Applied Statistical Regression

HS 2010 – Week 10

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, November 29, 2010

Applied Statistical Regression
HS 2010 – Week 10

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:

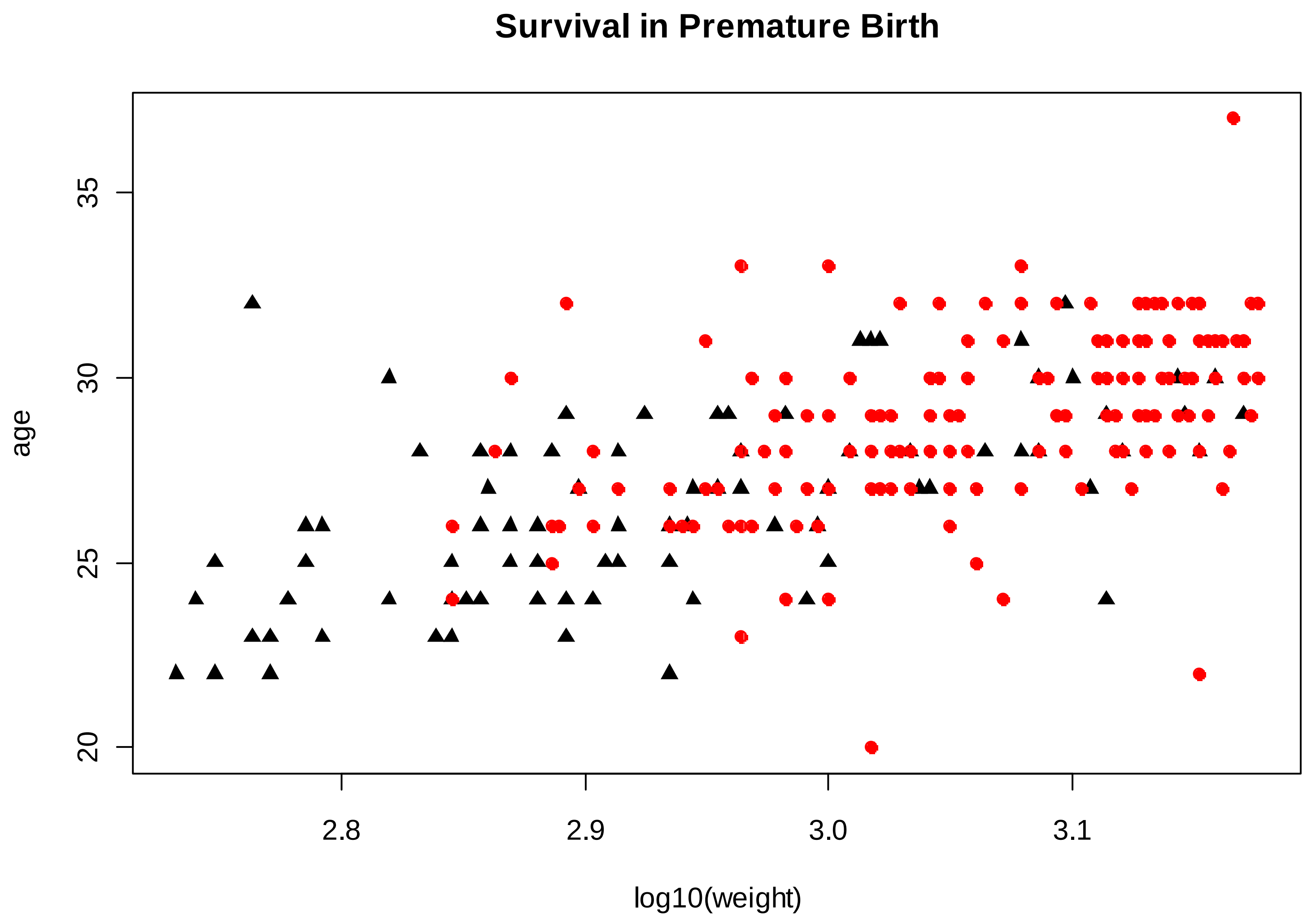
$$p_i = P(Y_i = 1) = E[Y_i]$$

→ 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: $p_i = \beta_0 + \beta_1 x_{i1} + ... + \beta x_{ip}$ is no good here!

