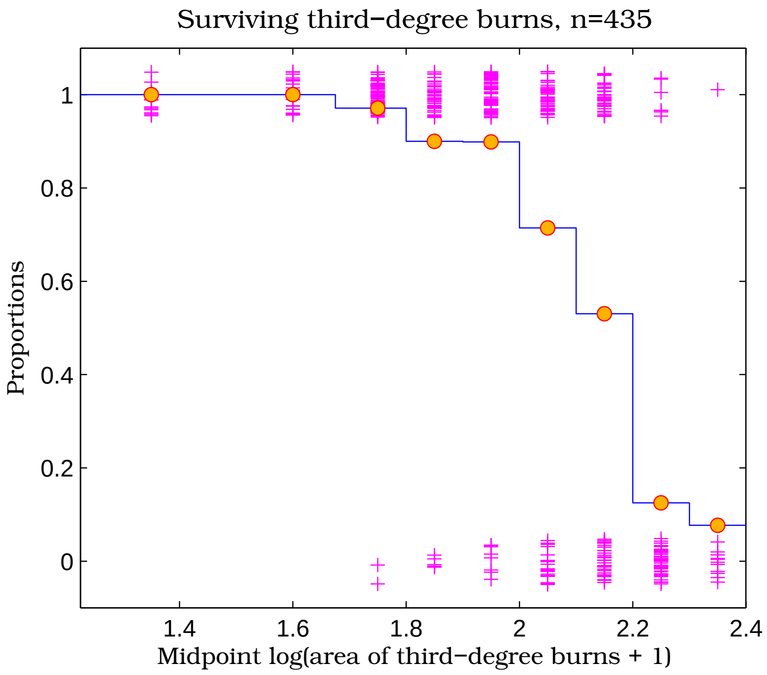
2. Samostatně analyzujte níže uvedené datové soubory. V souborech s koncovkou **txt** je popsána zkoumaná problematika. V souborech s koncovkou **dat** jsou uložena samotná data.

(a)	burns.txt	a	burns.dat
(b)	sirds.txt	a	sirds.dat
(c)	SpaceShuttleData.txt	a	SpaceShuttleData.dat
(d)	vaso.txt	a	vaso.dat
(e)	insecticides.txt	a	insecticides.dat
(f)	HeliothisVirescens.txt	a	HeliothisVirescens.dat

Surviving third-degree burns

Soubor: burns.txt

```
Surviving third-degree burns
============================
Data: burns.dat
These data refer to 435 adults who were treated for third-degree
burns by the University of Southern California General Hospital
Burn Center. The patients were grouped according to the area of
third-degree burns on the body. In the table below are recorded,
for each midpoint of the groupings ■log(area +1),, the number of
patients in the corresponding group who survived, and the number
who died from the burns.

Number of observations: 435
Variable   Description
midpoint   Midpoint of the group corresponding to the patients burn.
survive    Binary variable: survived=1, died=0

Source: Fan, J., Heckman, N.E. and Wand, M.P. (1995) Local polynomial
kernel regression for generalised linear models and quasi-likelihood
functions, Journal of the American Statistical Association,
90, pp. 141-50.
```



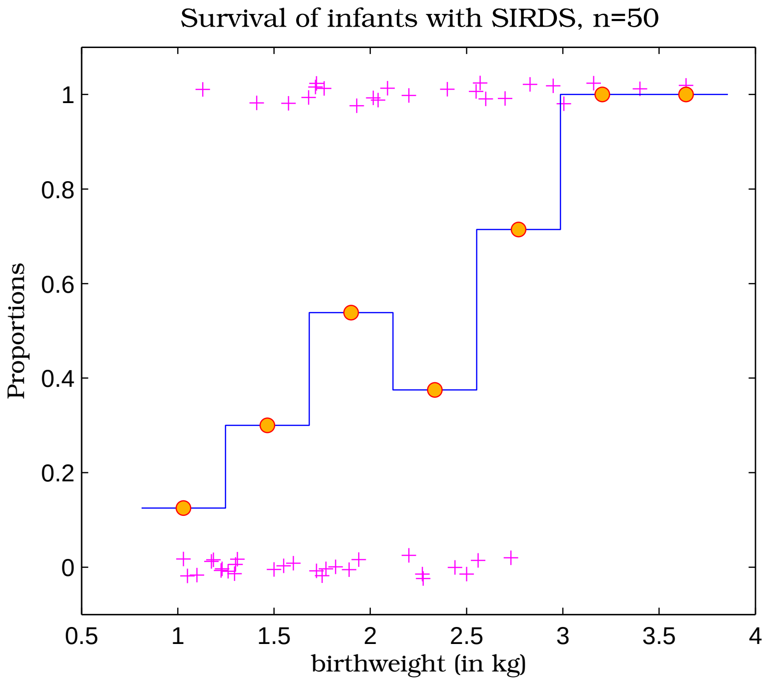
Soubor: burns.dat

```
midpoint,survive
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.35   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
1.60   1
  .
  .
  .
  .
  .
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.25   0
2.35   1
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
2.35   0
```

Transient vasoconstriction in skin of fingers

Soubor: sirds.txt

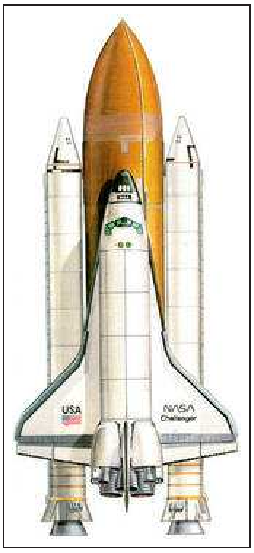
```
Survival of infants with SIRDS
==============================
Data: sirds.dat
Keywords: Logistic regression.
Description: This data set contains the birth weights of fifty infants
who exhibited severe idiopathic respiratory distress syndrome (SIRDS).
This is a serious condition that may result in death, and in fact
of the fifty children sampled only 23 survived.

Number of observations: 50
Variable       Description
birthweight    Weight at birth (in kg)
survival       Binary variable: survived=1, died=0

Source: van Vliet, P.K. and Gupta, J.M. (1973) Sodium bicarbonate
in idiopathic respiratory distress syndrome, Archives of Disease
in Childhood, 48, pp. 249-255.
```



Soubor: sirds.dat

```
birthweight survival
1.130   1
1.575   1
1.680   1
1.760   1
1.930   1
2.015   1
2.090   1
2.600   1
2.700   1
2.950   1
3.160   1
3.400   1
3.640   1
2.830   1
1.410   1
1.715   1
1.720   1
2.040   1
2.200   1
2.400   1
2.550   1
2.570   1
3.005   1
1.050   0
1.175   0
1.230   0
1.310   0
1.500   0
1.600   0
1.720   0
1.750   0
1.770   0
2.275   0
2.500   0
1.030   0
1.100   0
1.185   0
1.225   0
1.262   0
1.295   0
1.300   0
1.550   0
1.820   0
1.890   0
1.940   0
2.200   0
2.270   0
2.440   0
2.560   0
2.730   0
```

O-Ring Damage During Pre-Challenger Shuttle Launches

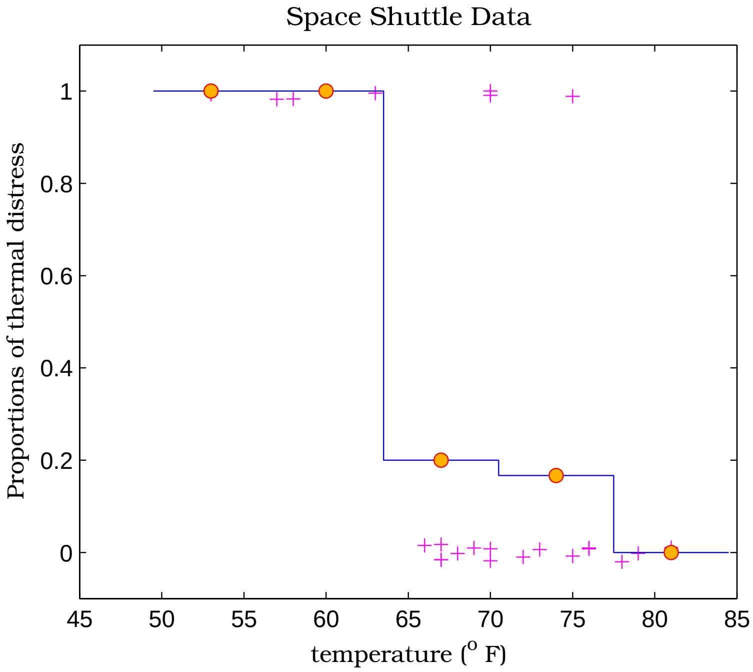


Soubor: `SpaceShuttleData.txt`

```
Space Shuttle Data
==================
This dataset gives information about the 23 space shuttle flights
before the Challenger disaster. We know the temperature of the time
of the flight and whether at least one primary O-ring suffered
thermal distress.

Ft   = flight no.
Temp = temperature
TD   = thermal distress (1 = yes, 0 = no)

Data based on Table 1 in J. Amer. Statist. Assoc, 84: 945-957, (1989),
by S. R. Dalal, E. B. Fowlkes, and B. Hoadley.
```

Soubor:
`SpaceShuttleData.dat`

Ft	Temp	TD
1	66	0
2	70	1
3	69	0
4	68	0
5	67	0
6	72	0
7	73	0
8	70	0
9	57	1
10	63	1
11	70	1
12	78	0
13	67	0
14	53	1
15	67	0
16	75	0
17	70	0
18	81	0
19	76	0
20	79	0
21	75	1
22	76	0
23	58	1



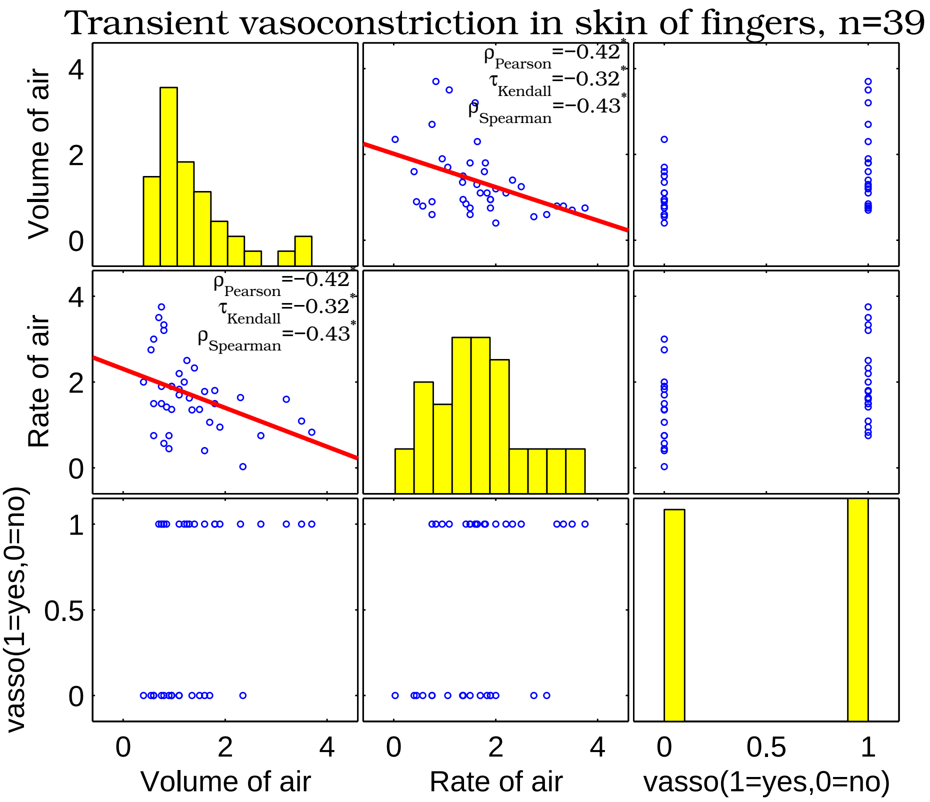
Transient vasoconstriction in skin of fingers

Soubor: vaso.txt

```
Transient vasoconstriction in skin of fingers
=============================================
Data: vaso.dat
Keywords: Logistic regression.
Description: A study was made into the effect of volume and rate of
air inspired by human subjects on the occurrence
of transient vasoconstriction in the skin of the fingers.
A total of 39 observations were
obtained on these variables from 3 subjects in a laboratory. The data are
assumed to be independent (including those on the same subject).
Number of observations: 39
Variable   Description
volume     Volume of air inspired by subject.
rate       Rate of air inspired by subject.
survive    Binary variable: occurrence of transient vasoconstriction
           in the skin of the fingers=1, no-occurence=0
Source: Krzanowski, W.J. (1998) An Introduction to Statistical Modelling,
        London: Arnold. pp. 201-2.
```

Soubor:
vaso.dat

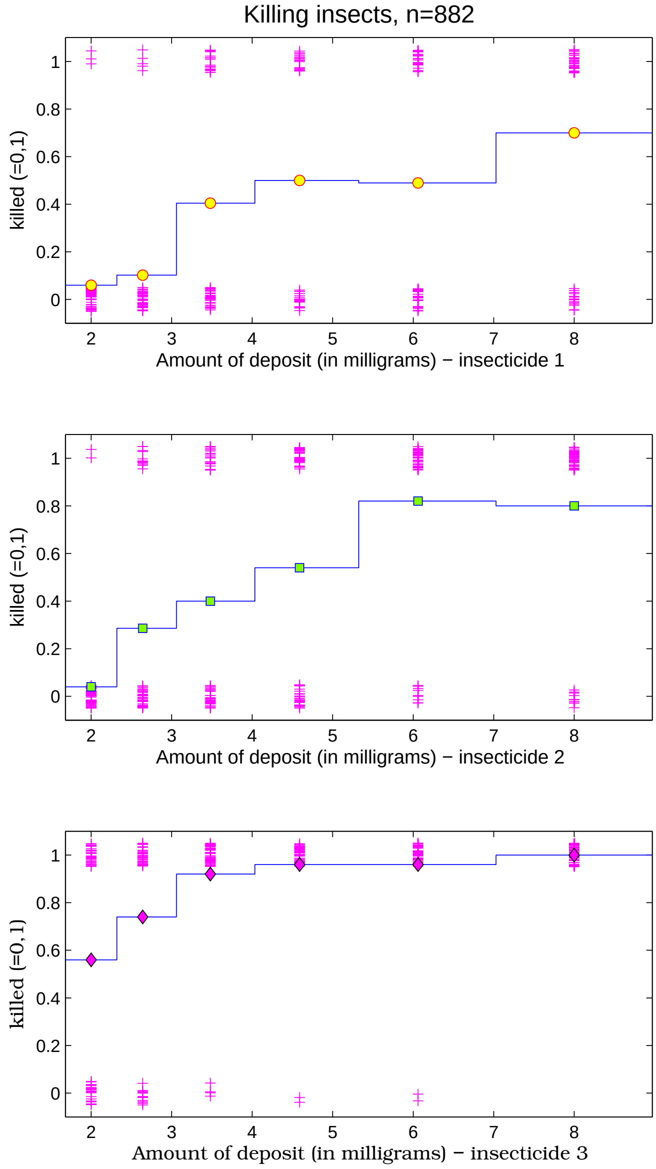
```
volume  rate  vaso
3.70  0.83  1
3.50  1.09  1
1.25  2.50  1
0.75  1.50  1
0.80  3.20  1
0.70  3.50  1
0.60  0.75  0
1.10  1.70  0
0.90  0.75  0
0.90  0.45  0
0.80  0.57  0
0.55  2.75  0
0.60  3.00  0
1.40  2.33  1
0.75  3.75  1
2.30  1.64  1
3.20  1.60  1
0.85  1.42  1
1.70  1.06  0
1.80  1.80  1
0.40  2.00  0
0.95  1.36  0
1.35  1.35  0
1.50  1.36  0
1.60  1.78  1
0.60  1.50  0
1.80  1.50  1
0.95  1.90  0
1.90  0.95  1
1.60  0.40  0
2.70  0.75  1
2.35  0.03  0
1.10  1.83  0
1.10  2.20  1
1.20  2.00  1
0.80  3.33  1
0.95  1.90  0
0.75  1.90  0
1.30  1.63  1
```



The trial of three insecticides

Soubor: insecticides.txt

```
Killing insects
===============
Data: insecticides.dat
Keywords: Logistic regression.
Description: In a trial of three insecticides, batches of about fifty insects
             were exposed to varying deposits of each insecticide.
Number of observations: 882
Variable       Description
killed         Binary variable: killed=1, not-killed=0
insecticide    Categorical variable identifying insecticide (numbered 1 to 3)
deposit        Amount of deposit (in milligrams)

Source: Krzanowski, W.J. (1998) An Introduction to Statistical Modelling,
        London: Arnold. pp. 198-9.
```



Soubor: insecticides.dat

```
killed  insecticide  deposit
1  1  2.00
1  1  2.00
1  1  2.00
0  1  2.00
0  1  2.00
0  1  2.00
   .
   .
   .
0  1  2.64
0  1  2.64
0  1  2.64
1  2  2.64
1  2  2.64
1  2  2.64
   .
   .
   .
0  2  3.48
0  2  3.48
0  2  3.48
1  3  3.48
1  3  3.48
1  3  3.48
   .
   .
   .
1  1  4.59
1  1  4.59
1  1  4.59
0  1  4.59
0  1  4.59
0  1  4.59
   .
   .
   .
0  1  6.06
0  1  6.06
0  1  6.06
1  2  6.06
1  2  6.06
1  2  6.06
   .
   .
   .
0  1  8.00
0  1  8.00
0  1  8.00
1  2  8.00
1  2  8.00
1  2  8.00
   .
   .
   .
1  3  8.00
1  3  8.00
1  3  8.00
1  3  8.00
1  3  8.00
1  3  8.00
```

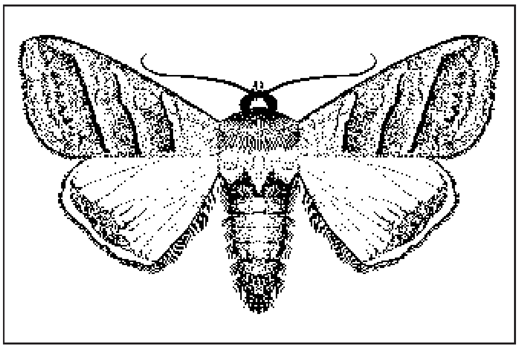
Toxicity of cypermethrin to months Heliothis virescens

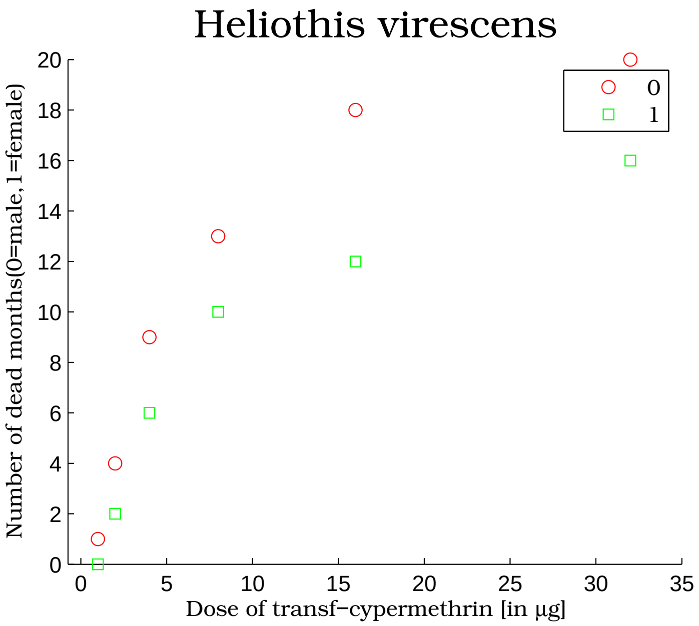
Soubor: HeliothisVirescens.txt

```
Toxicity of cypermethrin to months Heliothis virescens
======================================================
Collett (1991) reports the results of an experiment
on the toxicity of the tobacco budworm Heliothis virescens
to doses of the pyrethoid trans-cypermethin to which
the moths were beginning to show resistance.
Batches of 20 moths of each sex were exposed for 3 days
to the pyretoid and the number in each batch which were dead
or knocked down was recorded.
We fit a logistic regression model using log2(dose)
since the doses are powers of two.
```

Soubor: HeliothisVirescens.dat

```
dose male female
 1    1    0
 2    4    2
 4    9    6
 8   13   10
16   18   12
32   20   16
```




Heliothis virescens
Number of dead months(0=male, 1=female)
Dose of transf−cypermethrin [in µg]
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30
35
○ 0
□ 1