Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Week 11: Modelling binary data
11./12. May 2017

Overview (todo: check)

- Binary response variables
- Contingency tables, χ^2 test
- Odds and (log) odds ratios
- Logistic regression
- Residual analysis / model checking / deviances
- Interpretation of the results

Course material covered today

- Repetition: Chapter 10 in the Stahel book from last semester.
- (Ev? Stahel GLM Script, (parts of) chapters 7 and 8)
- Chapters 9.1 - 9.3 from “The new statistics with R”.

Recap of last week: GLMs and Poisson regression

- We introduced generalized linear models (GLMS) and key terms:
  Family **Linear predictor** **Link function**

- GLMs are useful when the response variable $\mathbf{y}$ is not continuous (→ residuals are not Gaussian).

- Count data usually lead to Poisson regression.

- However, for count data it may sometimes be ok to use linear regression with $\log(\mathbf{y})$ in the response.

Introduction

- Today, we will look at the case where the response variable is binary (0 or 1) or binomial (*e.g.* 5 out of 7 trials).
- In binary/binomial regression, the question will be: "Which variables influence the probability p of the outcome?"

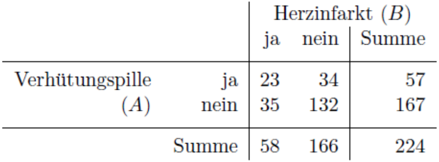
Examples:

- Outcome: Heart attack (yes=1, no=0).
  Question: which variables lead to higher or lower risk of heart attack?
- Outcome: Survival (yes=1, no=0).
  Question: which variables influence the survival probability of premature babies (Frühgeburten)?

Some repetition: The χ^2 test

You have dealt with binary (categorical) data in Mat183! Remember the χ^2 test...

Example: Heart attack and hormonal contraception (Verhütungspille), see Stahel 7.3.j:



		Herzinfarkt (B)		
		ja	nein	Summe
Verhütungspille	ja	23	34	57
(A)	nein	35	132	167
	Summe	58	166	224

"Hormonal contraception" is the predictor (x) and "heart attack" the outcome (y).

Question: Does hormonal contraception (x) have an influence on heart attacks (y)?

This question is equivalent to asking whether the proportion of patients with heart attack is the same in both groups.

The respective test-statistic can be calculated as

$$T = \sum_{\text{all entries}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \ .$$

By hand, T is obtained as

$$\frac{(23 - 14.8)^2}{14.8} + \frac{(34 - 42.2)^2}{42.2} + \frac{(35 - 43.2)^2}{43.2} + \frac{(132 - 123.8)^2}{123.8} = 8.329$$

and is expected to be χ^2_1 distributed (one degree of freedom: $(2 - 1) \cdot (2 - 1)$).

The p-value of this test is given as $\Pr(X \geq 8.329) = 0.003902$.

```
> 1-pchisq(8.329,1)

[1] 0.003901713
```

Of course, R can do this for us:

```
> contYes <- c(23,34)
> contNo <- c(35,132)
> data.table <- data.frame(rbind(contYes,contNo))
> chisq.test(data.table,correct=FALSE)

	Pearson's Chi-squared test

data:  data.table
X-squared = 8.3288, df = 1, p-value = 0.003902
```

Please note: You would usually prefer to have `correct=TRUE` to obtain a better approximation to the χ^2 distribution (continuity correction of Yates):

```
> chisq.test(data.table,correct=TRUE)

	Pearson's Chi-squared test with Yates' continuity correction

data:  data.table
X-squared = 7.3488, df = 1, p-value = 0.006711
```

In any case, there is strong evidence for an association of hormonal contraception with heart attacks!

Quantification of a dependency

(Stahel GLM chapter 7.4)

If two variables are not independent, it is often desired to quantify the dependency.

Let one variable be the grouping variable (e.g., hormonal contraception vs no hormonal contraception). Then π_1 and π_2 are the relative frequencies (proportions) observed in the two groups. For example:

$$\begin{aligned} \pi_1 = 23/57 &= 0.404 \\ \pi_2 = 35/167 &= 0.210 \end{aligned}$$

are the proportions of females with a heart attack in the two groups.

There are at least three numbers that can be calculated to quantify how the two groups differ:

- Risk difference: $\pi_1 - \pi_2 = 0.404 - 0.210 = 0.194$
- Relative risk: $\pi_1/\pi_2 = 0.404/0.210 = 1.92$
- Odds ratio ("Chancenverhältnis"):

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{0.404/(1-0.404)}{0.210/(1-0.210)} = 2.55 \ ,$$

where $\pi/(1-\pi)$ is the chance (die "Chance").

Interpretation:

1. $OR = 1 \rightarrow$ the two groups are independent.
2. $OR > 1 \rightarrow$ positive dependency.
3. $OR < 1 \rightarrow$ negative dependency.

The odds and the odds ratio

- The **odds** (“Wetteverhältnis”): For a probability π the odds is $\pi/(1-\pi)$. For example, if the probability to win a game is 0.75, then the odds is given as 0.75/0.25 or 3:1.

- The **odds ratio** is given on the previous slide. It is a ratio of two ratios, or, the **ratio of two odds**.

- Often the **log odds ratio** is used:

$$\log(OR) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) .$$

  Why is this simpler? Look at the interpretation:

  1. $\log(OR) = 0 \rightarrow$ the two groups are independent.
  2. $\log(OR) > 0 \rightarrow$ positive dependency.
  3. $\log(OR) < 0 \rightarrow$ negative dependency.

Binomial and binary regression

Usually the situation is more complicated than

binary covariate $(\mathbf{x}) \rightarrow$ **binary outcome** $(\mathbf{y})$

Often, we are interested in a relationship

Continuous/categ./binary variables $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$,.. $\rightarrow$ **binary outcome** $(\mathbf{y})$

→ A regression model is needed again!

Illustrative/working example

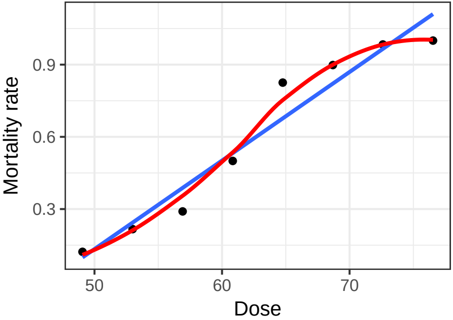
Let us look at an example from chapter 9.2 in Hector (2015):

Eight groups of beetles were exposed to carbon disulphide (an insecticide) for 5h. For each beetle it was then reported if it was killed or not (1 or 0), but the data were reported aggregated:

```
   Dose Number_tested Number_killed Mortality_rate
1 49.06            49             6      0.1224490
2 52.99            60            13      0.2166667
3 56.91            62            18      0.2903226
4 60.84            56            28      0.5000000
5 64.76            63            52      0.8253968
6 68.69            59            53      0.8983051
7 72.61            62            61      0.9838710
8 76.54            60            60      1.0000000
```

Question: (How) does the insecticide affect the survival of the beetles?

As always, start with a graph:



with linear (blue) and smoothed line (red).

What can we see from the plot?

- Mortality increases with higher doses of the herbicide (not surprising, right?).
- The linear line seems unreasonable. In particular, if one extrapolates to lower or higher doses, mortality would become < 0 or > 1, which is not possible. (Remember: A probability is between 0 and 1 by definition.)

How does one analyze these data correctly?

- So far, we know linear and Poisson regression.
- Both of these are not the correct approaches here.

The ‘wrong’ analyses

Linear regression

We could simply use

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 Dose_i$$

with $\mathrm{E}(y_i) = \pi_i =$ probability to die for individuals i with $Dose_i$.

R does this analysis without complaining (!!):

```
> lm(Mortality_rate ~ Dose, data=beetle)
```

This leads to $\hat{\beta}_0 = -1.71$ and $\hat{\beta}_1 = 0.037$. This means for instance that, for a zero dose, the probability to die would be $\mathrm{E}(y_i) = -1.71$.

Problem: Linear regression leads to impossible predicted probability values!
$\Rightarrow$ Unrealistic predictions!

Poisson regression

What about Poisson regression with the counts "`Number_killed`" in the response? We could use

$$\log(\mathrm{E}(y_i)) = \beta_0 + \beta_1 Dose_i$$

with $\mathrm{E}(y_i)$ = number killed. Again, R does this analysis without complaining (!!):

```
> glm(Number_killed ~ Dose, data=beetle,family=poisson)
```

This leads to $\hat{\beta}_0 = -0.77$ and $\hat{\beta}_1 = 0.067$.

Problem: This means for instance that, for a dose of 76, one expects that $\mathrm{E}(y_i) = \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 76) = 73.80$ beetles die. However, there are only around 60 beetles in each group, so the predicted number killed is more than what is available. ⇒ Unrealistic predictions!

A model for binary data?

I hope you remember the Bernoulli distribution from Mat183:

The probability distribution of a binary random variable $Y \in \{0, 1\}$ with parameter π is defined as

$$\Pr(Y = 1) = \pi \ , \quad \Pr(Y = 0) = 1 - \pi \ .$$

Characteristics of the Beroulli distribution:

- $\mathrm{E}(Y) = \pi$
- $\mathrm{Var}(Y) = \pi(1 - \pi)$.

  $\rightarrow$ The variance of the distribution is determined by its mean.

Doing it right: Logistic regression

We can again use the GLM machinery from last week! The linear predictor is again (as always):

$$\eta_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)} \ .$$

We again need a link function that relates the linear predictor η_i to the expected value $\mathrm{E}(y_i)$.

Remember we used the log link last week, but that seems a bad idea here (see slide 17).

The link function must be chosen such that the expected value $\mathrm{E}(y_i)$ is always between 0 and 1!

Link function: The logistic transformation

Your turn!

```
> path <- "../../data_examples/WBL/"
> d.heart <- read.table(paste(path,"heart.dat",sep=""),header=T,sep=",")
> data.table

        heartYes heartNo cont
contYes       23      34    1
contNo        35     132    0

> r.heart.glm <- glm(cbind(heartYes,heartNo) ~ cont,data.table,family=binomial)
> summary(r.heart.glm)

Call:
glm(formula = cbind(heartYes, heartNo) ~ cont, family = binomial,
    data = data.table)

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3275     0.1901  -6.982 2.91e-12 ***
cont          0.9366     0.3302   2.836  0.00456 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.8676e+00  on 1  degrees of freedom
Residual deviance: 3.7303e-14  on 0  degrees of freedom
AIC: 13.629

Number of Fisher Scoring iterations: 3
```

Summary