Applied Statistical Regression
HS 2010 – Week 10

Example



Inference with 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

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:

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

Example:

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

Applied Statistical Regression
HS 2010 – Week 10

Model Diagnostics

Diagnostics are:

- 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
HS 2010 – Week 10

Pearson Residuals

Take the difference between observed an fitted value and divide 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 *i*th 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

Applied Statistical Regression
HS 2010 – Week 10

Deviance Residuals

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

$$d_i = -2 \cdot \left(y_i \log\left(\hat{p}_i \right) + (1 - y_i) \log\left(1 - \hat{p}_i \right) \right)$$

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

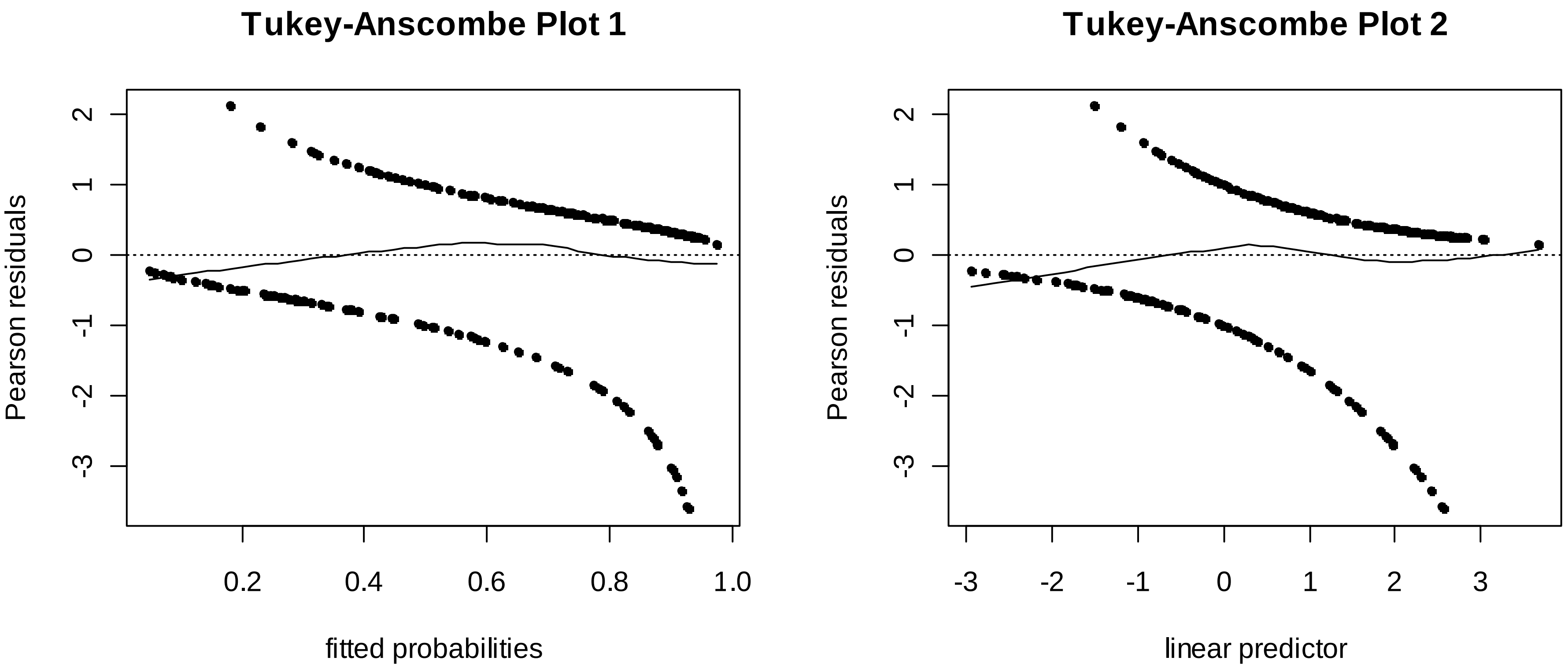
$$D_i = sign(y_i - \hat{p}_i) \cdot \sqrt{d_i}$$

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

Applied Statistical Regression
HS 2010 – Week 10

Tukey-Anscombe Plot

Remark: sometimes studentized residuals are used!



