Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 10: 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.
- 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 **y** is not continuous ($\rightarrow$ 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 for contingency tables (simplest: 2×2 tables).

Example: Heart attack and hormonal contraception (Verhütungspille) (from Stahel):

		Herzinfarkt (B)		
		ja	nein	Summe
Verhütungspille (A)	ja	23	34	57
	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

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 odds (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

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

Often, we are interested in a relationship

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

$\rightarrow$ A regression model is needed again!

Illustrative/working example

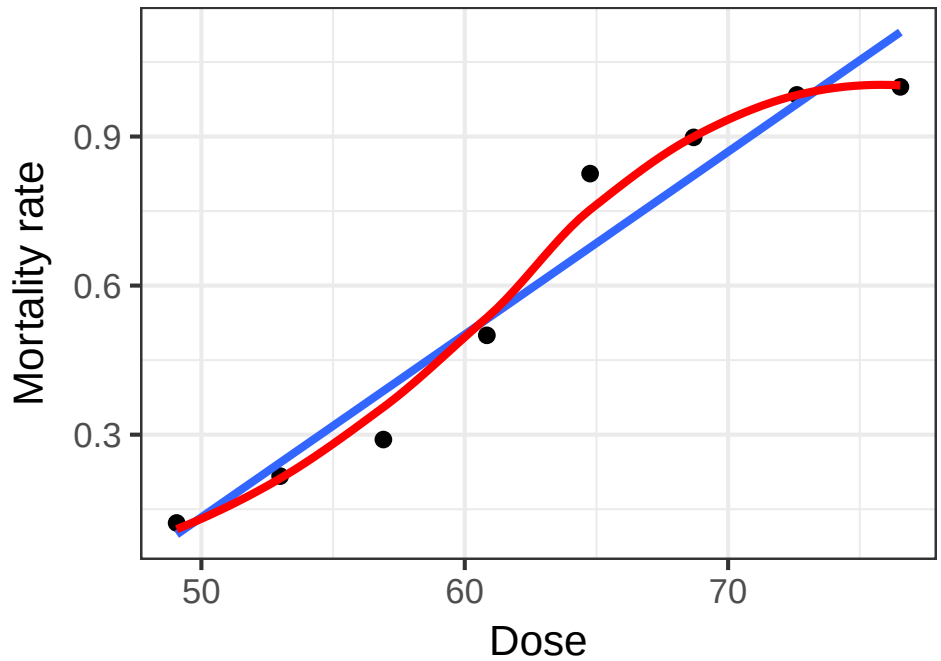
Let us look at an example from chapter 9.2 in Hector (2015):

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 (**x**) affect the survival probability (**y**) 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 $\mathrm{E}(y_i) = -1.71$.

Problems:

- Linear regression leads to impossible predicted probability values!
  $\Rightarrow$ Unrealistic predictions!
- For $y_i = \beta_0 + \beta_1 Dose_i + e_i$, residuals e_i are **not** normally distributed!

