Lecture 10: Modeling count data

BIO144 Data Analysis in Biology

Stephanie Muff, Owen Petchey, Erik Willems

University of Zurich

06 May, 2024

Recap

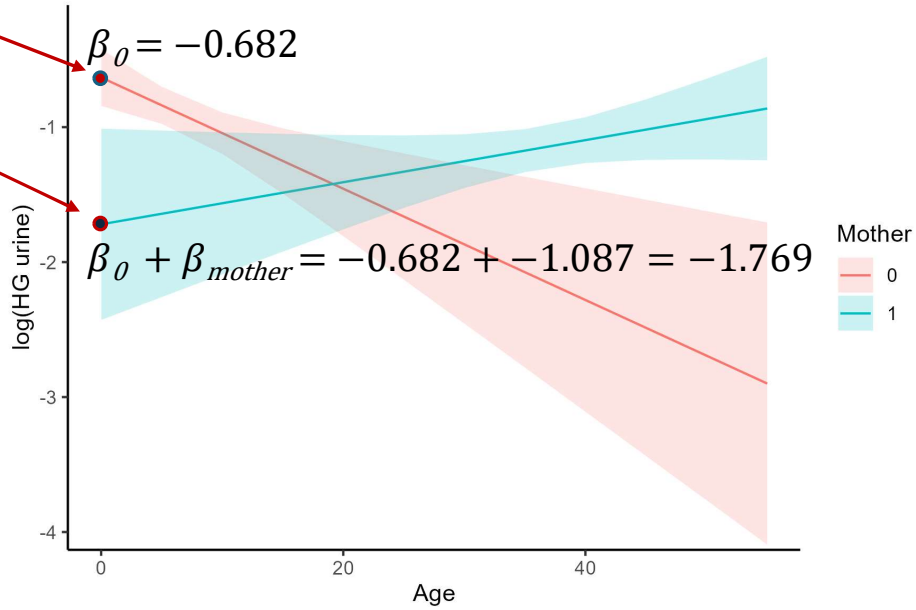
In linear models, the p-value is often used as an indicator of explanatory model importance. Remember the mercury example:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
~~age~~	~~0.041~~	~~0.013~~	~~3.289~~	~~0.001~~
~~mother~~	~~-1.087~~	~~0.373~~	~~-2.912~~	~~0.004~~
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000

A common practice is to look only at the p-value and use $p < 0.05$ to decide whether a variable has an influence or not ("is significant or not").

Recap

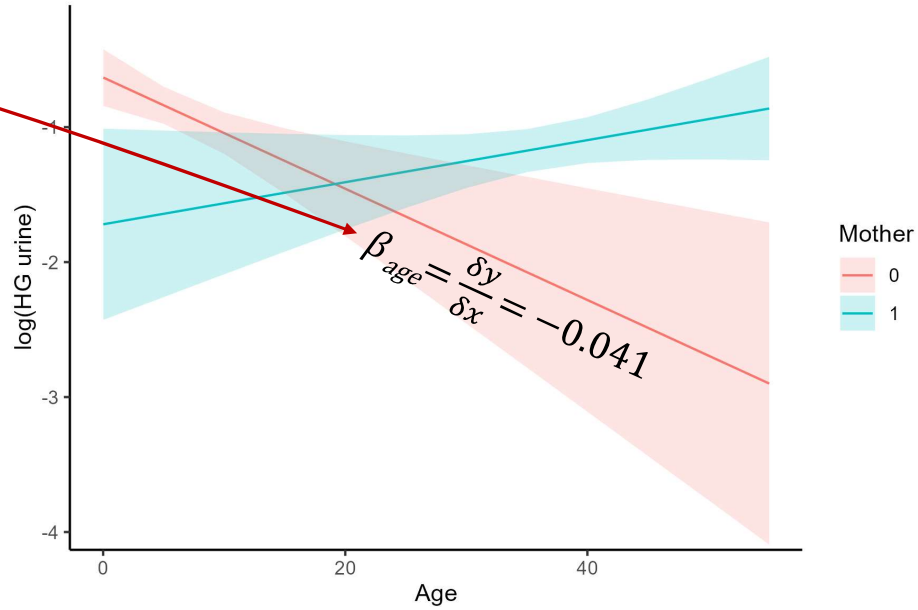
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
age	-0.041	0.013	-3.289	0.001
mother	-1.087	0.373	-2.912	0.004
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000