Tukey-Anscombe Plot

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

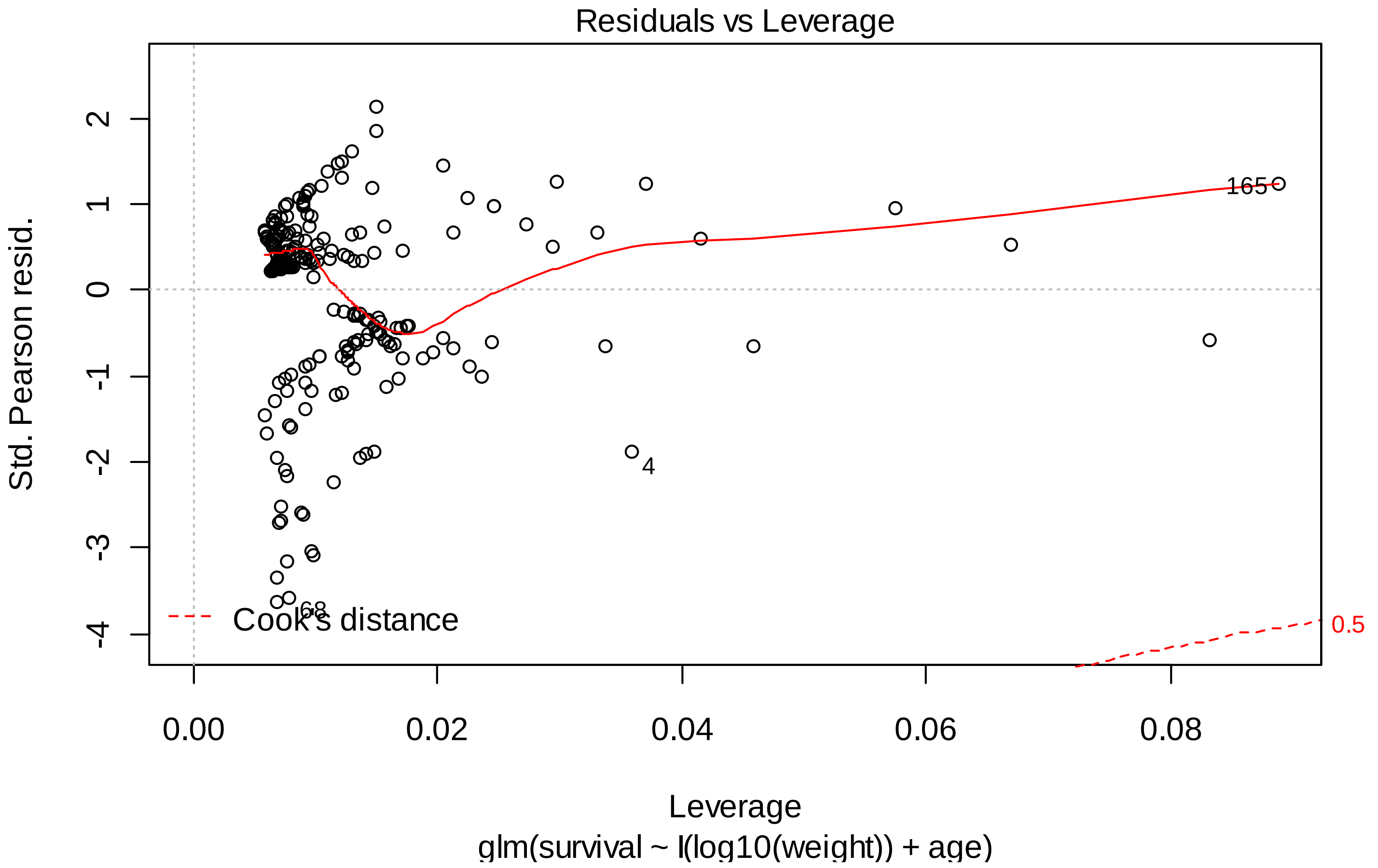
```
xx <- predict(fit, type="response")
yy <- residuals(fit, type="pearson")
scatter.smooth(xx, yy, family="gaussian", pch=20)
abline(h=0, lty=3)
```

