# Applied Statistical Regression
## AS 2012 – Week 12

*Marcel Dettling*

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, December 10, 2012

# Applied Statistical Regression
## AS 2012 – Week 12

# *Binomial Regression Models*

## Example: Effectiveness of Insecticide

| Concentration in log of mg/l $x_i$ | Number of insects $n_i$ | Number of killed insects $y_i$ |
|---|---|---|
| 0.96 | 50 | 6 |
| 1.33 | 48 | 16 |
| 1.63 | 46 | 24 |
| 2.04 | 49 | 42 |
| 2.32 | 50 | 44 |

→ the response is the number of killed insects: $Y_i \mid x_i \sim Bin(n_i, p_i)$

→ our main interest is in the proportion of insects that survive

→ while this could be treated as a logistic regression problem with repeated measurements, we gain efficiency by working with grouped data and a binomial regression approach

# Applied Statistical Regression
## AS 2012 – Week 12

## *Model and Estimation*

The goal is to find a relation:

$$p_i(x) = P(Y_i = 1 | X = x) \sim \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

We will again use the logit link function such that $\eta_i = g(p_i)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

Here, $p_i$ is the expected value $E[Y_i / n_i]$, and thus, also this model here fits within the GLM framework. The log-likelihood is:

$$l(\beta) = \sum_{i=1}^{k} \left[ \log\binom{n_i}{y_i} + n_i y_i \log(p_i) + n_i(1-y_i)\log(1-p_i) \right]$$

# Applied Statistical Regression
## AS 2012 – Week 12

## *Fitting with R*

We need to generate a two-column matrix where the first contains the “successes” and the second contains the “failures”

```
> killsurv
     killed surviv
[1,]      6     44
[2,]     16     32
[3,]     24     22
[4,]     42      7
[5,]     44      6

> fit <- glm(killsurv~conc, family="binomial")
```

# *Summary Output*

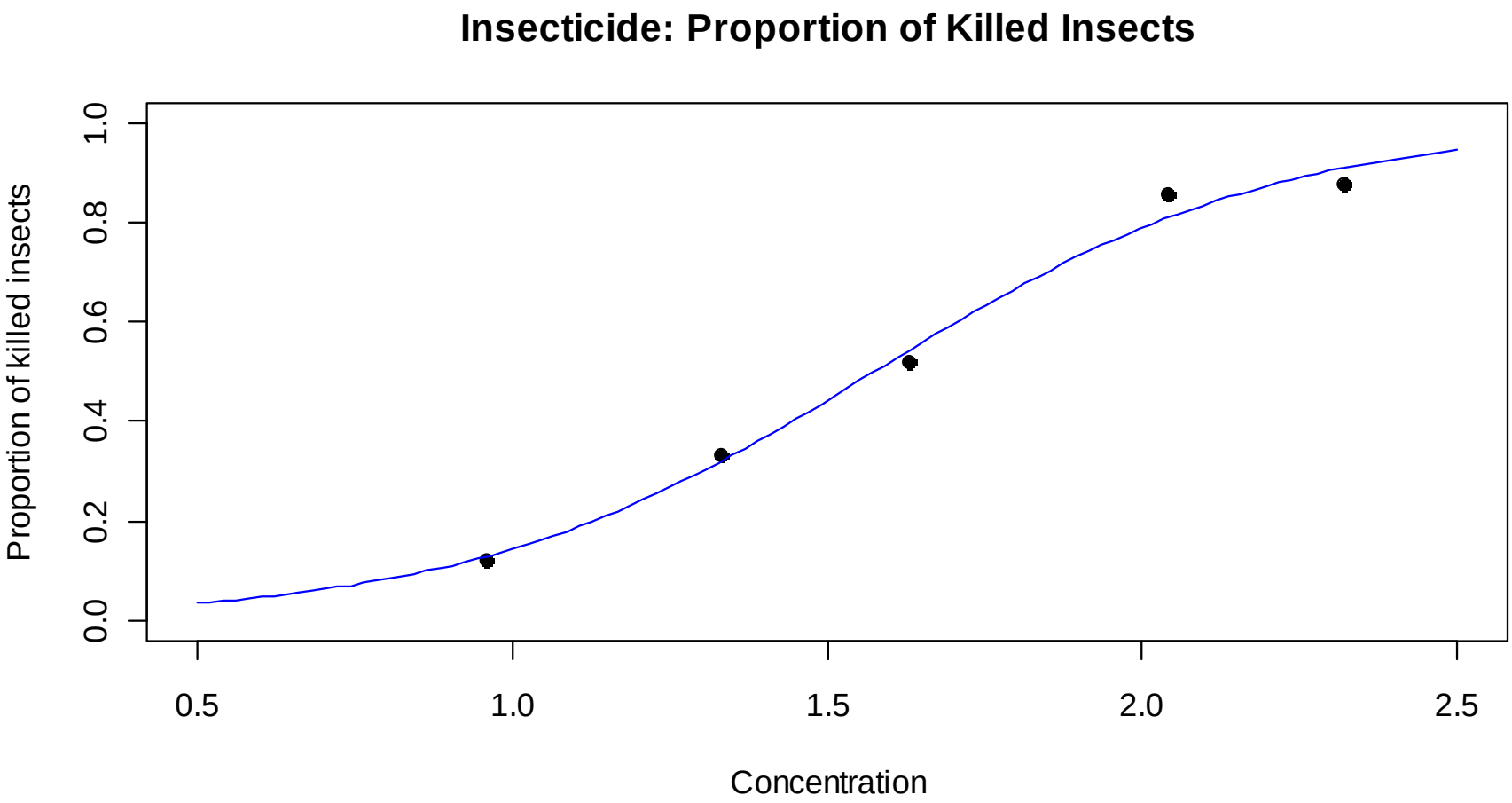
The result for the insecticide example is:

```
> summary(glm(killsurv ~ conc, family = "binomial")


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8923     0.6426  -7.613 2.67e-14 ***
conc          3.1088     0.3879   8.015 1.11e-15 ***
---
Null deviance: 96.6881  on 4  degrees of freedom
Residual deviance:  1.4542  on 3  degrees of freedom
AIC: 24.675
```

# Applied Statistical Regression
## AS 2012 – Week 12

# *Proportion of Killed Insects*



# Applied Statistical Regression
## AS 2012 – Week 12

# *Global Tests for Binomial Regression*

For GLMs there are three tests that can be done:

- **Goodness-of-fit test = model evaluation test**
  - based on comparing against the saturated model
  - not suitable for non-grouped, binary data
- **Comparing two hierachical models**
  - likelihood ratio test leads to deviance differences
  - test statistics has an asymptotic Chi-Square distribution
- **Global test**
  - comparing versus an empty model with only an intercept
  - this is a nested model, take the null deviance

# Applied Statistical Regression
## AS 2012 – Week 12

# *Model Evaluation vs. Saturated Model*

**Null hypothesis:** The fitted model with $p$ predictors is correct

→ **the *residual deviance* will be our test statistic!**

**Paradigm**: take twice the difference between the log-likelihood for our current model and the saturated one, which fits the proportions perfectly, i.e. $\hat{p}_i = y_i / n_i$

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

Because the saturated model fits as well as any model can fit, the residual deviance given in the summary output measures how close our model comes to perfection.

# Applied Statistical Regression
## AS 2012 – Week 12

# *Evaluation of the Test*

**Asymptotics:**

If $Y_i$ is truly binomial and the $n_i$ are large, the deviance is approximately $\chi^2$ distributed. The degrees of freedom is:

$$k - (\#\ of\ \ predictors) - 1$$

```
> pchisq(deviance(fit), df.residual(fit), lower=FALSE)
[1] 0.69287
```

**Quick and dirty:**

$Deviance \gg df$ : → model is not worth much.

More exactly: check $df \pm 2\sqrt{2df}$

→ only apply this test if at least all $n_i \geq 5$

# Applied Statistical Regression
## AS 2012 – Week 12

# *Overdispersion*

**What if** $Deviance \gg df$ **???**

**1) Check the structural form of the model**

- model diagnostics
- predictor transformations, interactions, …

**2) Outliers**

- should be apparent from the diagnostic plots

**3) IID assumption for** $p_i$ **within a group**

- unrecorded predictors or inhomogeneous population
- subjects influence other subjects under study

# Applied Statistical Regression
## AS 2012 – Week 12

# *Overdispersion: a Remedy*

We can deal with overdispersion by estimating:

$$\hat{\phi} = \frac{X^2}{n-p} = \frac{1}{n-p} \cdot \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1-\hat{p}_i)}$$

This is the sum of squared Pearson residuals divided with the df

**Implications:**

- regression coefficients remain unchanged
- standard errors will be different: inference!
- need to use an F-test for comparing nested models

# Applied Statistical Regression
## AS 2012 – Week 12

## *Results when Correcting Overdispersion*

```
> phi <- sum(resid(fit)^2)/df.residual(fit)
> phi
[1] 0.4847485
> summary(fit, dispersion=phi)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.8923     0.4474  -10.94   <2e-16 ***
conc          3.1088     0.2701   11.51   <2e-16 ***
---
(Dispersion parameter taken to be 0.4847485)
Null deviance: 96.6881  on 4  degrees of freedom
Residual deviance:  1.4542  on 3  degrees of freedom
AIC: 24.675
```

# Applied Statistical Regression
## AS 2012 – Week 12

# *Global Tests for Binomial Regression*

For GLMs there are three tests that can be done:

- **Goodness-of-fit test**
  - based on comparing against the saturated model
  - not suitable for non-grouped, binary data

- **Comparing two nested models**
  - likelihood ratio test leads to deviance differences
  - test statistics has an asymptotic Chi-Square distribution

- **Global test**
  - comparing versus an empty model with only an intercept
  - this is a nested model, take the null deviance

# Applied Statistical Regression
## AS 2012 – Week 12

# *Testing Nested Models and the Global Test*

For binomial regression, these two tests are conceptually equal to the ones we already discussed in binary logistic regression.

→ *We refer to our discussion there and do not go into further detail here at this place!*

## Null hypothesis and test statistic:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$$

$$2\left(ll^{(B)} - ll^{(S)}\right) = D\left(y, \hat{p}^{(S)}\right) - D\left(y, \hat{p}^{(B)}\right)$$

## Distribution of the test statistic:

$$D^{(S)} - D^{(B)} \sim \chi^2_{p-q}$$

# Applied Statistical Regression
## AS 2012 – Week 12

# *Poisson-Regression*

### When to apply?

- Responses need to be counts
  - for bounded counts, the binomial model can be useful
  - for large numbers the normal approximation can serve

- The use of Poisson regression is a must if:
  - unknown population size and small counts
  - when the size of the population is large and hard to come by, and the probability of “success”/ the counts are small.

### Methods:

Very similar to Binomial regression!

# *Extending...: Example 2*

## Poisson Regression

*What are predictors for the locations of starfish?*

→ analyze the number of starfish at several locations, for which we also have some covariates such as water temperature, ...

→ the response variable is a count. The simplest model for this is a Poisson distribution.

We assume that the parameter $\lambda_i$ at location *i* depends in a linear way on the covariates:

$$Y_i \sim Pois(\lambda_i), \text{ where } \log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$