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, although these are not 'real' counts:

```
> 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. $\Rightarrow$ Unrealistic predictions!

Sidenote: count vs. binomial data

Clarification of the difference between count data and binomial data:

Count data:

- Theoretically no upper limit on number of times an "event" (e.g., number of birds observed in a forest plot)
- Counts cannot be expressed as a proportion.

Binomial data:

- Aggregated version of many binary experiments, that is, each can be 0 or 1.
- Therefore, there is an upper limit on the number of times an "event" can be observed (e.g., number of deaths cannot be greater than number of living individuals).
- Successes can be expressed as proportion (number of successes/number of trials).

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:

- $\mathrm{E}(Y) = \pi = \Pr(Y = 1)$ (useful to remember)
- $\mathrm{Var}(Y) = \pi(1 - \pi)$.

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

From binary to binomial data

Binomial data is an aggregation of binary data:

- Repeate the experiment with $\Pr(Y = 1) = \pi$ a total number of n times, calculate how often a success was observed (k times).
- The expected proportion of successes ("success rate", here k/n) has then the same expectation as the success probability of a single experiment:

$$\mathrm{E}\left(\frac{\sum_{i=1}^{n} Y}{n}\right) = \pi = \mathrm{E}(Y) \ .$$

Example: In the beetle data 49 beetles were tested for the lowest dose, of which $k = 6$ died.