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
HS 2010 – Week 10

More Diagnostics



Applied Statistical Regression
HS 2010 – Week 10

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

Applied Statistical Regression
HS 2010 – Week 10

Model and Estimation

The goal is to find a relation:

$$p_i = P(Y_i = 1 \mid x_1, \ldots, x_p) \;\sim\; \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

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:

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

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")
```

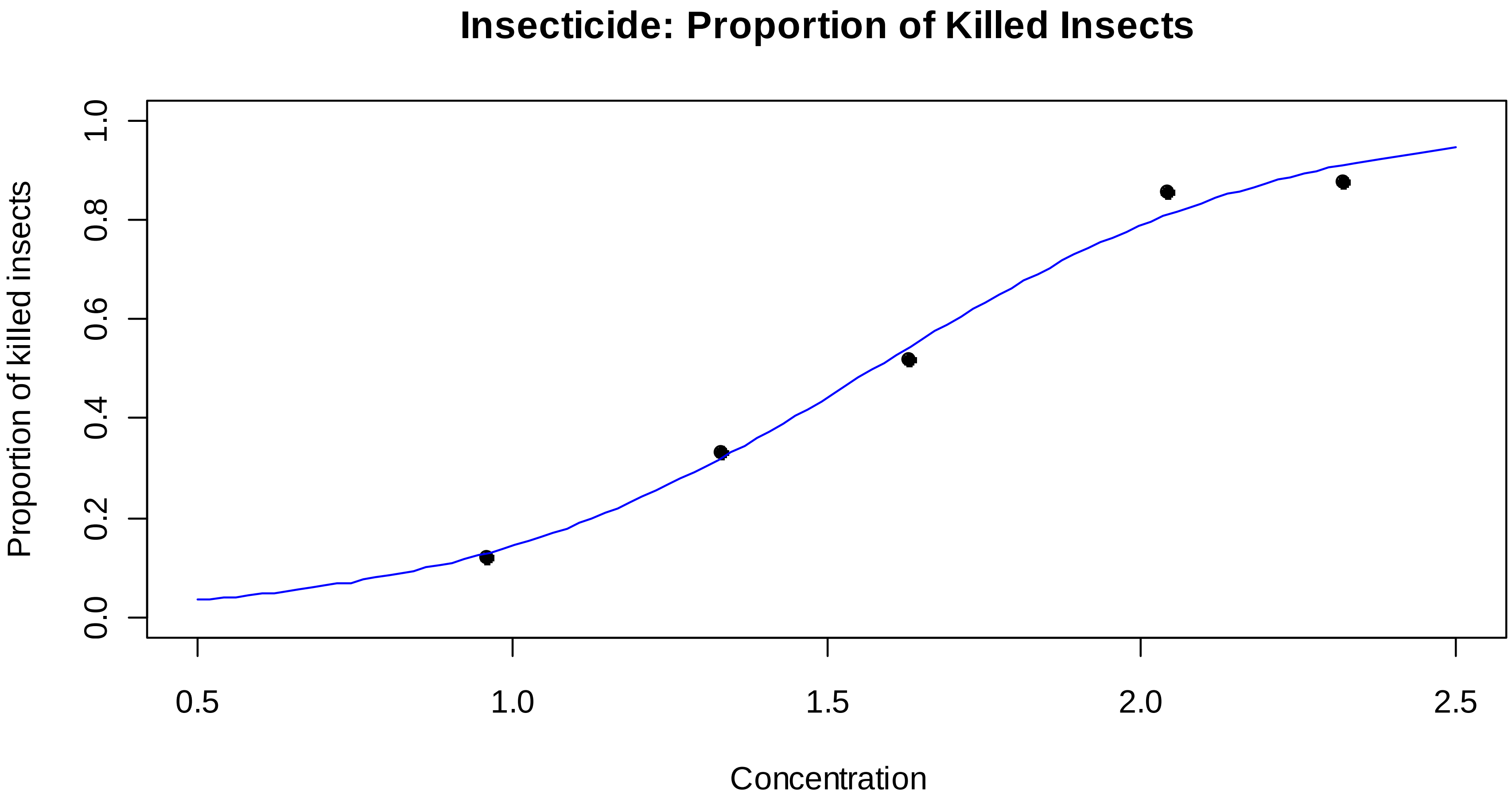
Summary Output

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
```

Proportion of Killed Insects



Inference with 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
HS 2010 – Week 10

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.

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
```

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!