Lecture 11: Modeling binary data

BIO144 Data Analysis in Biology

Stephanie Muff, Owen Petchey & Erik Willems

University of Zurich

13 May, 2024

Overview

- Generalized Linear Models (GLMs)
  - Count outcome $\rightarrow$ Poisson regression
  - Binary outcome $\rightarrow$ logistic regression
- Contingency tables, χ^2-test
- Odds and (log) odds ratios
- Residual analysis / model checking / deviances
- Interpretation of the results

Course material covered today

The lecture material of today is based on the following literature:

- Repetition: Chapter 9.4 about χ^2-Tests in the Luchsinger script
- Chapters 9.1 - 9.3 from *The new statistics with R* (Hector book).

Note: on OLAT you'll also find the continuation of the Stahel script, chapters 7-9 that cover GLMs. This is **not** mandatory literature.

Recap of last week: GLMs and Poisson regression

- We encountered Generalized Linear Models (GLMS) and their key components:
  - Linear predictor
  - Family
  - Link function
- GLMs are useful when the response variable y is not continuous, or more generally, when we can't assume that residuals follow a Gaussian distribution.
- Count data (without known maximum) can be modelled using Poisson regression.

Introduction

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

Examples:

- Outcome: Heart attack (yes= 1, no= 0).
  Question: which variables affect the 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 may recall binary (categorical) data from Mat183, where you encountered the χ^2 test for contingency tables (simplest case is a 2 x 2 table).

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) influence the risk of heart attack (y)?

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

The test-statistic can be calculated as:

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

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

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

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

```
pchisq(8.329, 1, lower.tail= F)
```

```
## [1] 0.003901713
```

$\rightarrow$ There is strong evidence for an association between hormonal contraception and the risk of heart attack.

Quantifying the strength of the association

If we consider hormonal contraception vs. no hormonal contraception as our two treatment groups, then π_1 and π_2 represent the relative frequencies (proportions) of women with heart attack in each group:

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

Using these values, there are at least 3 metrics that can be used to quantify how the two groups differ...

Three measures to quantify the association/difference between groups:

- 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 do not differ.
2. $OR > 1 (< 1) \rightarrow$ group 1 $>$ (or $<$) group 2

The odds and the odds ratio

- The **odds** ("Wettverhältnis"): For a probability π the odds is

$$\frac{\pi}{(1-\pi)} = \frac{\text{Wahrscheinlichkeit}}{\text{Gegenwahrscheinlichkeit}} . \quad (1)$$

  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:

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

  1. $\log(OR) = 0 \rightarrow$ the two groups do not differ.
  2. $\log(OR) > 0 (< 0) \rightarrow$ group 1 $>$ (or $<$) group 2

Binomial and binary regression

Often, the situation is more complicated than:

binary outcome (y)$\sim$ binary explanatory variable (x)

Often, we are interested in:

binary outcome (y)$\sim$ continuous/categorical variables $x^{(1)}$, $x^{(2)}$, ...

$\rightarrow$ A (multiple) regression model is needed!

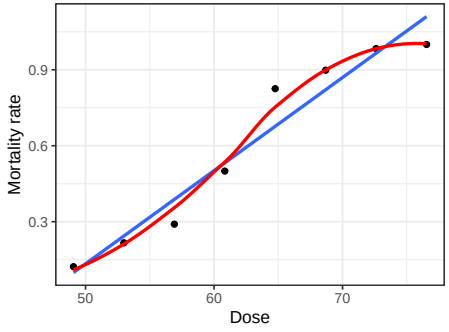
Illustrative/working example

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 whether it was killed or not (1 or 0), but the data were reported in **aggregated** form:

```
##    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
```

As always, start with a graph:



with linear (blue) and smoothed line (red). **Question:** (How) does the dose of the insecticide (x) affect the mortality (y) of the beetles?

What can we see from the plot?

- Mortality increases with higher doses of the herbicide
- The linear line seems unreasonable. In particular, extrapolation to lower or higher doses leads to mortalities < 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.
- Neither are the correct approach here.