Categorical terms represent a difference in intercept

Recap

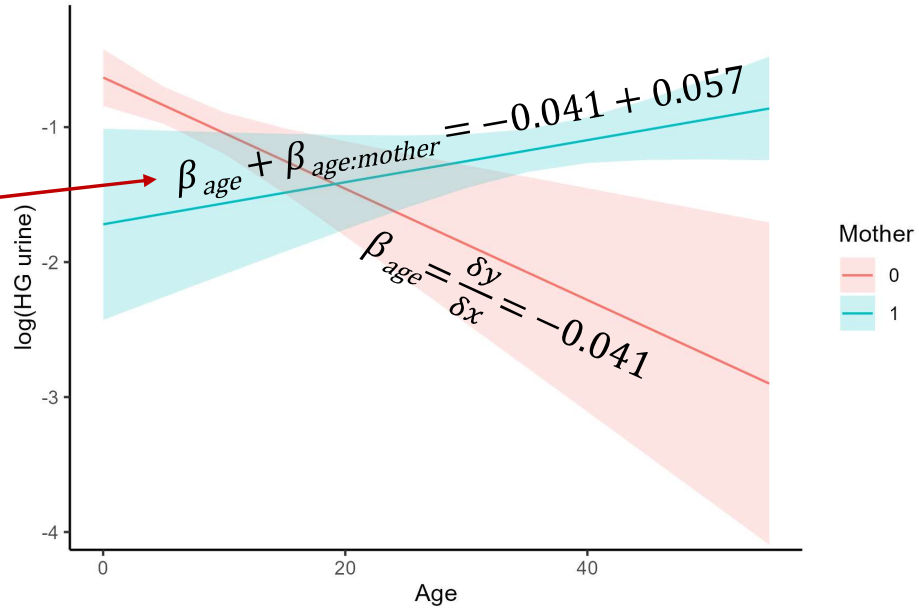
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
age	-0.041	0.013	-3.289	0.001
mother	-1.087	0.373	-2.912	0.004
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000



Continuous terms represent a slope

Recap

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.682	0.113	-6.057	0.000
log10(Hg_soil)	0.030	0.043	0.691	0.491
vegetables	0.058	0.057	1.019	0.310
migration	-0.013	0.087	-0.150	0.881
smoking	0.351	0.116	3.014	0.003
sqrt(amalgam)	0.290	0.052	5.616	0.000
age	-0.041	0.013	-3.289	0.001
mother	-1.087	0.373	-2.912	0.004
sqrt(fish)	0.070	0.030	2.352	0.020
last_fish	0.311	0.081	3.844	0.000
age:mother	0.057	0.016	3.627	0.000



An interaction between a categorical and continuous term represents a difference in slope

Overview

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

outcome variable does not follow a normal distribution

Course material covered today

Today's lecture is largely based on the following:

- Chapter 7 of GSWR (Beckerman et al.)
- Hothorn T and Everitt BS (2014), A Handbook of Statistical Analyses Using R

Introduction

- We have seen that explanatory variables in regression models can be continuous, categorical (binary or multi-level), or count (non-negative integers).
- The response variable (y), so far, has always been continuous, with the assumption that residuals $\epsilon_i \sim N(0, \sigma^2)$.
- Often, however, the response variable will be a count, or categorical variable.
- Today we will look at the case when 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)
  - Number of offspring in animals
  - Number of pathological structures in humans (e.g. polyps)

In such cases, the research question is:

How do the explanatory variables influence the probability of a given count of the outcome?

Example: Soay sheep

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

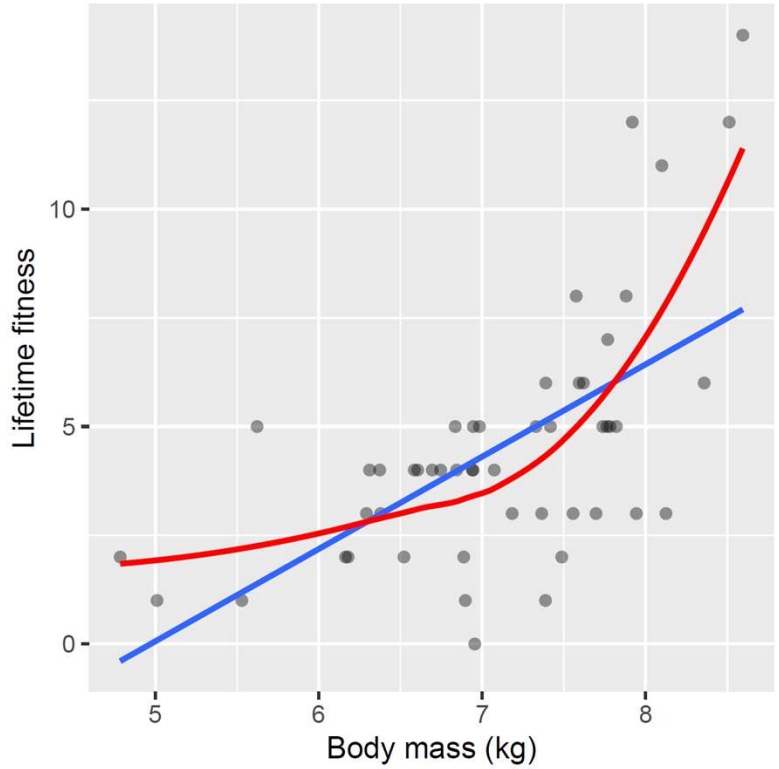
Ecologists were interested whether body mass of females influences their fitness, measured as their lifetime reproductive success (the number of offspring over their lifespan).

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

```
glimpse(soay)
```

```
## Rows: 50
## Columns: 2
## $ fitness   <int> 4, 3, 2, 14, 5, 2, 2, 5, 8, 4, 12, 6, 3, 2, 3, 0, 5, 3, 5, 6
## $ body.size <dbl> 6.373546, 7.183643, 6.164371, 8.595281, 7.329508, 6.179532,
```

As always, explore the data with a graph (code in GSWR book):



What can we see from the figure?

- Reproductive success indeed seems to increase with female body weight.
- The straight line does not capture the pattern very well, the red (smoothed) line seems better. Maybe a quadratic term would improve things?
- However, the "problem" with these data is more subtle.

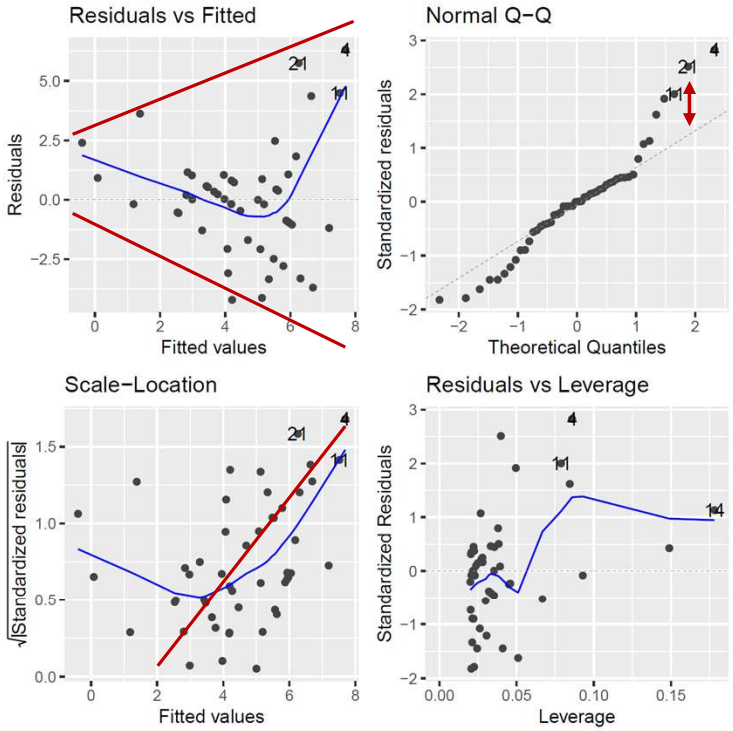
How does one analyze these data correctly?

- The outcome represents a count (non-negative integer)!
- So far, we have always used normal linear regression.
- This is not the correct approach here, but let's do it anyway.

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:

heteroscedasticity



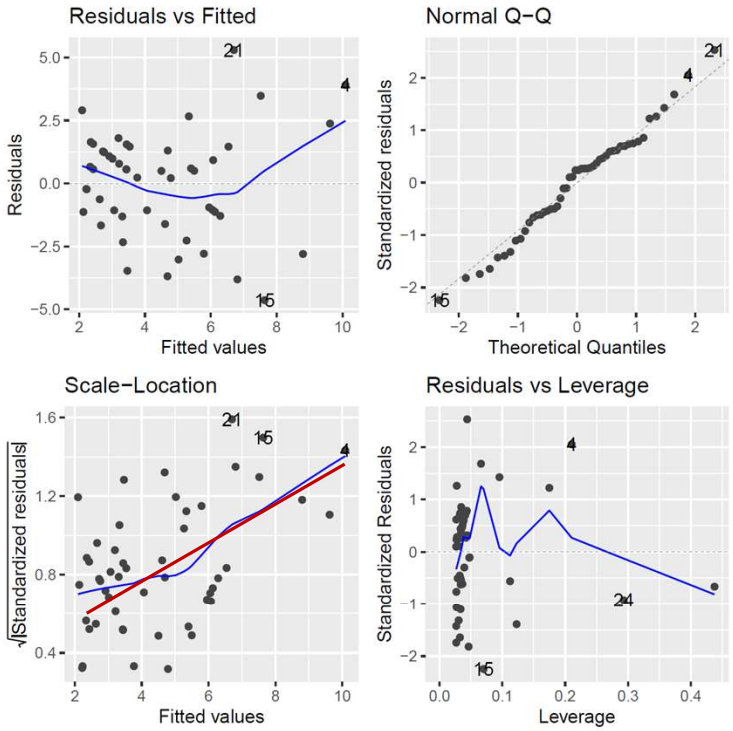
higher values at right tail of distribution

variability increases with fitted values

→ The diagnostic plots indicate that the **linear regression assumptions are violated**!

What about the model with a quadratic term?

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



variability increases with fitted values

→ 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 get larger.

Moreover:

- The normal distribution is for continuous variables.
- The normal distribution allows values < 0, count data are non-negative integers.
- The normal distribution is symmetrical, counts often are not!
- The variability in count data tends to increases with higher values.

A different distribution for count data I

Remember from Mat183?

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

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

In short:

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

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

A different distribution 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.