Doing it right: Logistic regression

We can again use the GLM machinery from last week! The linear predictor is 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).

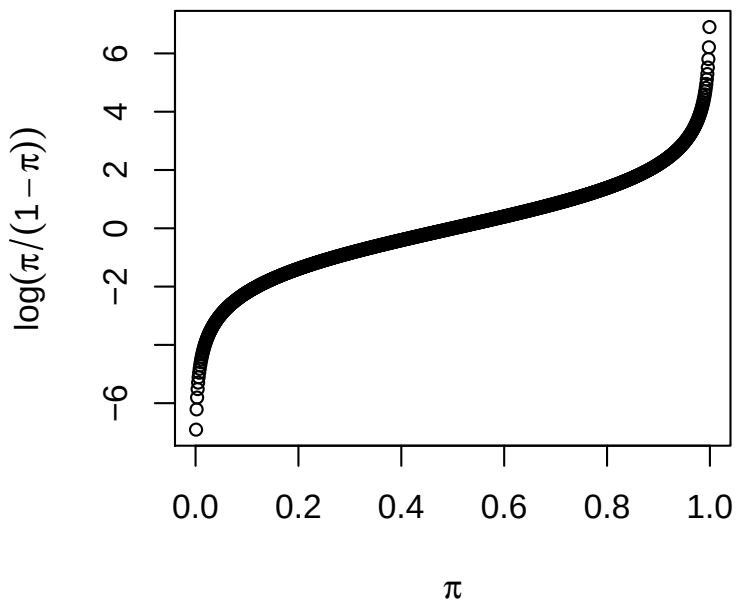
The link function must be chosen such that the expected value $\mathrm{E}(y_i)$ is always between 0 and 1!

Link function: The logistic transformation

A transformation that assigns a probability (π) between 0 and 1 a value between $-\infty$ and ∞ is the logit-transformation:

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\pi) - \log(1-\pi) \ .$$

A graph depicts the functional form of $g(\cdot)$:



The logistic regression model

In order to prevent the expected value $\mathrm{E}(y_i)$ to attain unreasonable values, we thus formulate the logistic regression model as

$$\log\left(\frac{\mathrm{E}(y_i)}{1-\mathrm{E}(y_i)}\right) = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)}$$

with $\mathrm{E}(y_i) = \Pr(y_i = 1)$.

- The link function is called the **logistic link**.
- The family is **binomial**.

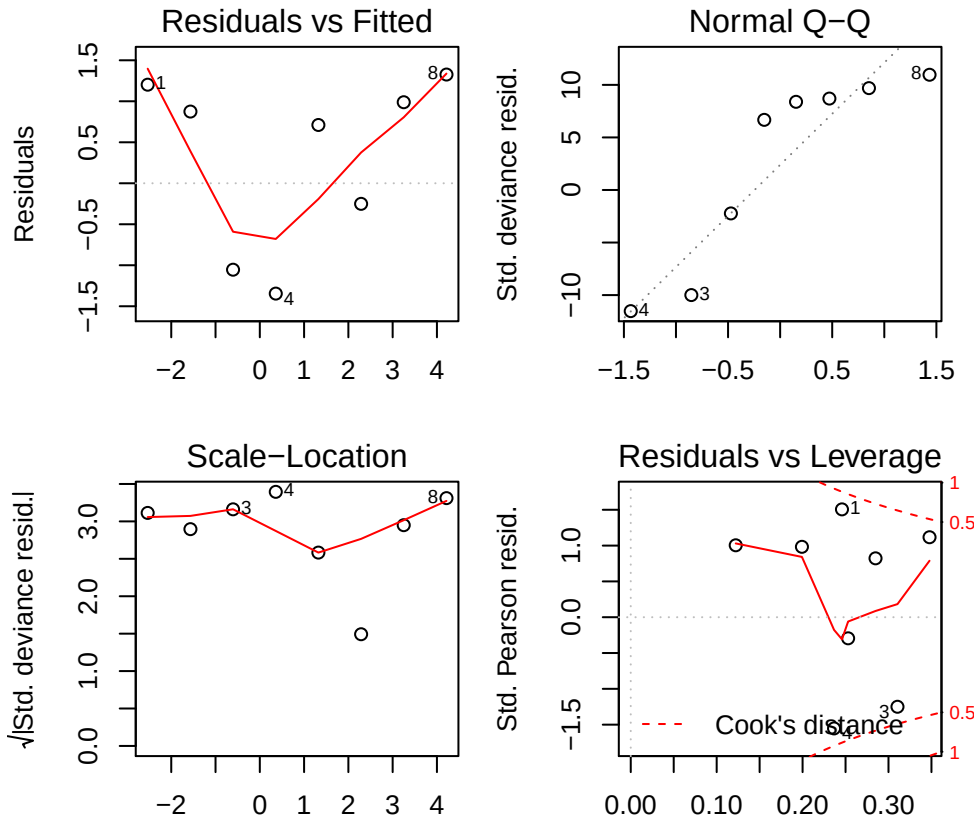
Doing it right: Fitting a logistic regression

- As for the Poisson GLM, we can estimate the parameters $\beta_0, \beta_1, \ldots$ by maximum-likelihood estimation.
- Luckily, the `glm()` function in R can also handle binomial and binary data!
- A complication comes from the fact that we need to tell the function two numbers for the response:
  - The number of successes, encoded as 1 (here: number killed)
  - The number of failures, encoded as 0 (here: number died)