The "wrong" analyses

Wrong analysis 1: Linear regression

We could simply use:

$$E(y_i) = \beta_0 + \beta_1 Dose_i$$

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

```
r.lm<- lm(Mortality_rate~ Dose, data= beetle)
```

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

Problems:

- Impossible predicted probabilities
- Residuals ϵ_i are **not** normally distributed

Wrong analysis 2: Poisson regression

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

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

with $E(y_i)$ = number killed.

```
r.pois<- glm(Number_killed~ Dose, data= beetle,family= poisson)
```

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

Problem: This model also makes impossible predictions as, for a dose of 76, it predicts that $E(y_i) = \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 76) = 73.80$ beetles die, which is more than the number of beetles that were tested!

Sidenote: Poisson vs. binomial distribution

Clarification of the difference between the Poisson and binomial distribution:

Poisson is appropriate when:

- There is no theoretical upper limit to the number of times an "event" can occur, or observed values are far from such an upper limit (e.g., number of birds observed in a forest plot)
- Counts cannot be expressed as a proportion.

Binomial is appropriate when:

- Aggregated version of many binary experiments, that is, each can be 0 or 1.
- There is an upper limit to the number of times an "event" can occur (e.g., number of deaths out of a known total number of individuals).
- Events can be expressed as a proportion (number of successes/number of trials).

A model for binary data?

Remember the Bernoulli distribution from Mat183:

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

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

Characteristics of the Bernoulli distribution:

- $E(Y) = \pi = P(Y = 1)$
- $Var(Y) = \pi(1 - \pi)$.

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

From binary to binomial data

Binomial data is an aggregation of binary data:

- Repeat the experiment with $P(Y = 1) = \pi$ 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 the same expectation as the success probability of a single experiment:

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

Example: In the beetle data $n = 49$ beetles were tested for the lowest dose, of which $k = 6$ died, thus the "success rate" is $6/49 = 0.122$.

The binomial distribution

The **binomial distribution** assigns the probability of seeing k successes out of n trials, where the success probability of a single trial is π.

$$P(Y = k) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \ , \quad k = 0, 1, 2, \ldots, n$$

In short:

$$Y \sim Binom(n, \pi) \ .$$

Characteristics of the binomial distribution:

- Mean: $E(Y) = n \cdot \pi$
- Variance: $Var(Y) = n \cdot \pi(1-\pi)$

$\rightarrow$ For given n, the variance is determined by its mean.

R functions: `rbinom()`, `dbinom()`

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 need a link function that relates the linear predictor η_i to the expected value $E(y_i)$.

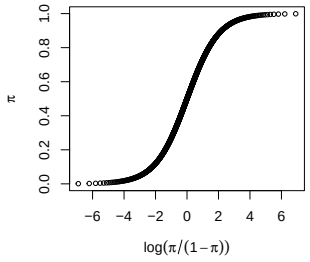
The link function must be chosen such that the expected value $E(y_i)$ is always between 0 and 1.

Link function: The logit transformation

The logit-transformation assigns a probability (π) between 0 and 1, to a value between $-\infty$ and ∞:

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

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



The logistic regression model

In order to prevent the expected value $E(y_i)$ of a binary outcome (1/0) to attain unreasonable values, we thus formulate the logistic regression model as

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

with $\pi_i = P(y_i = 1)$.

- The link function is called the **logit link**.
- To backtransform from the link scale (log odds) to the response scale (probability), the **logistic function** is applied
- The family is **binomial**.

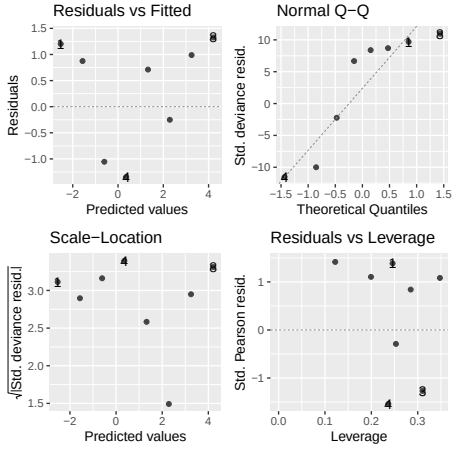
Doing it right: Fitting a logistic regression

- As for the Poisson GLM, we can estimate the parameters $\beta_0, \beta_1, \ldots$ by maximizing the likelihood (ML estimation).
- The `glm()` function in R can also handle binomial and binary data
- For `glm(..., family= binomial)`, the default link function is the logistic link.
- Another novelty lies in the fact that we need to give 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 survived)

```
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 output, assess model diagnostics:



→ Hard to see anything due to low number of data points.

- As in Poisson regression, it is not clear how to define residuals, there are many ways (data scale, linear predictor scale, likelihood scale).
- Again, different types of residuals are used in the plots generated by `autoplot()`.
- Be careful: such plots are only reasonable for aggregated data (which we have here)! The larger the groups, the more precise 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 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; be careful, this is wrong in the Hector book!!).

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

Note: The β coefficients are approximately normally distributed as $N(\hat{\beta}, \hat{\sigma}^2_{\hat{\beta}})$.

$\rightarrow$ confidence intervals etc. can be calculated as in the linear case

Quantitative interpretation of the coefficients

Remember the regression model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Dose_i \ . \tag{2}$$

To understand what β_1 tells us, let's solve for π_i:

$$\pi_i = P(y_i = 1|Dose_i) = \frac{\exp(\beta_0 + \beta_1 Dose_i)}{1 + \exp(\beta_0 + \beta_1 Dose_i)} \ . \tag{3}$$

From model (2) it is possible to calculate the odds ("Chance"):

$$odds(y_i = 1|Dose_i) = \frac{\pi_i}{1-\pi_i} = \frac{P(y_i = 1|Dose_i)}{P(y_i = 0|Dose_i)} = \exp(\beta_0 + \beta_1 Dose_i) \ .$$

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

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

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

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

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

Doing it right: The anova() table

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

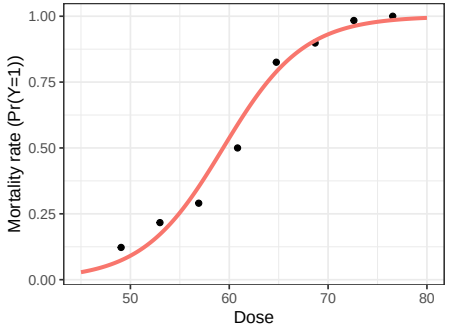
```
anova(beetle.glm, test= "Chisq")
```

```
## 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 reallly small.

Plotting the fit

A fitted curve can be added to the raw data by plotting $P(y_i = 1)$ against the Dose, using equation (3):



(Compare to Figure 9.1 in the Hector book *The new statistics with R*.)

Overdispersion

Remember :

- Slides 18 and 20: $E(Y) = \pi$ and $Var(Y) = \pi(1 - \pi)$, thus the variance is determined by the mean
- "Overdispersion" means "extra variability" (larger than the model predicts or allows).
- Probable reason: Missing variables in the model
- Overdispersion leads to *p*-values that are too small.
- Can be detected by looking at the residual deviance:

  `Residual deviance` $>>$ `df` $\rightarrow$ 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?

```
pchisq(8.438, 6, lower.tail=F)
```

```
## [1] 0.2077375
```

$\rightarrow$ $p = 0.21$ does not seem problematic.

One can nevertheless account for overdispersion by switching to a 'quasibinomial' model, which estimates the dispersion parameter separately.

```r
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 dosis, we had several beetles. For instance, 49 beetles at lowest dose (49.06), of which 6 died (1) and 43 survived (0). This was binomial data, an aggregated version of many (here 49) trials with 0 or 1 outcome.
- In reality, one often has only one trial (0/1) for a (combination of) explanatory variable(s).
- The analysis is the same as for aggregated data, however there are a few complications with graphical descriptions and model checking.

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

Blood screening example

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

Introduction: Individuals with a low ESR (Erythrocyte Sedimentation Rate) are generally considered healthy, while those with an ESR $> 20mm/hr$ are not. We are interested whether concentrations of the plasma proteins Fibrinogen and Globulin (which can be disease indicators) are associated with an increased probability that an individual has a high ESR (ESR$> 20mm/hr$, encoded as $y_i = 1$)?

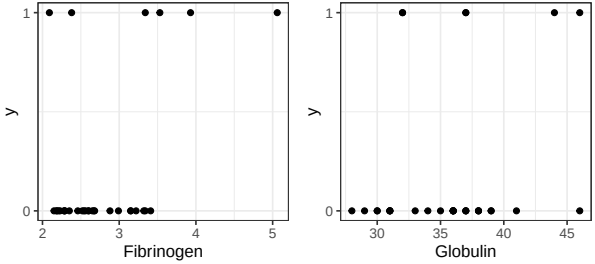
The model to be fitted:

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

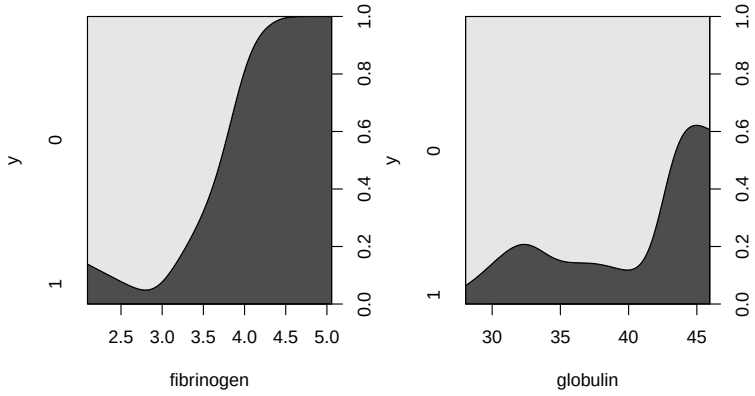
with $E(y_i) = P(y_i = 1) = \pi_i$. \ Equivalently: $y_i \sim \mathit{Bern}(\pi_i)$

Complication 1 with binary data: Graphical description

Plotting the response y ($y = 1$ if ESR$>$ 20 and $y = 0$ if ESR$<$ 20) against the covariates does not lead to very informative graphs:



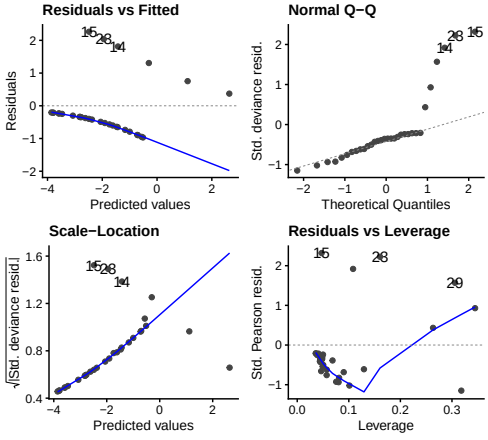
It is a bit more informative to look at a conditional density plot (`cdplot()`):



Complication 2: Model diagnostics

a) Residual plots:

Plotting the residuals is possible, but not meaningful. Why? Because 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**!!

Why? Becasue for a single binary (0/1) variable it is impossible to estimate a variance, thus it is also impossible to say if the variance is too high/too low.

Your turn!

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

```r
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
```

```r
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

- Logistic regression is also known as a binary/binomial GLM
- The link function is the logistic link.
- The coefficients of logistic regression are log odds ratios

  $\Leftrightarrow$ $\exp(\beta)$ is an odds ratio
- You can think of binomial data as aggregated binary data
- Overdispersion does not apply to binary data