Applied Statistical Regression
AS 2012 – Week 11

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 3, 2012

Applied Statistical Regression
AS 2012 – Week 11

Extending the Linear Model

What is the problem?

→ **So far, we exclusively considered continuous response variables. Now, we wish to extend this to binary and categorical response, proportions or counts!**

- This does not fit within the current framework
- Counterexamples: → see next slides

We need some additional techniques which can deal with these types of situations. Depending on how the response variable is, there are several different approaches.

Extending...: Example 1

Logistic Regression, i.e. 0/1 response:

In human medicine, we are often interested in the question for how much „dose" of a medication we have an effect, i.e. a reduction in pain or symptoms.

Data:

Patients, where each one obtains some „dose" and as a response, either has a reduction (1), or not (0).

There may be some further predictors such as age, sex, … that contribute towards predicting the response.

Applied Statistical Regression

AS 2012 – Week 11

Extending...: Example 1

Logistic Regression, i.e. 0/1 response:

- A statistical model for this example takes into account that for a given “dose” resp. predictor configuration, we will only have an effect on some of the subjects, but not on all of them.
- We thus need to model the relation between the binary response and a number of predictors.

The perhaps ***simplest, but faulty approach*** is:

$$P(Y_i = 1 \mid X) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta x_{ip}$$

→ **This will ultimately lead to probabilities beyond [0,1]**.

Extending...: Example 1

- We obtain a better model if we transform the response variable to a scale that ranges from minus to plus infinity.

- Usual choice is the so-called logit transformation:

$$p \mapsto \log(p/(1-p))$$

We obtain the **logistic regression model**:

$$\log\left(\frac{P(Y_i=1|X)}{1-P(Y_i=1|X)}\right)=\beta_0+\beta_1 x_{i1}+...+\beta x_{ip}$$

→ **all fitted values are within [0,1]**.

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 λ_i at location *i* depends in a linear way on the covariates:

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

Extending...: Example 3

Log-Linear-Models

Question:

Prediction of a nominal response variable

Example:

Which party does a person favor, depending on covariates such as education, age, sex, region, …

→ such data can be summarized with contingency tables

→ and they can be modeled using log-linear models

Applied Statistical Regression
AS 2012 – Week 11

Generalized Linear Models

What is it?

- General framework for regression type modeling
- Many different response types are allowed
- Notion: the expected value of the response has a monotone relation to a linear combination of the predictors.

$$E[Y_i \mid X] = g(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})$$

- Some further requirements on variance and density of Y

→ **may seem complicated, but is very powerful!**

Binary Logistic Regression

What is it?

- Response $Y_i \in \{0,1\}$

What do we need to take care of?

- Formulation of the model
- Estimation
- Inference
- Model diagnostics
- Model choice

Applied Statistical Regression
AS 2012 – Week 11

Example

Premature Birth, by Hubbard (1986)

$Y_i \in \{0,1\}$ survival (1) /death (0) after premature birth.

Predictors:

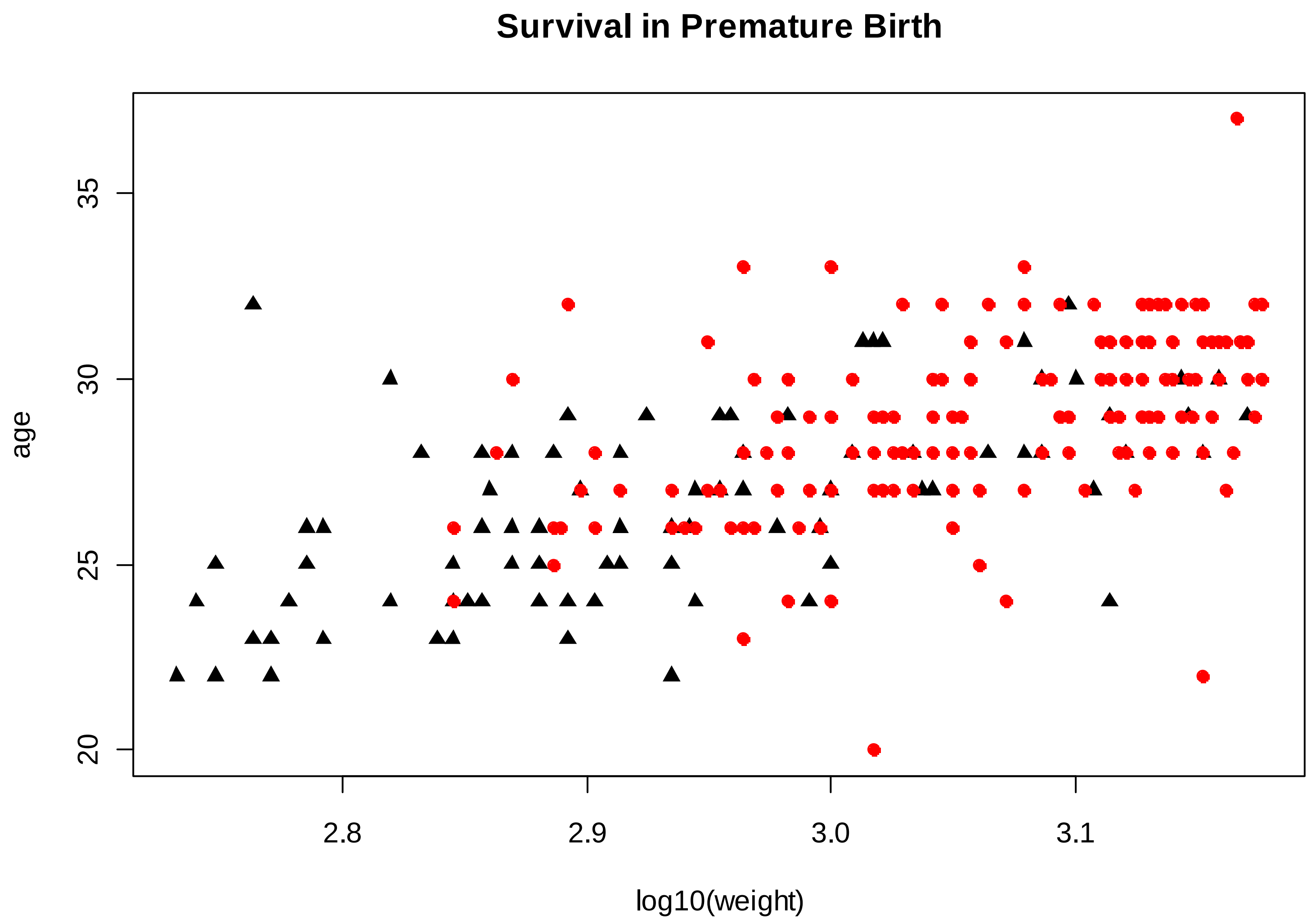
- weight (in grams) at birth
- age at birth (in weeks of pregnancy)
- apgar scores (vital function after 1 and 5 min)
- pH-value of the blood (breathing)

Observations:

- there are 247 instances

Applied Statistical Regression
AS 2012 – Week 11

Example



Logistic Regression Model

- $Y_i \in \{0,1\}$ has a Bernoulli distribution.
- The parameter of this distribution is p_i, the success rate

Now please note that:

$$p_i = P(Y_i = 1 \mid X) = E[Y_i \mid X]$$

→ the most powerful notion of the logistic regression model is to see it as a model where we try to find a relation between the expected value of Y_i and the predictors!

Important: $P(Y_i = 1) = \beta_0 + \beta_1 x_{i1} + ... + \beta x_{ip}$ is no good here!

