Kurs Bio144: Datenanalyse in der Biologie

Lecture 10: Modeling count data

Stefanie Muff (Lecture) & Owen L.Petchey (Practical)

University of Zurich

26 December, 2020

Overview

- When the outcome (y) is a count
- Generalized linear models
- Poisson regression
- Link function
- Residual analysis / model checking / deviances
- Interpretation of the results
- Overdispersion, zero-inflation

Course material covered today

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

- Chapter 7 of GSWR (Beckerman et al.)

Introduction

- We have seen: Covariates in regression models can be continuous, categorical (binary or multi-level) or counted numbers (integers).
- However, the response variable (y) was so far always continuous, and the assumption was that the residuals $\epsilon_i \sim N(0, \sigma^2)$.
- In reality, the response variable can also be a count (an integer ≥ 0) or a categorical variable (e.g., presence/absence data).
- Today we will look at the case where the response variable is a count, that is, $y_i = 0, 1, 2, \ldots$.

Count data

In biological or medical data, the outcome of interest is quite often a count:

- Counting items in time or space (animals, plants, species)
- Parasites in animals or humans
- Number of offspring in animals or humans
- Number of adverse health events (e.g., exacerbations) or health-related numbers (e.g., polyps)

The research question then is:

How do the covariates influence the probability of a certain count outcome?

Illustrative/working example: Soay sheep

Hirta, a small island of Scotland, is inhabited by an unmanaged and feral population of Soay sheep.

Ecologists were interested in the **question** whether body mass of female animals influences their fitness, measured as their lifetime reproductive success (the number of offspring over their lifespan).

Question: "Are heavier females fitter than lighter females?"

```
## [1] "missing DATA"
```

```
#glimpse(soay)
```

As always, start with a graph (see GSWR book for code to produce the figure):

What can we see from the plot?

- Reproductive success seems to increase with female body weight (not surprising, right?).
- The linear line seems unreasonable, the red (smoothed line) seems better. Maybe a quadratic term would be useful?
- However, the problem with these data is more subtle.

How does one analyze these data correctly?

- So far, we always used linear regression.
- This is not the correct approach here.
- We nevertheless start doing the 'wrong' analysis!

The “wrong” analysis

Use the `lm()` function to fit the linear model $y = \beta_0 + \beta_1 \cdot \text{bodySize} + \epsilon$ and look at the diagnostic plots:

$\rightarrow$ The diagnostic plots indicate that the **linear regression assumptions are violated**!

The same plot with smoothed lines (makes the problems more obvious):

What about the model with a quadratic term?

$$y = \beta_0 + \beta_1 \cdot \text{bodySize} + \beta_2 \cdot \text{bodySize}^2 + \epsilon$$

$\rightarrow$ This looks a bit better. However...

What is the problem?

There is still a clear upward trend in the scale-location plot, indicating that the variance increases as the fitted values grow larger (I used the smoother to make this obvious).

Moreover, the normal distribution for count data is obviously violated. Why?

- The normal distribution is for continuous variables.
- The normal distribution allows values < 0, count data doesn't.
- The normal distribution is symmetrical, counts are often not! In particular, they cannot be negative.

A model for count data? I

Do you remember a distribution for count values from Mat183?

The probability distribution[a] of Poisson-distributed random variable Y with parameter λ is defined as

$$P(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda} , \quad k = 0, 1, 2, \dots .$$

In short:

$$Y \sim Po(\lambda) .$$

[a] A probability distribution is just a mathematical statement of how likely different events are, see GSWR book p. 169.

A model for count data? II

Characteristics of the Poisson distribution:

- Suitable to model unbounded counts ($k = 0, 1, 2, ...$).
- $E(Y) = Var(Y) = \lambda$. In words:

  Mean = variance = λ.

  $\rightarrow$ The variance of the distribution increases with the mean.

A model for count data? III

Some examples:

MISSING IMAGE

So: How can one use the Poisson distribution for regression modeling?

Doing it right: The generalized linear model (GLM) for count data

The aim of the GLM approach is that we can still use a linear predictor η_i in the form of the linear model:

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

- In **linear regression**, η_i is the predicted value for the mean $E(y_i) = \eta_i$.
  (Why $E(y_i)$ and not y_i? Because the residual/error term $+\epsilon_i$ is missing!)
- However, for counts we cannot simply set η_i equal to $E(y_i)$, if y_i is a count!