```
> beetle$Number_survived <- beetle$Number_tested - beetle$Number_killed
> beetle.glm <- glm(cbind(Number_killed,Number_survived) ~ Dose,
+               data = beetle, family = binomial)
```

Doing it right: Model diagnostics

As always, before looking at the regression output, let's do some model diagnostics:



- As in Poisson regression, it is not clear how do define residuals, there are many ways...
- Again, different types of residuals are used in the plots.
- Be careful: such plots are only reasonable for aggregated data (which we have here)! The larger the groups, the more precise are the underlying assumptions (approximate equality of distributions).
- See example on slide 40 for an example with non-aggregated (binary) data.

Doing it right: Interpreting the coefficients

Let's look at the results, for the moment on the coefficients:

```
> summary(beetle.glm)$coef
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -14.5780604  1.2984622 -11.22717 2.999201e-29
Dose          0.2455399  0.0214937  11.42381 3.179900e-30
```

The intercept and slope are estimated as

$$\hat{\beta}_0 = -14.578 \quad \text{and} \quad \hat{\beta}_1 = 0.246 \ ,$$

with standard errors and p-values. Very clearly, the dose influences the survival probability ($p << 0.001$), and $\hat{\beta}_1 > 0$, thus, the larger the dose, the larger the mortality probability (positive relation).

This is a **qualitative interpretation** of the coefficients.

Note: The β coefficients are approximately normally distributed as $\mathrm{N}(\hat{\beta}, \hat{\sigma}^2_\beta)$.
$\rightarrow$ confidence intervals etc. can be calculated as in the linear case!

Quantitative interpretation of the coefficients

Remember the regression model

$$\log\left(\frac{\mathrm{E}(y_i)}{1-\mathrm{E}(y_i)}\right) = \beta_0 + \beta_1 \mathit{Dose}_i \ .$$

To understand what β_1 tells us, let's transform the equation such that $\mathrm{E}(y_i) = \Pr(y_i = 1)$ is on the left. Some algebra leads to

$$\Pr(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 \mathit{Dose}_i)}{1 + \exp(\beta_0 + \beta_1 \mathit{Dose}_i)} \ . \tag{1}$$

It is then possible to calculate the odds ("Chance")

$$\mathrm{odds}(y_i = 1 \,|\, \mathit{Dose}_i) = \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} = \exp(\beta_0 + \beta_1 \mathit{Dose}_i) \ .$$

If the $Dose_i$ is then increased by 1 unit in concentration (from x to $x+1$), the odds ratio is given as

$$\frac{\text{odds}(y_i = 1 \,|\, Dose_i = x+1)}{\text{odds}(y_i = 1 \,|\, Dose_i = x)} = \exp(\beta_1) = \exp(0.246) = 1.28 \; .$$

Interpretation: When the dose is increased by 1 unit, the odds to die increases by a factor of 1.28.

Moreover, taking the log on the above equation shows that β_1 can be interpreted as a log odds ratio:

$$\beta_1 = \log\left(\frac{\text{odds}(y_i = 1 \,|\, Dose_i = x+1)}{\text{odds}(y_i = 1 \,|\, Dose_i = x)}\right)$$

Doing it right: The `anova()` table

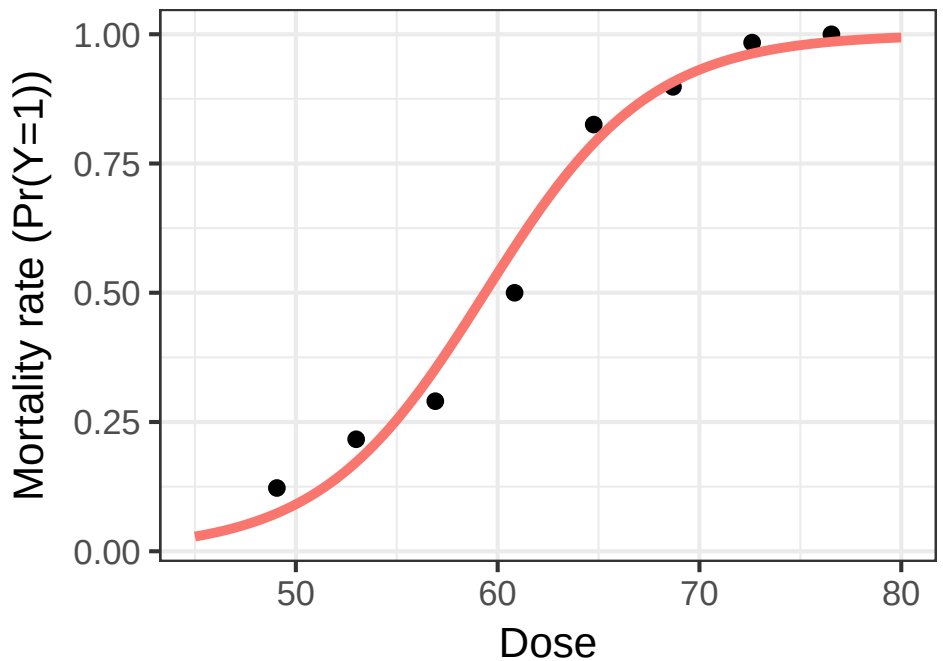
We can look at the Analysis of Deviance table (directly using `test="Chisq"`):

```
> anova(beetle.glm,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Number_killed, Number_survived)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     7    267.662
Dose  1   259.23         6      8.438 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: The total deviance is 267.66, and of this 259.23 is explained by `Dose` (using 1 degree of freedom). This seems really good, because the χ^2 test gives a p-value that is reeeeallly small ($< 2.2e-16$).