Logit Transformation

Goal: mapping from $[0,1] \mapsto (-\infty, +\infty)$

Logit transformation: $g(p) = \log\left(\frac{p}{1-p}\right)$

Interpretation: probabilities are mapped to logged odds ("Wettverhältnisse") which can then be modeled linearly.

$$\log\left(\frac{P(Y_i = 1 \mid X)}{1 - P(Y_i = 1 \mid X)}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta x_{ip}$$

→ where is the error term?

...

Applied Statistical Regression

AS 2012 – Week 11

Some Remarks

- For estimating the regression coefficients, we require the observations to be independent.

- There is no restriction for the predictors. They can be continuous, categorical, transformed, interactions, …

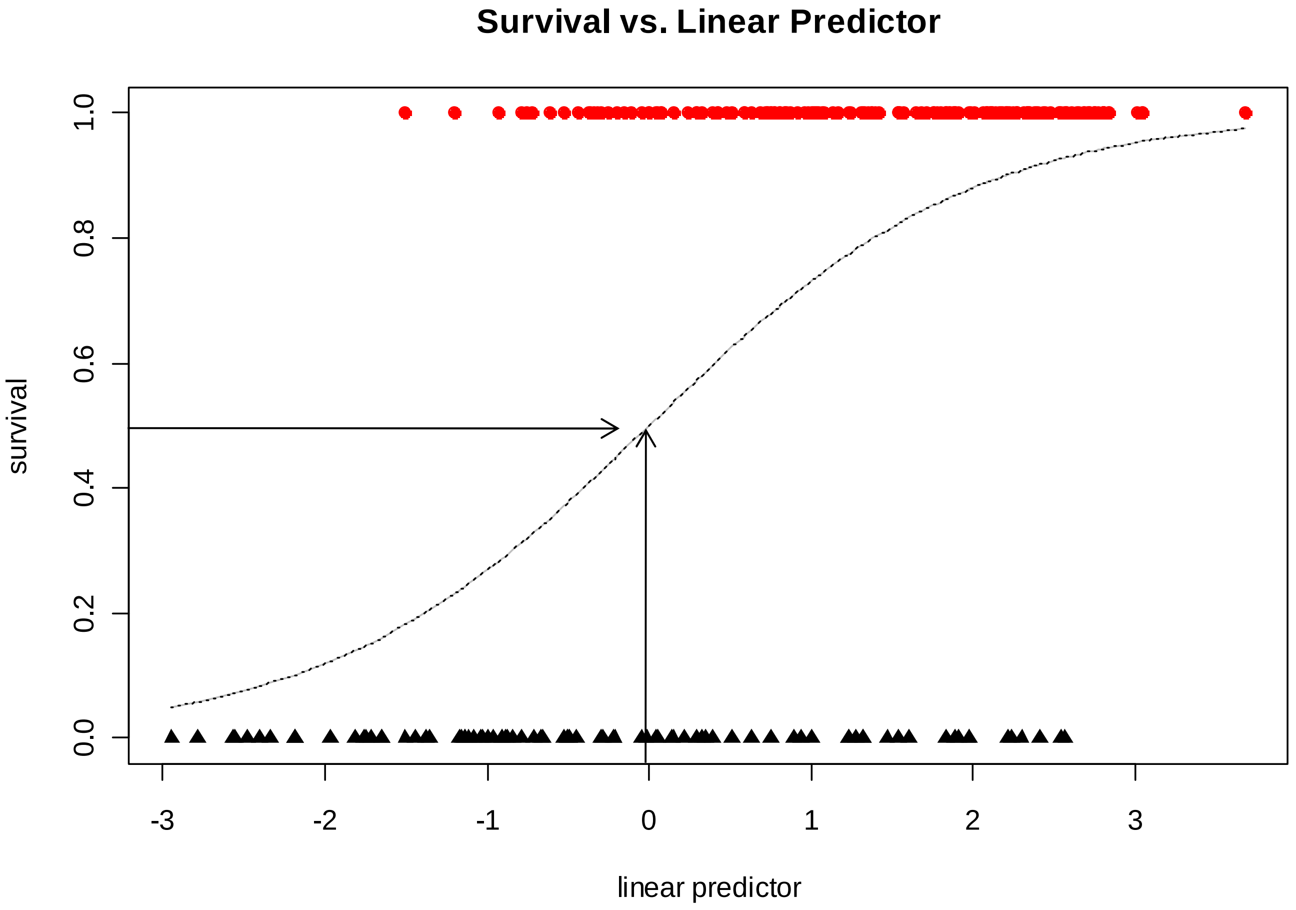
- $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ is called the linear predictor