Why can't we set $E(y_i) = \eta_i$ for count data?

$\rightarrow$ **Because nothing prevents η_i from being negative!**\[4mm]

Let us try to use the same approach as in linear regression, assuming that $E(y_i) = \eta_i$ for our soay sheep counts, that is,

$$E(y_i) = \beta_0 + \beta_1 \cdot \text{bodySize}_i \ ,$$

using a model with $\beta_0 = -2$ and $\beta_1 = 1.2$.

What is the predicted number of offspring for a 1kg sheep? Plug-in:
$-2 + 1.2 \cdot 1 = -0.8$, thus a negative prediction!

This is very unreasonable, right?

$\rightarrow$ We need a trick!

The trick: Use a link function

Instead of using $E(y_i) = \eta_i$ as in linear regression, we simply log-transform the expected value.

In order to prevent the expected value $E(y_i)$ to be negative for count data y_i, we use

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

- The log is called the link function.
- The **advantage**: The predicted fitness $E(y_i)$ is now always positive, because equation (1) is identical to

$$E(y_i) = \exp(\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_p x_i^{(p)}) \ ,$$

which is now **always** > 0! (Plot the exp() function if you forgot what it looks like...)

The probability model

- Finally, we need a reasonable probability model for the response variable.
- Remember: we always used the normality assumption $y_i \sim N(\eta_i, \sigma^2)$ in linear regression.
- Given that y_i are counts, a Poisson model seems more reasonable:

$$y_i \sim Po(E(y_i)) \ .$$

  In words: y_i is a realization of a Poisson random variable distributed as $Po(\lambda_i)$, where $\lambda_i = E(y_i)$.
- We say: The model belongs to the Poisson family.

Key terms for GLMs

In summary, we have introduced three terms related to GLMs:

- **Family**: The family corresponds to the likelihood model that is used for the (transformed) response. Here, the family is Poisson. Another very common distribution is the binomial family for binary outcome (see next week). Other families are the gamma or the negative binomial. The family is determined by the data type!
- **Linear predictor**: The linear predictor is always given as

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

  In the soay sheep example, it is $\eta_i = \beta_0 + \beta_1 \cdot \text{bodySize}_i$ for sheep i.
- **Link function**: It defines how the linear predictor η_i is related to $E(y_i)$. In Poisson regression this is typically the log: $\log(E(y_i)) = \eta_i$! We will see another important link function next week (for binary data).

Doing it right: Fitting a Poisson GLM

Uff... that was hard! But we finally have all the tools for fitting a Poisson GLM. A statistician would now say:

"Let's fit a Poisson GLM with a log-link!"

- Basically, the idea is again to perform maximum-likelihood estimation, but the details are a bit more complicated (nothing for us).
- Luckily, there is an R-function that works (almost) like 'lm()', namely `glm()` :

```
soay.glm <- glm(fitness   body.size, data = soay, family =
poisson(link=log))
```

You **must** specify the family, but you could leave away the `link=log` option here, because R automatically picks the log link for the Poisson family...

Doing it right: Model diagnostics

Before we look at the output, let's do some model diagnostics (as we did it for linear regression):

- Model diagnostics seem ok. In particular, the scale location plot is a bit better than when using the quadratic term in linear regression (slide 11).
- Do you find it strange that the same diagnostic plot can be used as in simple linear regression?
- The definition of a "residuum" is no longer clear when link-functions are used (on which scale should residuals be calculated? On the link scale or on the observed scale?)
- We don't care too much about this aspect, but you should remember:
  - There are different types of residuals.
  - `autoplot()` automatically picks the residuals that "make sense".

Doing it right: Interpreting the coefficients

Let's now look at the `summary()` output (yes! this works also for glm objects):

```
#summary(soay.glm)
```

You see several (familiar and less familiar) components in the output. For the moment, we are interested in the coefficients, which are estimated as

with respective standard errors and p-values. In particular, $p < 0.001$ for $\hat{\beta}_1$ indicates very strong evidence for a positive (because $\hat{\beta}_1 > 0$!) effect of female weight on reproductive success (number of offspring)!

Good to know: Theory says that the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately normally distributed around the true values:

$$\hat{\beta} \sim N(\beta, \sigma^2_{\hat{\beta}}) \tag{2}$$

Thus a 95% CI can be approximated by the usual $\hat{\beta} \pm 2 \cdot \hat{\sigma}_{\hat{\beta}}$ idea.