Plotting the fit

A fitted curve can be added to the raw data by plotting $\Pr(y_i = 1)$ against the Dose, using equation (1):



Overdispersion

Remember :

- Slide19: $\mathrm{E}(Y) = \pi$ and $\mathrm{Var}(Y) = \pi(1-\pi)$, thus the variance is determined by the mean!
- "Overdispersion" means "extra variability" (larger than the model predicts).
- Reason: Variables are missing in the model!
- Overdispersion leads to too small *p*-values.
- Detectable by looking at the residual deviance:

  `Residual deviance` $>>$ `df` $\quad \rightarrow$ Overdispersion

- Also possible: underdispersion (dependency in the data), if:

  `Residual deviance` $<<$ `df`

Here, the residual deviance is **8.44** with **6** degrees of freedom. Is this good or bad?

```
> 1-pchisq(8.438,6)
```

```
[1] 0.2077375
```

$\rightarrow$ $p = 0.21$ seems not problematic.

One can nevertheless account for overdispersion by switching to a `quasibinomial` model. This allows to estimate the dispersion parameter separately.

```
> beetle.glm2 <- glm(cbind(Number_killed,Number_survived) ~ Dose,
+                data = beetle, family = quasibinomial)
> summary(beetle.glm2)

Call:
glm(formula = cbind(Number_killed, Number_survived) ~ Dose, family = quasibinomial,
    data = beetle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3456  -0.4515   0.7929   1.0422   1.3262

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.57806    1.46611  -9.943 5.98e-05 ***
Dose          0.24554    0.02427  10.118 5.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.274895)

    Null deviance: 267.6624  on 7  degrees of freedom
Residual deviance:   8.4379  on 6  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Binary response / non-aggregated data

- In the beetle example, we were in a comfortable situation: For each level of the does, we had several beetles. For instance, 55 beetles at lowest does (49.06), of which 6 died (1) and 49 survived (0). This was binomial data, an aggregated version of 55 trials with 0 or 1 outcome.
- In reality, one often has only one trial (0/1) for a (combination of) covariate(s).
- The analysis is the same as for aggregated data, however there are a few complications with graphical descriptions and model checking.

Example: Blood screeing (see week 1; data from Hothorn & Everitt 2014, chapter 7.3)

Blood screening example