- $g(\cdot)$ is the link function, mapping between $E[Y_i \mid X]$ and η_i

- **There are other (less important) link functions:**
  - probit link
  - c-log-log link

Survival vs. Linear Predictor

- $g\left(P(Y=1 \mid \log_{10}(weight), age)\right) = -33.97 + 10.17 \cdot \log_{10}(weight) + 0.14 \cdot age$



Estimation

Multiple linear regression:

→ *minimize sum of squared residuals!*

can be solved in closed form

Logistic regression:

→ *maximum likelihood approach!*

leads to a non-linear equation system that needs to be solved with an iterative approach by weighted multiple linear regressions.

Important:

→ seems like a very different paradigm, but is it?

Applied Statistical Regression

AS 2012 – Week 11

Interpretation of the Coefficients

→ see blackboard…

Applied Statistical Regression

AS 2012 – Week 11

Summary Output from R

```
> summary(glm(survival ~ I(log10(weight)) + age,
              family  = "binomial", data = baby)

Deviance Residuals: ...

Coefficients:     Estimate Std. Error z value Pr(>|z|)
(Intercept)      -33.97108    4.98983  -6.808 9.89e-12 ***
I(log10(weight))  10.16846    1.88160   5.404 6.51e-08 ***
age                0.14742    0.07427   1.985   0.0472 *
---
Null deviance: 319.28  on 246  degrees of freedom
Residual deviance: 235.94  on 244  degrees of freedom
AIC: 241.94
```

Inference: Individual Parameter Tests

Multiple Linear Regression:

Gaussian errors → $\hat{\beta}_j$ are normally distributed

Logistic Regression:

There are no errors, variability arises from Bernoulli distribution

The regression coefficients $\hat{\beta}_j$ are only approximately normally distributed with a covariance matrix V that can be derived from the coefficients.

Hence: $$Z = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{jj}}} \sim N(0,1)$$

Inference: Global Tests with GLMs

There are three tests, two can be done with logistic regression:

- **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 11

Goodness-of-Fit

Multiple Linear Regression:

Sum of Squared Residuals

Logistic Regression:

Residual Deviance

$$D(y, \hat{p}) = -2\sum_i (y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i))$$

- based on the log-likelihood
- in principle: comparison against fully saturated model
- for logistic regression, there is no formal test here

Applied Statistical Regression

AS 2012 – Week 11

Comparing Nested Models

Model 1: small model S, with q parameters

Model 2: big model B, with p parameters

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:

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

Example with drop1()

```
> drop1(fit, test="Chisq")
Single term deletions
Model: survival ~ I(log10(weight)) + age
                 Df Deviance    AIC    LRT   Pr(Chi)
<none>               235.94 241.94
I(log10(weight))  1   270.19 274.19 34.247 4.855e-09 ***
age               1   239.89 243.89  3.948   0.04694 *
```

Question:

- where is the difference to the summary output?
- it exists, though it‘s not obvious and asymptotically vanishes

AIC and Variable Selection

General remark:

All comparison between models of different size can also be done using the AIC criterion. Not only in logistic regression, but also here.

The criterion:

$$AIC = D(y_i, \hat{p}) + 2p$$

Variable selection:

- stepwise approaches as with multiple linear regression
- factor variables need to be treated the right way!

Null Deviance

Smallest model:

- The smallest model is without predictors, only with intercept
- Fitted values will all be equal to $\hat{\pi}_0$
- Our best fit (F) and the smallest model (0) are nested

A global test:

$$2\left(ll^{(F)} - ll^{(0)}\right) = D\left(y, \hat{p}^{(0)}\right) - D\left(y, \hat{p}^{(F)}\right)$$

Example and "Quick Check": → **see blackboard...**

```
Null deviance: 319.28  on 246  degrees of freedom
Residual deviance: 235.94  on 244  degrees of freedom
```

Model Diagnostics

Diagnostics are:

- in principle as important with logistic regression as they are with multiple linear regression models
- again based on differences between fitted & observed values

→ we now have to take into account that the variances are not equal for the different instances.

→ we have to come up with novel types of residuals:

Pearson and **Deviance residuals**

Applied Statistical Regression
AS 2012 – Week 11

Pearson Residuals

Take the difference between observed and fitted value and divides by an estimate of the standard deviation:

$$R_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

→ R_i^2 is the contribution of the *ith* observation to the Pearson statistic for model comparison.

→ It is important to note that Pearson residuals exceeding a value of two in absolute value warrant a closer look

Deviance Residuals

Take the contribution of the i*th* observation to the log-likelihood, i.e. the chi-square statistic for model comparison.

$$d_i = \left(y_i \cdot \log(\hat{p}_i) + (1 - y_i) \cdot \log(1 - \hat{p}_i) \right)$$

For obtaining a well interpretable residual, we take the square root and the sign of the difference between true and fitted value:

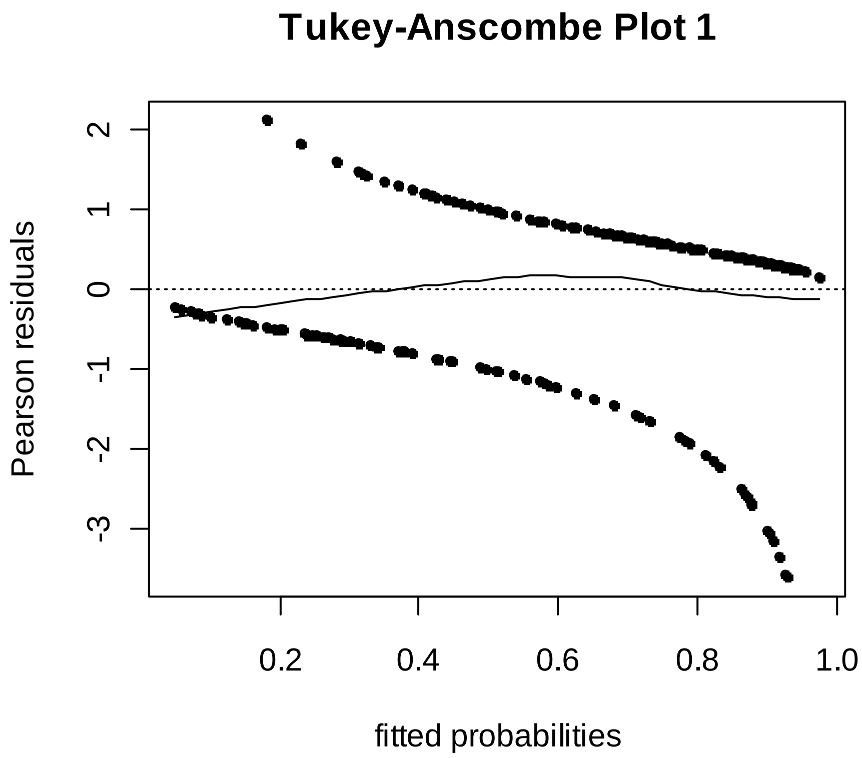
$$D_i = sign(y_i - \hat{\pi}_i) \cdot \sqrt{d_i}$$

→ - *deviance residuals > 2 warrant a closer look*.

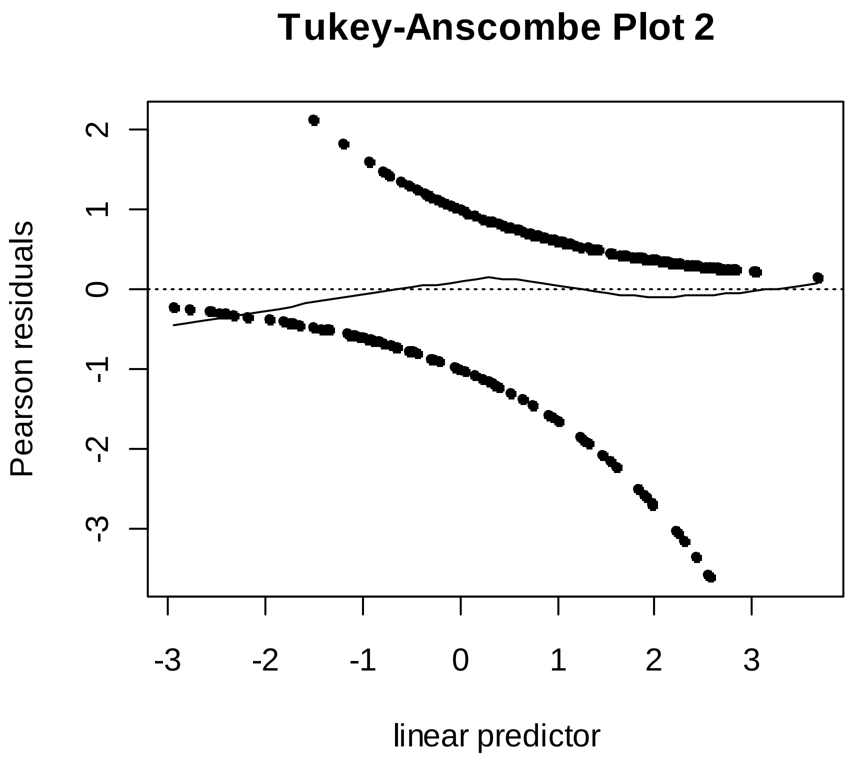
- *the distribution of the deviance residuals is not known*.

Tukey-Anscombe Plot

Remark: sometimes studentized residuals are used!



Tukey-Anscombe Plot 1



Tukey-Anscombe Plot 2

Applied Statistical Regression

AS 2012 – Week 11

Tukey-Anscombe Plot

The Tukey-Anscombe plots in R are not perfect. Better use:

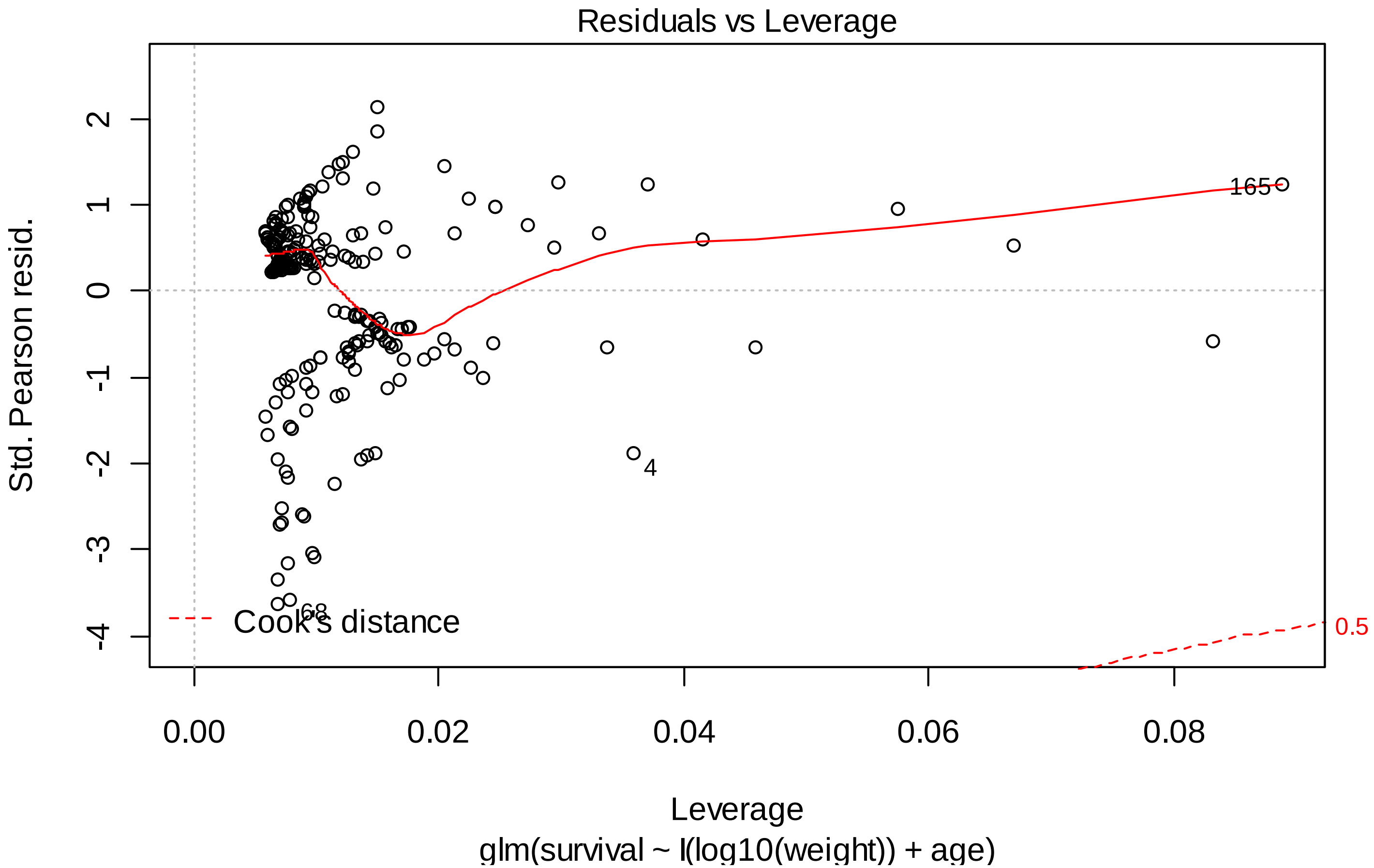
```
xx <- predict(fit, type="response")
yy <- residuals(fit, type="pearson")
loess.smooth(xx, yy, family="gaussian", pch=20)
abline(h=0, lty=3)
```

Reasons:

- using a non-robust smoother is a must
- different types of residuals can be used
- on the x-axis: probs or linear predictor

Applied Statistical Regression
AS 2012 – Week 11

More Diagnostics



Binomial Regression Models

Concentration in log of mg/l	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

→ for the number of killed insects, we have $Y_i \sim Bin(n_i, p_i)$

→ we are mainly interested in the proportion of insects surviving

→ these are grouped data: there is more than 1 observation for a given predictor setting

Model and Estimation

The goal is to find a relation:

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

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 11

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

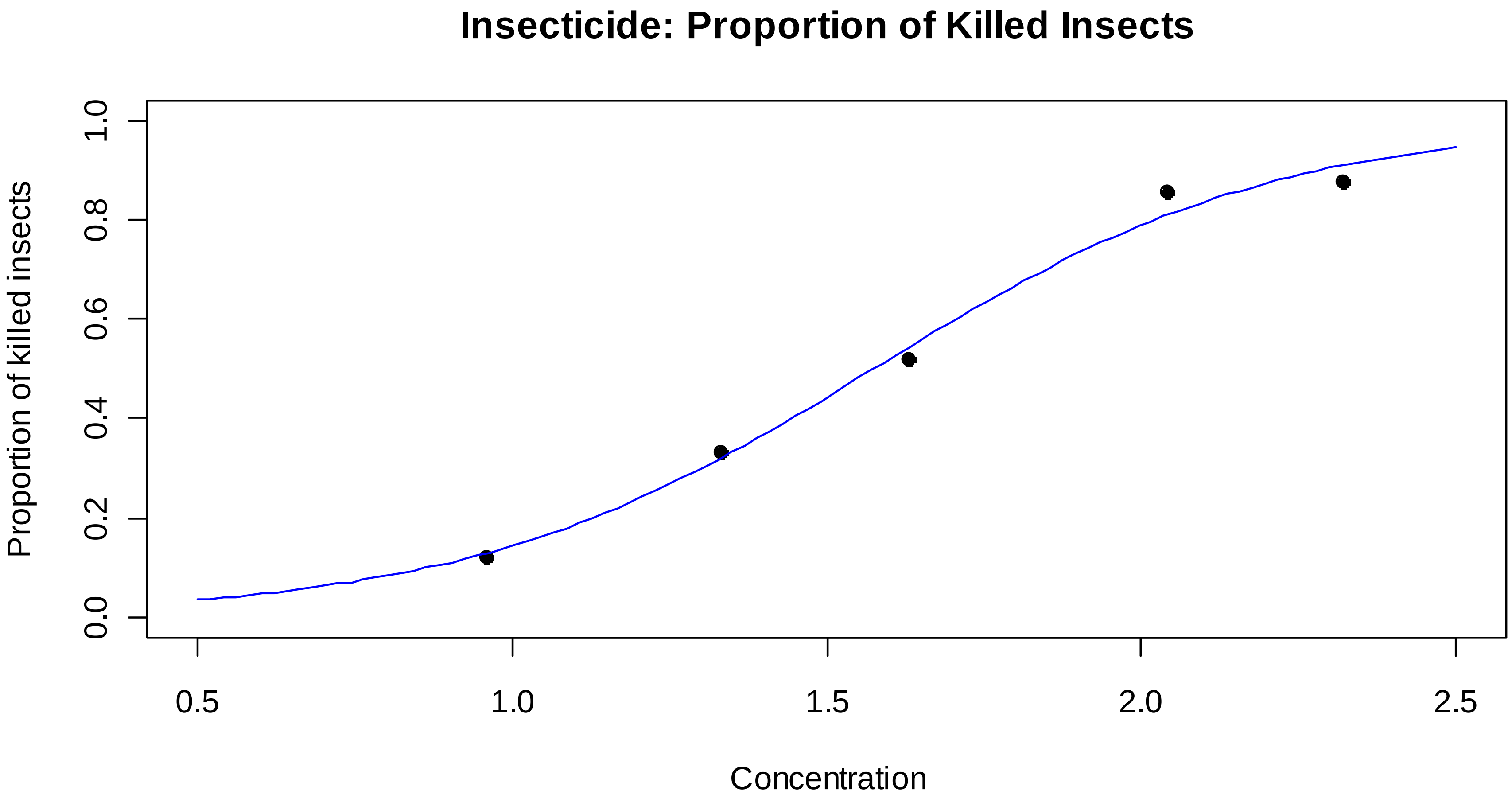
The result for the insecticide example is:

```
> 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 11

Proportion of Killed Insects



Applied Statistical Regression

AS 2012 – Week 11

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

Goodness-of-Fit Test

→ **the residual deviance will be our goodness-of-fit measure!**

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 deviance measures how close our model comes to perfection.

Applied Statistical Regression

AS 2012 – Week 11

Evaluation of the Test

Asymptotics:

If Y_i is truly binomial and the n_i are large, the deviance is approximately χ^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 $df \pm 2\sqrt{df}$

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

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

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

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
```

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

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:

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