What do the coefficients tell us?

Remember our model

$$\log(E(y_i)) = \beta_0 + \beta_1 \cdot \text{bodySize}_i$$

Thus the $\exp(\cdot)$ transformation leads to

$$E(y_i) = \exp(\beta_0 + \beta_1 \cdot \text{bodySize}_i) \ .$$

Our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be plugged into this equation!

For example, a 5~kg female has **expected fitness** (i.e., produces)

while a 7~kg female would already have **expected fitness**

Makes sense, right?

Doing it right: The anova() table

Remember: We were sometimes using ANOVA tables, for instance for linear regression output with categorical covariates.

anova() is useful again, but it gives us the Analysis of Deviance table:

```
#anova(soay.glm)
# MISSING DATA
```

Note: The deviance is essentially a difference of likelihoods. In brief, the larger the likelihood, the better fits your model to the data...

Here, the total deviance is given by the so-called NULL deviance: 85.081. It is the analogon to the total variability of the data in linear regression...

Of this, 37.041 is explained by bodysize.

The question is, whether this is "much''? This can be tested by a χ^2 test, because the value is χ^2 distributed with 1 degree of freedom:

```
pchisq(37.041,1,lower.tail=F)
```

```
## [1] 1.156712e-09
```

You would get the same if you directly specify the test you want to carry out in the `anova()` call:

```
#anova(soay.glm, test="Chisq")
```

Making a beautiful graph

See section 7.4.5 in the GSWR book. Work through the code to obtain this one:

Your turn!

Let's look at another data example, taken from Hothorn and Everitt (2014), chapter 7:

A new drug was tested in a clinical trial (Giardiello et al. 1993, Piantodsi 1997), aiming at reducing the number of polyps in the colon (Dickdarm). The data are publicly available from the Hothorn/Everitt book package:

```
library(HSAUR3)
data("polyps")
```

Scientific question: Does the drug influence (reduce) the number of polyps?

Data: Number of polyps (outcome), the binary variable for the treatment, and the continuous covariate age.

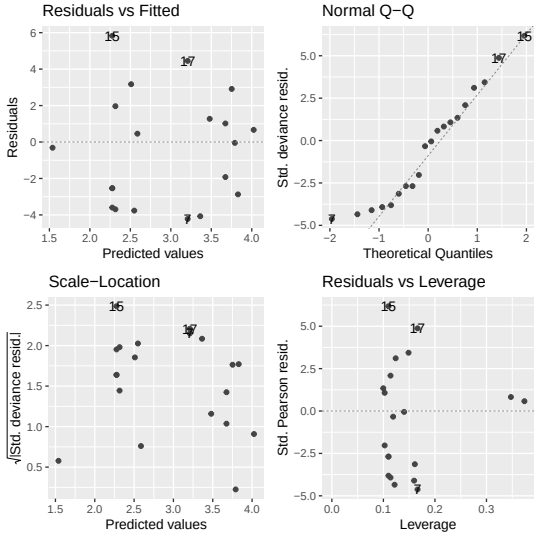
number	treat	age
63	placebo	20
2	drug	16
28	placebo	18
17	drug	22
61	placebo	13
1	drug	23
7	placebo	34
15	placebo	50
44	placebo	19
25	drug	17

number	treat	age
3	drug	23
28	placebo	22
10	placebo	30
40	placebo	27
33	drug	23
46	placebo	22
50	placebo	34
3	drug	23
1	drug	22
4	drug	42

Your tasks (in teams, if you like):

Look at the analysis done on the next three slides, and answer the following questions:

1. Are there any problems visible from the diagnostics plots?
2. Does the treatment seem to be effective? If yes, can you quantify the effect?
3. Is age a relevant variable? If yes, what happens when patients grow older?


Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
Predicted values
Residuals vs Leverage
Std. Pearson resid.
Leverage

```r
anova(polyps.glm,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: number
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     19     378.66
## treat  1  150.101        18     228.56 < 2.2e-16 ***
## age    1   49.018        17     179.54 2.536e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(polyps.glm)
```

```
##
## Call:
## glm(formula = number ~ treat + age, family = "poisson", data = polyps)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2212  -3.0536  -0.1802   1.4459   5.8301
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.529024   0.146872   30.84  < 2e-16 ***
## treatdrug   -1.359083   0.117643  -11.55  < 2e-16 ***
## age         -0.038830   0.005955   -6.52 7.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: 273.88
##
## Number of Fisher Scoring iterations: 5
```

Overdispersion

"Overdispersion" means **"extra variability"**. Why could this be a problem?

Remember: The variance of the Poisson distribution increases with the mean, because **mean=variance** (see slide 14).

In Poisson regression it is assumed that, for each observation i,

$$y_i \sim Po(E(y_i)) \ .$$

However, the variance is often larger than the mean in reality, because there are factors that influence the response ("cause variability in it") that cannot be captured by the covariates!

Why? Maybe you simply cannot observe the variable or it is too expensive to monitor, or...

Detecting overdispersion

Look at the summary output from your GLM object, check the "Residual deviance" and compare it to the "degrees of freedom".

Soay sheep data: Res. deviance: 48.040, df=48

Polyps data: Res. deviance: 178.54, df=17

The residual deviance should be approximately χ^2 distributed with df degrees of freedom. This means that one should check whether

$$\texttt{Residual deviance} \approx \texttt{df} \ .$$

The sheep data seem fine, but the polyps data have

$\texttt{Residual deviance} >> \texttt{df} \quad \rightarrow$ overdispersion!

If unclear, use the χ^2 test with df degrees of freedom.

What is the problem with overdispersion?

When there is unaccounted overdispersion, the p-values that are calculated are usually too small!

Possible solutions:

- Use as your regression family `quasipoisson` instead of `family = poisson`.

  This allows to estimate the variance parameter (denoted as dispersion parameter) separately.

- Use a negative binomial regression (there is the `glm.nb()` fuction from the `MASS` package in R to fit it, check it out by `?glm.nb()`).

Reanalyzing the polyps data with quasipoisson

```
polyps.glm2 <- glm(number ~ treat + age,data=polyps, family="quasipoisson")
summary(polyps.glm2)
```

```
##
## Call:
## glm(formula = number ~ treat + age, family = "quasipoisson",
##     data = polyps)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2212  -3.0536  -0.1802   1.4459   5.8301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.52902    0.48106   9.415 3.72e-08 ***
## treatdrug   -1.35908    0.38533  -3.527  0.00259 **
## age         -0.03883    0.01951  -1.991  0.06284 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.72805)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

- The dispersion parameter is now estiamted as 10.73, which is calculated as

  `Residual deviance / df`.

- The p-values for the coefficients are now larger! In particular, there is only weak evidence ($p = 0.063$) for an effect of age!

  $\rightarrow$ The `poisson` family gave too optimistic p-values!

  We say: The p-values were anti-conservative or non-conservative.

(Anti-conservative results are really **the worst** that can happen. Guess why?)

Underdispersion?

Can it happen that the observations are less variable than expected?

Yes: Especially when there are dependent observations!

You can **detect it** by checking if ; `Residual deviance` $<$ `df`.

In that case, your p-values are usually too large, that is, the results are conservative (just the opposite of the overdispersion problem).

The `quasipoisson` regression is a pragmatic solution in that case, too.

Zero-inflation

A special type of overdispersion may be caused by an overrepresentation of zeros in the obervations.

Example: Numbers of cigarettes smoked

Some people are never-smokers, so they will always produce a zero obervation, while smokers may smoke any number of cigarettes.

Please read chapter 7.5.2 of GSWR for some ideas how to handle this case.

A note on interpretation and model selection

Just as a reminder:

The same remarks and warnings from the last weeks for linear models regarding

- **Caution with model selection**
- **Interpretation of p-values**
- **Reproducibility aspects**

also apply to GLMs!

Summary

- Poisson regression is useful to model counted outcomes.
- Pretending that counted outcomes are continuous may lead to wrong results.
- The main ingredients of GLMs are
  - The family
  - The linear predictor
  - The link function.
- Model diagnostics are similar as in the linear case (`autoplot()`).
- Interpret the coefficients by back-transforming to the original scale.
- Analysis of deviance is the analogon to ANOVA (`anova()`).
- Over-(under-)dispersion, how to detect it and what do do.

References

Hothorn, T. and B.S. Everitt (2014). *A Handbook of Statistical Analyses Using R* (3 ed.). Boca Raton:Chapman & Hall/CRC Press.