```
> library(HSAUR3)
> data("plasma",package="HSAUR3")
> plasma$y <- as.integer(plasma$ESR)-1
```

fibrinogen	globulin	ESR	y	fibrinogen	globulin	ESR	y
2.52	38	ESR $<$ 20	0	2.88	30	ESR $<$ 20	0
2.56	31	ESR $<$ 20	0	2.65	46	ESR $<$ 20	0
2.19	33	ESR $<$ 20	0	2.28	36	ESR $<$ 20	0
2.18	31	ESR $<$ 20	0	2.67	39	ESR $<$ 20	0
3.41	37	ESR $<$ 20	0	2.29	31	ESR $<$ 20	0
2.46	36	ESR $<$ 20	0	2.15	31	ESR $<$ 20	0
3.22	38	ESR $<$ 20	0	2.54	28	ESR $<$ 20	0
2.21	37	ESR $<$ 20	0	3.34	30	ESR $<$ 20	0
3.15	39	ESR $<$ 20	0	2.99	36	ESR $<$ 20	0
2.60	41	ESR $<$ 20	0	3.32	35	ESR $<$ 20	0
2.29	36	ESR $<$ 20	0	5.06	37	ESR $>$ 20	1
2.35	29	ESR $<$ 20	0	3.34	32	ESR $>$ 20	1
3.15	36	ESR $<$ 20	0	2.38	37	ESR $>$ 20	1
2.68	34	ESR $<$ 20	0	3.53	46	ESR $>$ 20	1
2.60	38	ESR $<$ 20	0	2.09	44	ESR $>$ 20	1
2.23	37	ESR $<$ 20	0	3.93	32	ESR $>$ 20	1

Question: Is a high ESR (erythrocyte sedimentation rate) an indicator for certain diseases (rheumatic disease, chronic inflammations)?

Specifically: Is there an association between ESR level ESR$< 20 mm/hr$ and the concentrations of the plasma proteins Fibrinogen and Globulin?

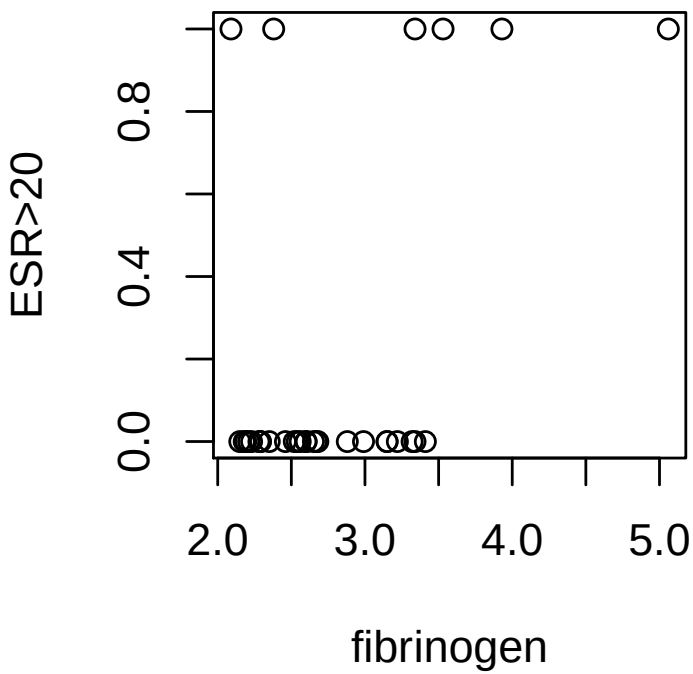
The model to be fitted:

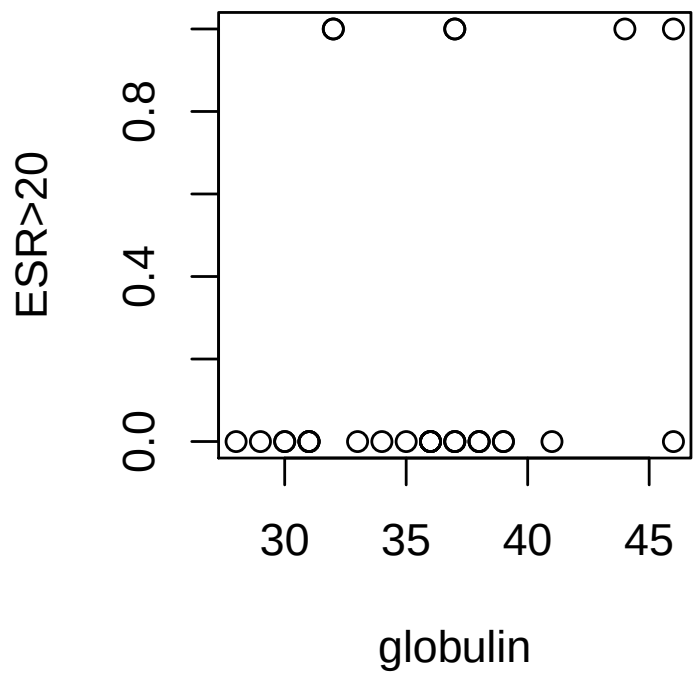
$$\log\left(\frac{\mathsf{E}(y_i)}{1-\mathsf{E}(y_i)}\right) = \beta_0 + \beta_1 \mathit{fibrinogen}_i + \beta_2 \mathit{globulin}_i \ ,$$

with $\mathsf{E}(y_i) = \mathsf{Pr}(y_i = 1)$.

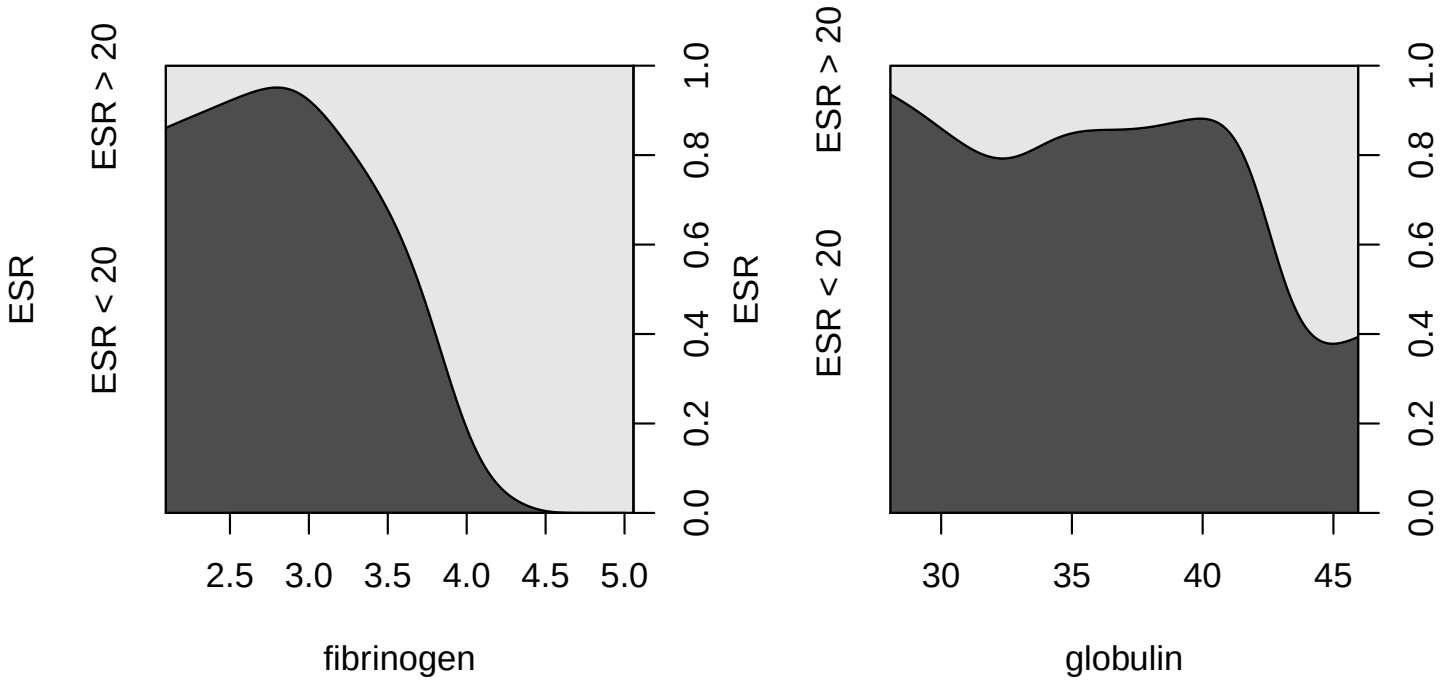
Complication 1 with binary data: Graphical description

Plotting the response y (indictor for ESR$>$ 20) against the covariats does not lead to very illustrative graphs:





It is a bit more illustrative to give a conditional density plot:

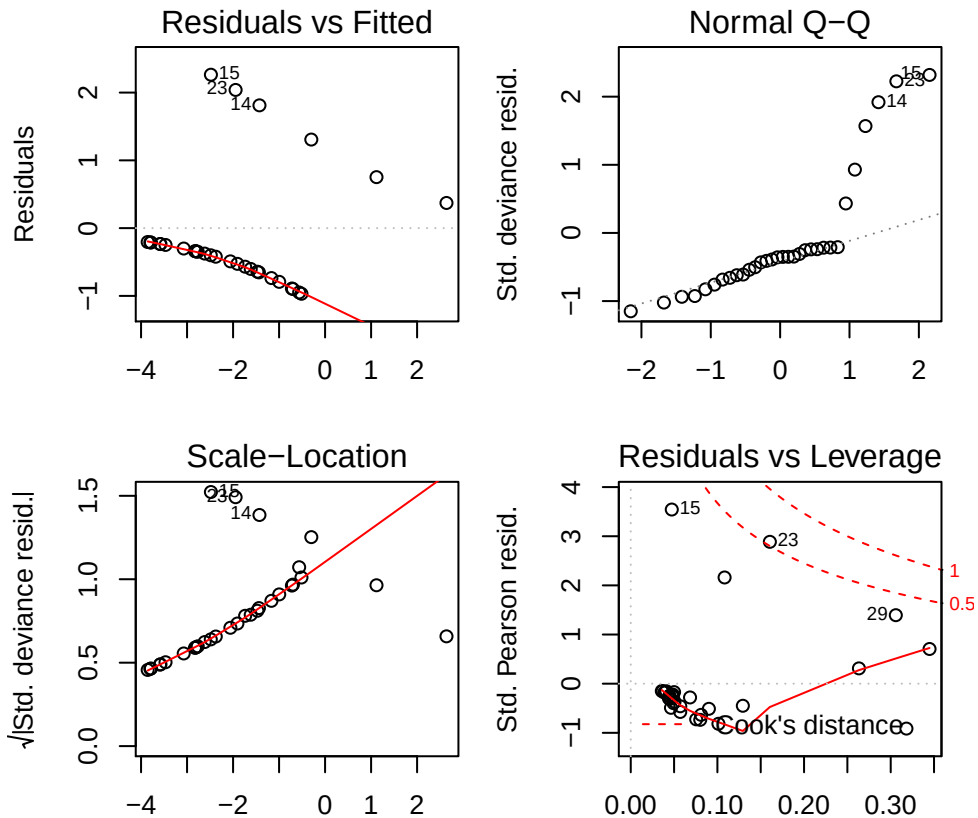


Complication 2: Model diagnostics

a) Residuals plots:

Plotting the residuals is possible, but not meaningful.

Why? Becasue the model checking assumptions rely on aggregated data!



b) Residual deviance:

For non-aggregated data, the `residual deviance` vs. `df` relation cannot be used to detect overdispersion!!

Your turn!

Apart from the above complications, fitting and interpreting the model is analogous to aggregated binary data. Let's continue with the blood screening example:

```
> plasma.glm <- glm(y ~ fibrinogen + globulin, data = plasma, family=binomial)
```

Please look at the model outcomes (summary and anova table) on the next slides and answer the following questions:

1. Is there an an effect of figrinogen and/or globulin on the outcome (ESR$>$ 20)?
2. What is the *quantitative* interpretation of the β_1 coefficient (what happens to Pr($ESR > 20$) when fibrinogen incrases by 1 unit)?
3. Is a `quaisbinomial` model more suitable for these data?

```
> summary(plasma.glm)

Call:
glm(formula = y ~ fibrinogen + globulin, family = binomial, data = plasma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9683  -0.6122  -0.3458  -0.2116   2.2636

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.7921     5.7963  -2.207   0.0273 *
fibrinogen    1.9104     0.9710   1.967   0.0491 *
globulin      0.1558     0.1195   1.303   0.1925
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 22.971  on 29  degrees of freedom
AIC: 28.971

Number of Fisher Scoring iterations: 5
```

```
> anova(plasma.glm,test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                          31     30.885
fibrinogen  1   6.0446        30     24.840  0.01395 *
globulin    1   1.8692        29     22.971  0.17156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary