Experimental Design and Biostatistics (TAMS38)
Lecture 11 – Linear models & Logistic regression

Martin Singull

Department of Mathematics
Mathematical Statistics
Linköping University, Sweden

LINKÖPING UNIVERSITY

Content

- Linear models
- Factorial design and regression analysis
- Logistic regression
- Deviance
- Two examples

Linear models

The models of our different factorial designs and models in the regression analysis are included in the class of *linear models*. In particular, the models in factorial design can be written as regression models by using dummy variables.

The linear model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} : n \times 1,$$

where $\boldsymbol{\beta} : (k+1) \times 1$ are unknown parameters, $\boldsymbol{X} : n \times (k+1)$ is a known design matrix and

$$\text{cov}(\boldsymbol{Y}) = \text{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}.$$

One-Way ANOVA with dummy variables

Let

$$
\begin{array}{llll}
y_1, & \ldots, & y_4 & \text{be observations from} \quad N(\mu_1, \sigma) \\
y_5, & \ldots, & y_7 & \text{be observations from} \quad N(\mu_2, \sigma) \\
y_8, & \ldots, & y_{10} & \text{be observations from} \quad N(\mu_3, \sigma)
\end{array}
$$

and

$$
\boldsymbol{Y} = (y_1, \ldots, y_4, y_5, \ldots, y_7, y_8, \ldots, y_{10})'.
$$

We have

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_4 \\ Y_5 \\ \vdots \\ Y_7 \\ Y_8 \\ \vdots \\ Y_{10} \end{pmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}}_{=\boldsymbol{\mu}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \\ \varepsilon_7 \\ \varepsilon_8 \\ \vdots \\ \varepsilon_{10} \end{pmatrix}}_{=\boldsymbol{\varepsilon}},$$

i.e., a regression model with no constant term and we get estimates

$$\widehat{\boldsymbol{\mu}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Exercise Show that the equation gives the ordinary μ-estimator.

A parameterization which is common in the regression analysis is that we let

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j,$$

where

$$z_{j1} = \begin{cases} 1, & \text{for sample 1,} \\ 0, & \text{otherwise,} \end{cases}$$

$$z_{j2} = \begin{cases} 1, & \text{for sample 2,} \\ 0, & \text{otherwise.} \end{cases}$$

Exercise. Give the design matrix $\boldsymbol{X}$.

Note that

$$\mathsf{E}(Y_j) = \begin{cases} \beta_0 + \beta_1, & \text{for sample 1,} \\ \beta_0 + \beta_2, & \text{for sample 2,} \\ \beta_0, & \text{for sample 3,} \end{cases}$$

where β_1 describes the difference between expectations of sample 1 and sample 3 and β_2 describes the difference between expectations of sample 2 and sample 3.

If we want to compare samples 1 and 2 we should study $\beta_1 - \beta_2$.

Example 1 - One-Way ANOVA

Measurements for the four laboratories from the example from Lecture 1 and Lecture 3.

A	B	C	D
0.25	0.18	0.19	0.23
0.27	0.28	0.25	0.30
0.22	0.21	0.27	0.28
0.30	0.23	0.24	0.28
0.27	0.25	0.18	0.24
0.28	0.20	0.26	0.34
0.32	0.27	0.28	0.20
0.24	0.19	0.24	0.18
0.31	0.24	0.25	0.24
0.26	0.22	0.20	0.28
0.21	0.29	0.21	0.22
0.28	0.16	0.19	0.21

Model:

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \beta_3 z_{j3} + \varepsilon_j,$$

where

$$z_{jk} = \begin{cases} 1, & \text{for laboratory } k \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, 2, 3$.

We have expectations

$$\mathrm{E}(Y_j) = \begin{cases} \beta_0 + \beta_1, & \text{for sample 1,} \\ \beta_0 + \beta_2, & \text{for sample 2,} \\ \beta_0 + \beta_3, & \text{for sample 3,} \\ \beta_0, & \text{for sample 4.} \end{cases}$$

Regression Analysis: y versus z1, z2, z3

```
The regression equation is
y = 0.250 + 0.0175 z1 - 0.0233 z2 - 0.0200 z3

Predictor      Coef  SE Coef      T      P
Constant    0.25000  0.01134  22.05  0.000
z1          0.01750  0.01604   1.09  0.281
z2         -0.02333  0.01604  -1.46  0.153
z3         -0.02000  0.01604  -1.25  0.219

S = 0.0392809   R-Sq = 16.1%   R-Sq(adj) = 10.4%

Analysis of Variance

Source           DF        SS        MS     F      P
Regression        3  0.013006  0.004335  2.81  0.050
Residual Error   44  0.067892  0.001543
Total            47  0.080898
```

```
MTB > print m1

Data Display

 Matrix XPXI1

 0.0833333  -0.083333  -0.083333  -0.083333
-0.0833333   0.166667   0.083333   0.083333
-0.0833333   0.083333   0.166667   0.083333
-0.0833333   0.083333   0.083333   0.166667
```

One-way ANOVA: C5 versus C6

```
Source  DF       SS       MS     F      P
C6       3  0.01301  0.00434  2.81  0.050
Error   44  0.06789  0.00154
Total   47  0.08090

S = 0.03928   R-Sq = 16.08%   R-Sq(adj) = 10.36%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level   N     Mean    StDev     --------+---------+---------+---------+-
A      12  0.26750  0.03388                   (--------*--------)
B      12  0.22667  0.04097     (--------*--------)
C      12  0.23000  0.03438      (--------*--------)
D      12  0.25000  0.04651              (--------*--------)
                                --------+---------+---------+---------+-
                                      0.225     0.250     0.275     0.300

Pooled StDev = 0.03928
```

Two-Way ANOVA

Let us have now two factors with two observations per cell

	B_1	B_2	B_3
A_1	y_1, y_2	y_3, y_4	y_5, y_6
A_2	y_7, y_8	y_9, y_{10}	y_{11}, y_{12}

Let

$$z_1 = \begin{cases} 1 & \text{for A-level 1} \\ 0 & \text{otherwise,} \end{cases} \qquad u_1 = \begin{cases} 1 & \text{for B-level 1} \\ 0 & \text{otherwise,} \end{cases}$$

$$u_2 = \begin{cases} 1 & \text{for B-level 2} \\ 0 & \text{otherwise.} \end{cases}$$

Two factor model can be written as

$$Y_j = \beta_0 + \alpha_1 z_{j1} + \gamma_1 u_{j1} + \gamma_2 u_{j2} + \delta_{11} z_{j1} \cdot u_{j1} + \delta_{12} z_{j1} \cdot u_{j2} + \varepsilon_j,$$

that is equivalent to the usual two-factor model.

Exercise. Give the design matrix $\boldsymbol{X}$.

Here, δ_{11} and δ_{12} are our parameters for interactions.

Observ that only the dummy variables that are related to **different** factors should be multiplied.

We obtain $(a-1)(b-1)$ parameters that corresponds to the interaction between pairs of factors.

Matrix of expectations for the cells is give by

	B_1	B_2	B_3
A_1	$\beta_0 + \alpha_1 + \gamma_1 + \delta_{11}$	$\beta_0 + \alpha_1 + \gamma_2 + \delta_{12}$	$\beta_0 + \alpha_1$
A_2	$\beta_0 + \gamma_1$	$\beta_0 + \gamma_2$	β_0

We have the additive model if and only if $\delta_{11} = \delta_{12} = 0$.

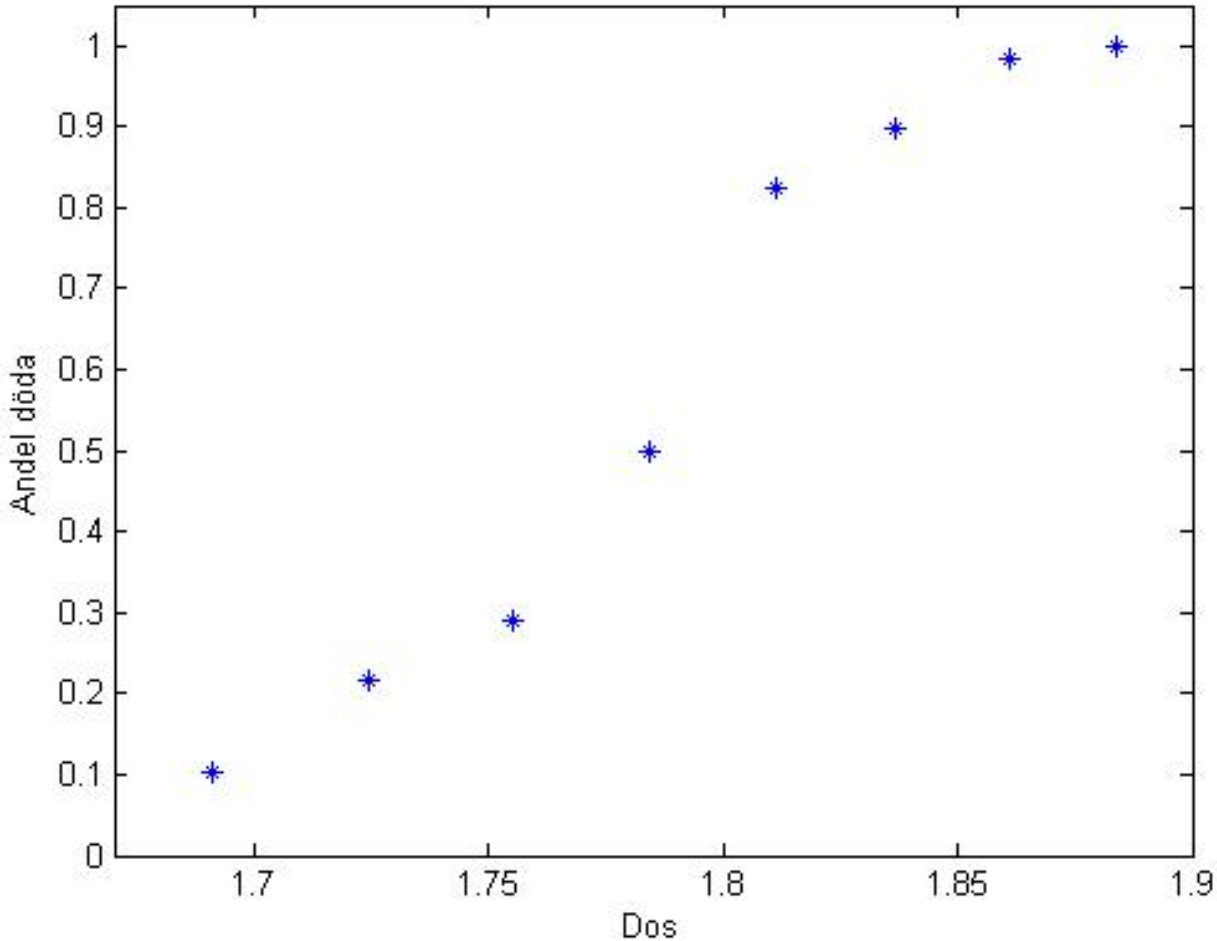
The regression analysis in the example above can be conducted even if we miss some y-observations, i.e., we have an **incomplete** factorial design.

If we analyze the factorial design as a regression model, the results are often much more difficult to interpret than in the standard analysis. The usual hypotheses must be translated into the new parameters etc.

Example 2 – Beetle mortality and logistic regression

The table below shows the number of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations (data from Bliss, 1935).

Dose, x_i ($log_{10} CS_2 mgl^{-1}$)	Number of beetles, n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60


Andel döda
Dos
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.7
1.75
1.8
1.85
1.9

Binomial distribution

A random variable Y follows Binomial distribution ($Y \sim Bin(n, p)$) if probability function is given by

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n,$$

Assume that we have random variables $Y_i \sim Bin(n_i, p_i)$ where Y_i is the number of successes among n_i trials, $i = 1, \ldots, m$.

Then one has m different parameters.

Log-likelihood function

Log-likelihood function (see Appendix) for the maximal model with m parameters is

$$l(p_1, \ldots, p_m; y_1, \ldots, y_m) = \sum_{i=1}^{N} \left(y_i \log \left(\frac{p_i}{1-p_i} \right) + n_i \log (1-p_i) + \log \binom{n_i}{y_i} \right).$$

Logistic regression

We want to explain the proportion of successes in each group and we make it using maximum-likelihood-estimator

$$P_i = \frac{Y_i}{n_i}$$

explained with the help of a number of explanatory variables.

Since the expectation is

$$\mathrm{E}(Y_i) = n_i p_i \quad \text{and} \quad \mathrm{E}(P_i) = p_i$$

we can use the following model for the probabilities p_i

$$g(p_i) = \boldsymbol{x}_i' \boldsymbol{\beta}.$$

Link function

The simplest case is the linear model

$$p = \boldsymbol{x}'\boldsymbol{\beta}.$$

Problem here is that $\boldsymbol{x}'\boldsymbol{\beta}$ can become negative or bigger than 1 and we know that obviously $0 \leq p \leq 1$.

If we let

$$p = g^{-1}\left(\boldsymbol{x}'\boldsymbol{\beta}\right) = \int_{-\infty}^{t} f(z)dz,$$

where $f(z)$ is the probability density function, so-called *tolerance-distribution*, we ensure that $p \in [0, 1]$.

Model: Linear

Tolerance function: $Re[a, b]$

$$p = \frac{x - a}{b - a}, \quad a \leq x \leq b$$

Link function:

$$g(p) = p = \frac{x - a}{b - a} = \beta_1 + \beta_2 x,$$

where $\beta_1 = -\frac{a}{b - a}$ and $\beta_2 = \frac{1}{b - a}$.

Model: Probit

Tolerance function: $N(\mu, \sigma)$

$$p = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Link function:

$$g(p) = \Phi^{-1}(p) = \frac{x-\mu}{\sigma} = \beta_1 + \beta_2 x \quad \textbf{Probit (Normit)},$$

where $\beta_1 = -\frac{\mu}{\sigma}$ and $\beta_2 = \frac{1}{\sigma}$.

Model: Logistic

Tolerance function: $f(z) = \beta_2 \dfrac{e^{\beta_1+\beta_2 z}}{(1+e^{\beta_1+\beta_2 z})^2}$

$$p = \frac{e^{\beta_1+\beta_2 x}}{1+e^{\beta_1+\beta_2 x}}$$

Link function:

$$g(p) = \log\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x \quad \textbf{Logit}$$

Model: Extreme value

Tolerance function: $f(z) = \beta_2 \exp\left\{\beta_1 + \beta_2 z - e^{\beta_1 + \beta_2 z}\right\}$

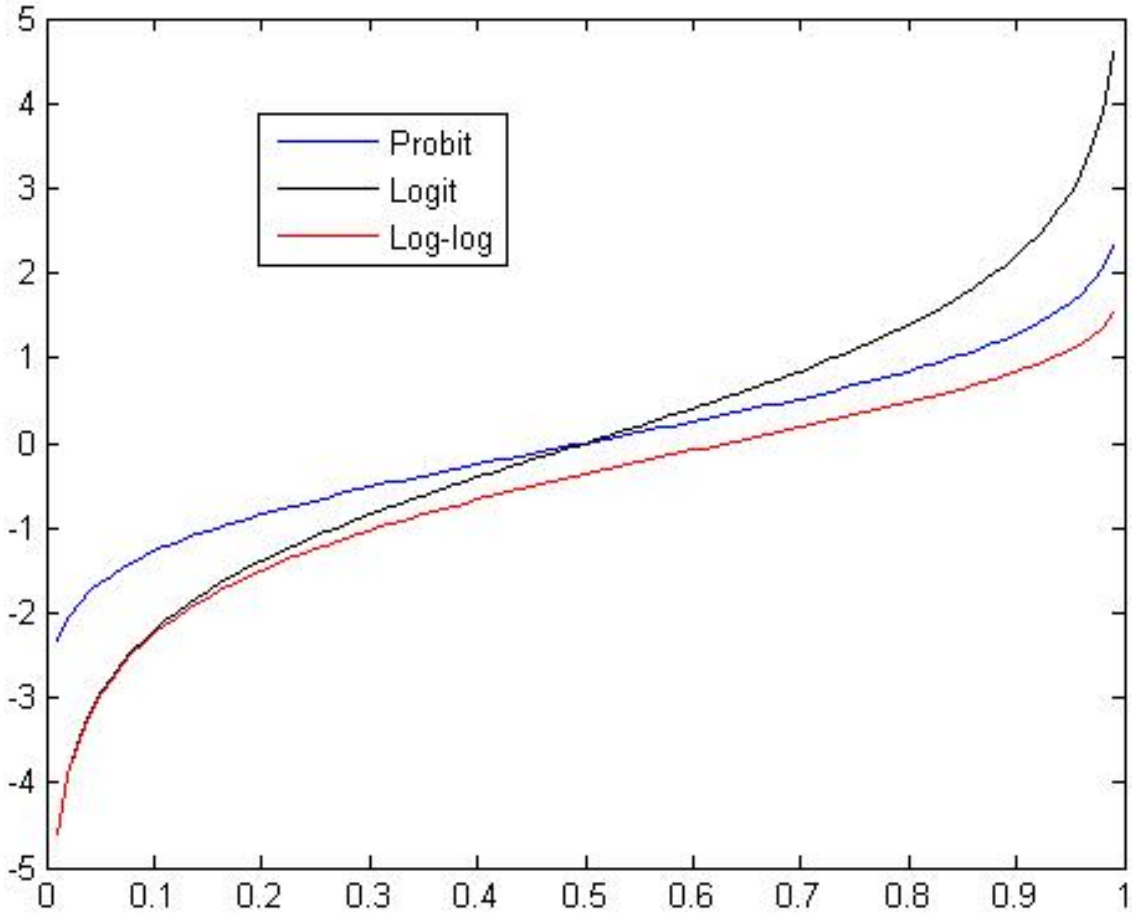
$$p = 1 - \exp\left\{-\exp\left(\beta_1 + \beta_2 x\right)\right\}$$

Link function:

$$g(p) = \log\left(-\log(1-p)\right) = \beta_1 + \beta_2 x$$

Complementary log-log (Gompit)

Link functions, cont.



Deviance

Assume that we have two models, one with p parameters and one with the maximal number of m parameters, where $m > p$.

Let parameters the be $\beta_0 : p \times 1$ and $\beta_1 : m \times 1$.

Assume also that the smaller model is a special case of the bigger. We want to test hypothesis

H_0 : Smaller model with p parameters is as good as the maximal model with m parameters,

vs.

H_1 : Maximal model is better.

To test the hypothesis we will use a quantity called the deviance.

Definition. The **deviance** is defined as

$$D = 2\left(l(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{Y}) - l(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{Y})\right).$$

One can show that under H_0 it holds that

$$D \approx \chi^2(m - p),$$

and we want to reject H_0 in favor of H_1 for large values of the deviance D.

Deviance - Binomial distribution

We have random variables $Y_i \sim Bin\,(n_i, p_i)$. The maximal model has m different parameters $p_1, \ldots, p_m$ with ML-estimates

$$\widehat{\boldsymbol{P}}_1 = (\hat{p}_1, \ldots, \hat{p}_m)\,,$$

where

$$\hat{p}_i = \frac{y_i}{n_i}.$$

Let $\widehat{\boldsymbol{P}}_0$ be ML-estimator for some other model (with fewer parameters).

The deviance is

$$
\begin{aligned}
D &= 2\left(l\left(\widehat{\boldsymbol{P}}_1, \boldsymbol{Y}\right) - l\left(\widehat{\boldsymbol{P}}_0, \boldsymbol{Y}\right)\right) \\
&= 2\sum_{i=1}^{m}\left(y_i \log\frac{\hat{p}_i}{\hat{p}_{0i}} + (n_i - y_i)\log\frac{1-\hat{p}_i}{1-\hat{p}_{0i}}\right) \\
&= 2\sum_{i=1}^{m}\left(y_i \log\frac{y_i}{n_i\hat{p}_{0i}} + (n_i - y_i)\log\frac{n_i - y_i}{n_i(1-\hat{p}_{0i})}\right) \\
&= 2\sum_{i=1}^{m}\left(y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right).
\end{aligned}
$$

The deviance

$$D = 2\sum_{i=1}^{n}\left(y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right)$$

has the form

$$D = 2\sum o_i \log\frac{o_i}{e_i},$$

where o_i are the observed values (y_i and $n_i - y_i$), and e_i are the fitted values ($\hat{y}_i$ and $n_i - \hat{y}_i$).

Example 2, cont.

The table below shows the number of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations (data from Bliss, 1935).

Dose, x_i ($log_{10} CS_2 mgl^{-1}$)	Number of beetles, n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

We will analyze the data using the different link functions.

We start with the logit link function.

$$logit(p) = log\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x.$$

The log-likelihood function with the logic link function is

$$l = \sum_{i=1}^{n} \left(y_i \left(\beta_1 + \beta_2 x_i\right) - n_i \log\left(1 + e^{\beta_1 + \beta_2 x_i}\right) + \log\binom{n_i}{y_i} \right).$$

We use MINITAB.

```
Binary Logistic Regression: y_i, n_i versus x_i

Link Function: Logit

Response Information

Variable  Value      Count
y_i       Event        291
          Non-event    190
n_i       Total        481

Logistic Regression Table
Predictor      Coef  SE Coef       Z      P
Constant   -60.7175  5.18071  -11.72  0.000
x_i         34.2703  2.91214   11.77  0.000

Log-Likelihood = -186.235
Test that all slopes are zero: G = 272.970, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             10.0268   6  0.124
Deviance            11.2322   6  0.081
Hosmer-Lemeshow     10.0268   6  0.124
```

Properties for the ML-estimators

Maximum-likelihood-estimators (MLE) have many *good* properties.

For example for large n we have

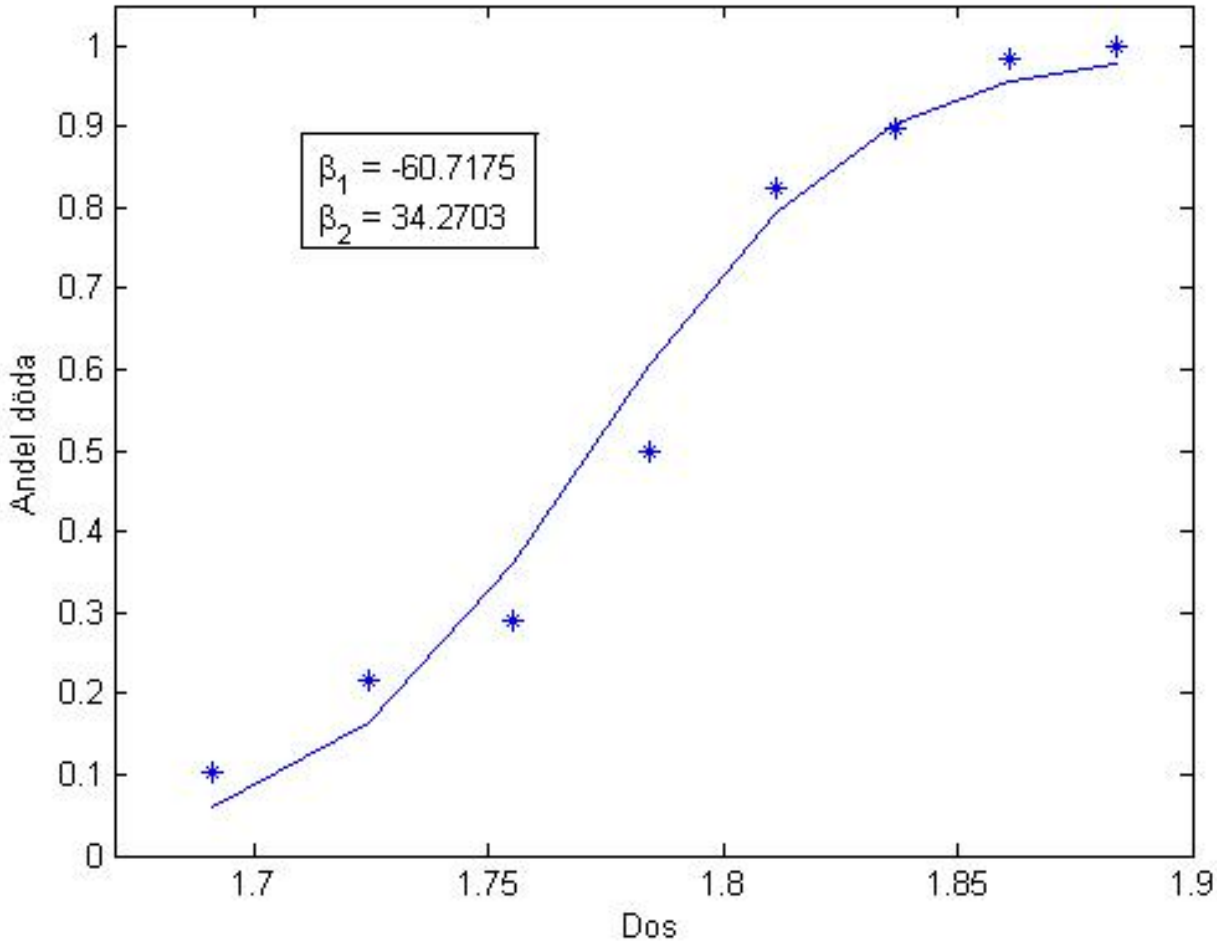
$$\widehat{\boldsymbol{\beta}} \approx N\left(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}^{-1}\right),$$

where *information matrix* $\boldsymbol{\mathcal{I}}$ is given by

$$\boldsymbol{\mathcal{I}} = \left(\mathrm{E}\left(U_j U_k\right)\right)_{j,k} = \ldots = \left(-\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta_j \partial\beta_k}\right)\right)_{j,k},$$

with $U_i = \dfrac{\partial l}{\partial \beta_i}$.

Example 2, cont.


β_1 = -60.7175
β_2 = 34.2703
Andel döda
Dos
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.7
1.75
1.8
1.85
1.9

Example 3 - Embryogenic anthers

The data in the table are taken from Sangwan-Norrell (1977).

The responses y_{jk} are the number of embryogenic anthers of the plant species *Datura innoxia* Mill. obtained when n_{jk} of anthers were prepared under several different conditions.

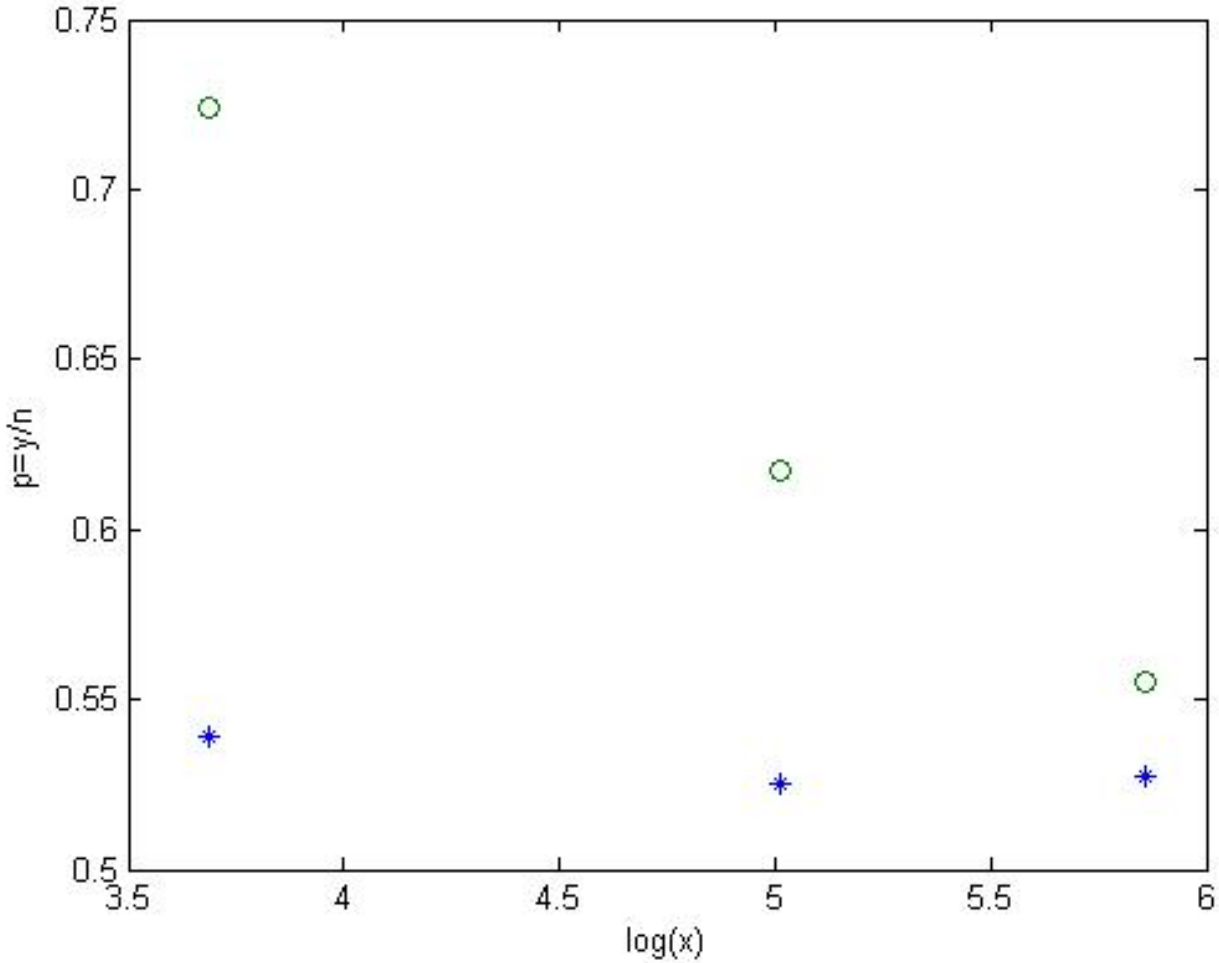
		Centrifuging force (g)		
Storage condition		40	150	350
Control	y_{1k}	55	52	57
	n_{1k}	102	99	108
Treatment	y_{2k}	55	50	50
	n_{2k}	76	81	90

We have one factor with two levels, storage in 3^oC under 48 hours (treatment) and a control group type of storage.

There is also a continuous explanatory variable corresponding to the different centrifuging forces.

We will investigate how the storage and centrifuging forces affect the number of embryogenic anthers.

If we plot $p_{jk} = y_{jk}/n_{jk}$ against the logarithm of those different centrifuging forces x_k, we obtain


0.75
0.7
0.65
0.6
0.55
0.5
p=y/n
3.5
4
4.5
5
5.5
6
log(x)

We will now consider two logistic models for p_{jk} (probability that anthers are *embryogenic*).

The first model is a model with different intercepts and different slopes for those two groups.

$$\begin{aligned} logit(p_{jk}) &= \beta_0 + \alpha_0 z_j + \beta_1 x_k + \alpha_1 z_j x_k \\ &= \beta_0 + \alpha_0 z_j + (\beta_1 + \alpha_1 z_j) x_k, \end{aligned} \tag{1}$$

where $z_j = 0$ for control group, $z_j = 1$ for *treatment group*.

The other model has a different intercepts but the same slope for the two groups

$$logit(p_{jk}) = \beta_0 + \alpha_0 z_j + \beta_1 x_k. \tag{2}$$

We use MINITAB.

Model (1)

```
Binary Logistic Regression: y, n versus z, x, zx

Link Function: Logit

Logistic Regression Table
Predictor         Coef   SE Coef      Z      P
Constant      0.233910  0.628418   0.37  0.710
z              1.97721  0.998079   1.98  0.048
x           -0.0227412  0.126851  -0.18  0.858
zx           -0.318628  0.198881  -1.60  0.109

Log-Likelihood = -374.109
Test that all slopes are zero: G = 10.424, DF = 3, P-Value = 0.015

Goodness-of-Fit Tests
Method            Chi-Square  DF      P
Pearson            0.0276564   2  0.986
Deviance           0.0276407   2  0.986
Hosmer-Lemeshow    0.0276564   4  1.000
```

Model (2)

```
Binary Logistic Regression: y, n versus z, x

Link Function: Logit

Logistic Regression Table
Predictor       Coef    SE Coef      Z      P
Constant    0.876775   0.487037   1.80  0.072
z           0.406841   0.174624   2.33  0.020
x          -0.154596  0.0970260  -1.59  0.111

Log-Likelihood = -375.404
Test that all slopes are zero: G = 7.833, DF = 2, P-Value = 0.020

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             2.59800   3  0.458
Deviance            2.61878   3  0.454
Hosmer-Lemeshow     2.59800   4  0.627
```

APPENDIX

Estimators

There are several ways to point to estimate the parameters of a probability model

- moment method,
- least square method,
- maximum-likelihood method.

We now want to look more closely at the maximum likelihood method since it is one the most often used by us.

Likelihood function

Let $x_1, \ldots, x_n$ be a random sample with independent observations from the distribution $f(x; \boldsymbol{\theta})$ that depends on the unknown parameters $\boldsymbol{\theta}$.

Definition. The function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = f(x_1; \boldsymbol{\theta}) \cdot \ldots \cdot f(x_n; \boldsymbol{\theta})$$

is called the **likelihood function**.

ML-estimator

Definition. A value $\hat{\theta}$, for which the likelihood function $L(\theta)$ obtains its **highest** value is called **maximum- likelihood-estimate** (ML-estimate) of θ.

Before one maximize it is often convenient to take the logarithm of the likelihood function

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

and then differentiate with respect to the parameters that you want to maximize.

Some of the maximum-likelihood-estimators (MLE) properties are given below.

If $\hat{\theta}$ is MLE of θ then, under *certain* (rather mild) conditions, for large n we have

$$\frac{\hat{\theta} - E\hat{\theta}}{\sqrt{Var\hat{\theta}}} \approx N(0,1).$$

This can be generalized to the multidimensional case where one can show that for large n we have

$$\widehat{\boldsymbol{\theta}} \approx N\left(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}\right),$$

where *information matrix* $\boldsymbol{\mathcal{I}}$ is given by

$$\boldsymbol{\mathcal{I}} = \left(E\left(U_j U_k\right)\right)_{j,k} = \ldots = \left(-E\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right)_{j,k},$$

with $U_i = \dfrac{\partial l}{\partial\theta_i}$.

Linköping University - Research that makes a difference