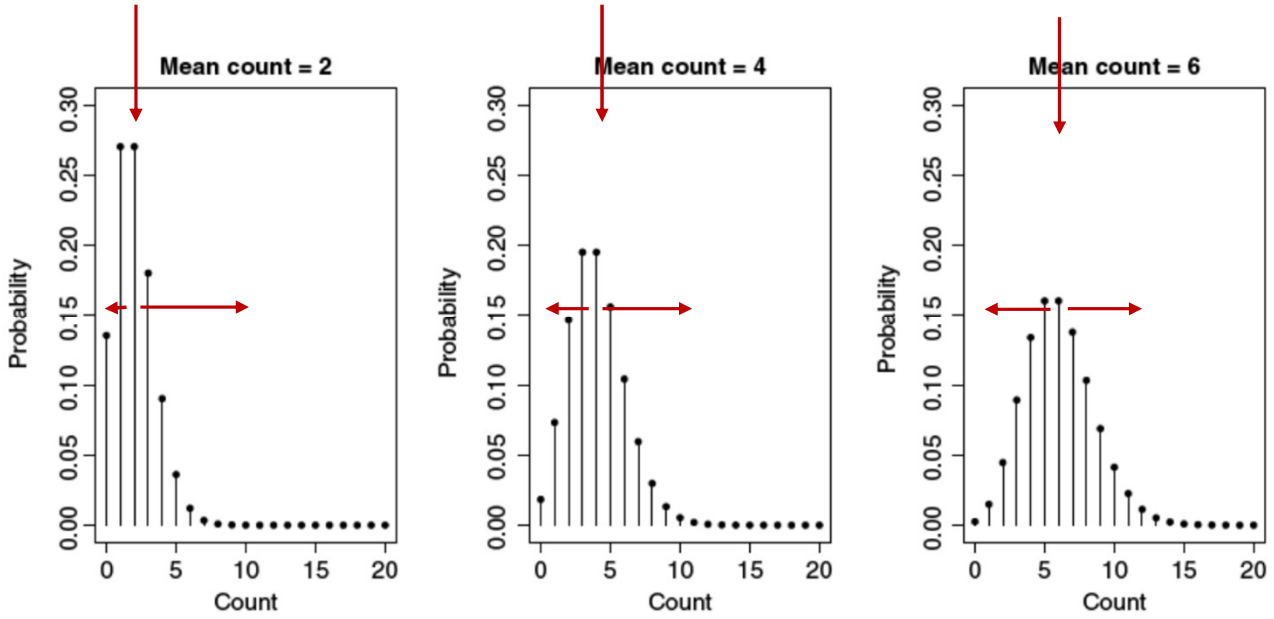
  $\rightarrow$ The variance of the distribution increases with the mean.

Thus:
- The Poisson distribution *Po(λ)* is described by one parameter: the rate

- The normal distribution *N(μ, σ)* is described by two parameters: the mean, and the standard deviation

A different distribution for count data III

Some examples:



So: How can one use the Poisson distribution in a regression model?

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, 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!**

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

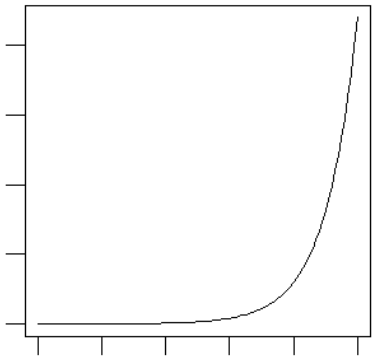
Instead of using $E(y_i) = \eta_i$ as in linear regression, we log-transform the expected value.

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

- 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 is more appropriate:

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

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

Key terms for GLMs

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

- **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.

- **Family**: The family corresponds to the likelihood model that is used for the (transformed) response. For count data, 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!

- **Link function**: This 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 using the logarithmic link-function"

- Basically, the idea is to perform maximum-likelihood estimation(we won't go into the details of this).

equivalent to OLS when outcome followed a normal distribution → another way to calculate the βs to best describe the observed data

- Luckily, there is an R-function that works (almost) like 'lm()', namely `glm()`:

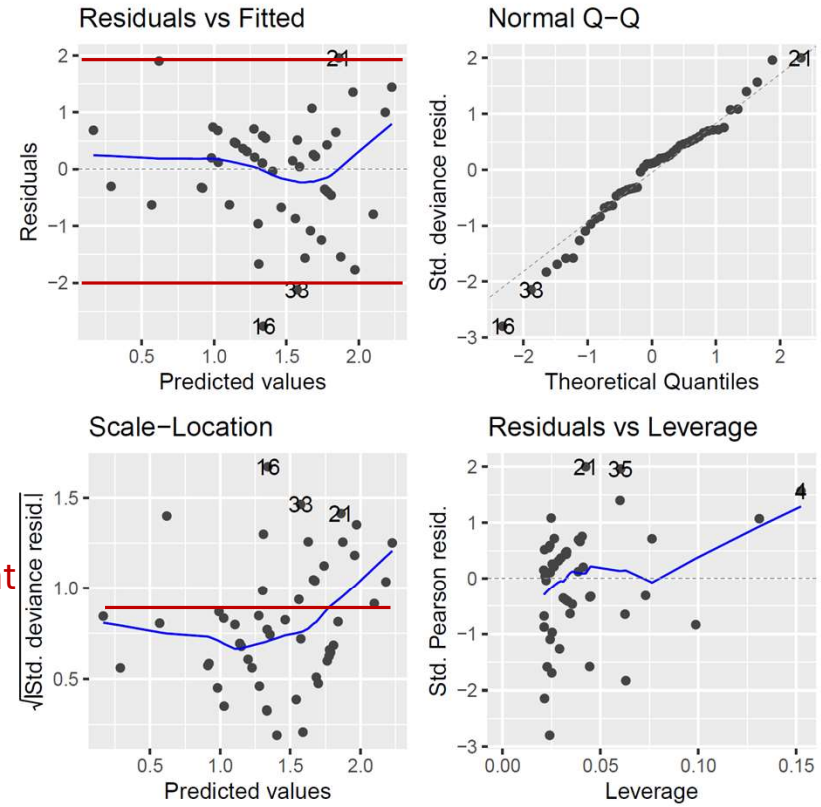
```
soay.glm<- glm(fitness~ body.size, data= soay, family= poisson(link= log))
```

- You **must** specify the family, but you can leave away the `link=log` argument (because R automatically picks the most appropriate link function).

Doing it right: Model diagnostics

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

homoscedasticity



Deviance residuals look normal

variability constant with fitted values

- 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).
- Note that the definition of a "residual" 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?)
- For now, we don't care too much about this, but you should remember:
  - There are different types of residuals.
  - `autoplot()` automatically picks the residuals that "make most sense".

Doing it right: Interpreting the coefficients

Let's now look at the `summary()` output:

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

```
##
## Call:
## glm(formula = fitness ~ body.size, family = poisson(link = log),
##     data = soay)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7634  -0.6275   0.1142   0.5370   1.9578
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.42203    0.69432  -3.488 0.000486 ***
## body.size    0.54087    0.09316   5.806 6.41e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 85.081  on 49  degrees of freedom
## Residual deviance: 48.040  on 48  degrees of freedom
## AIC: 210.85
##
## Number of Fisher Scoring iterations: 4
```

*β*s are on the link-scale →

1kg increase in weight is associated with a:
exp(0.54)= 1.72 times increase in fitness

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

$$\hat{\beta}_0 = -2.422 \quad \text{and} \quad \hat{\beta}_1 = 0.541$$

with respective standard errors and p-values. In particular, $p < 0.001$ for $\hat{\beta}_1$ indicates very strong evidence for a positive 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:

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

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

What do the coefficients tell us?

Remember our model:

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

Which we can back-transform to obtain expected counts, by:

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

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

For example, a 5~kg female has **expected fitness**

$$\exp(-2.422 + 0.541 \cdot 5) = 1.33 \text{ lambs} \ ,$$

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

$$\exp(-2.422 + 0.541 \cdot 7) = 3.91 \text{ lambs} \ .$$

Doing it right: The anova() table

anova() gives us the Analysis of Deviance table:

```
anova(soay.glm)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: fitness
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                         49     85.081
## body.size  1   37.041         48     48.040
```

Maximum Likelihood equivalent to the Analysis of Variance Table

family

"SS_{total}"

"$SS_{residual}$"

"SS_{model}"

Note: The deviance is essentially a difference of likelihoods. Think of it as the "Maximum-Likelihood Estimation" (MLE) equivalent to the Sums of Squares in OLS regression.

Here, the total deviance is the so-called NULL deviance: 85.081. It is analogue 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 (deviance follows a χ^2- rather than an F-distribution:

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

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

Or, directly specify the test in the anova() call:

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

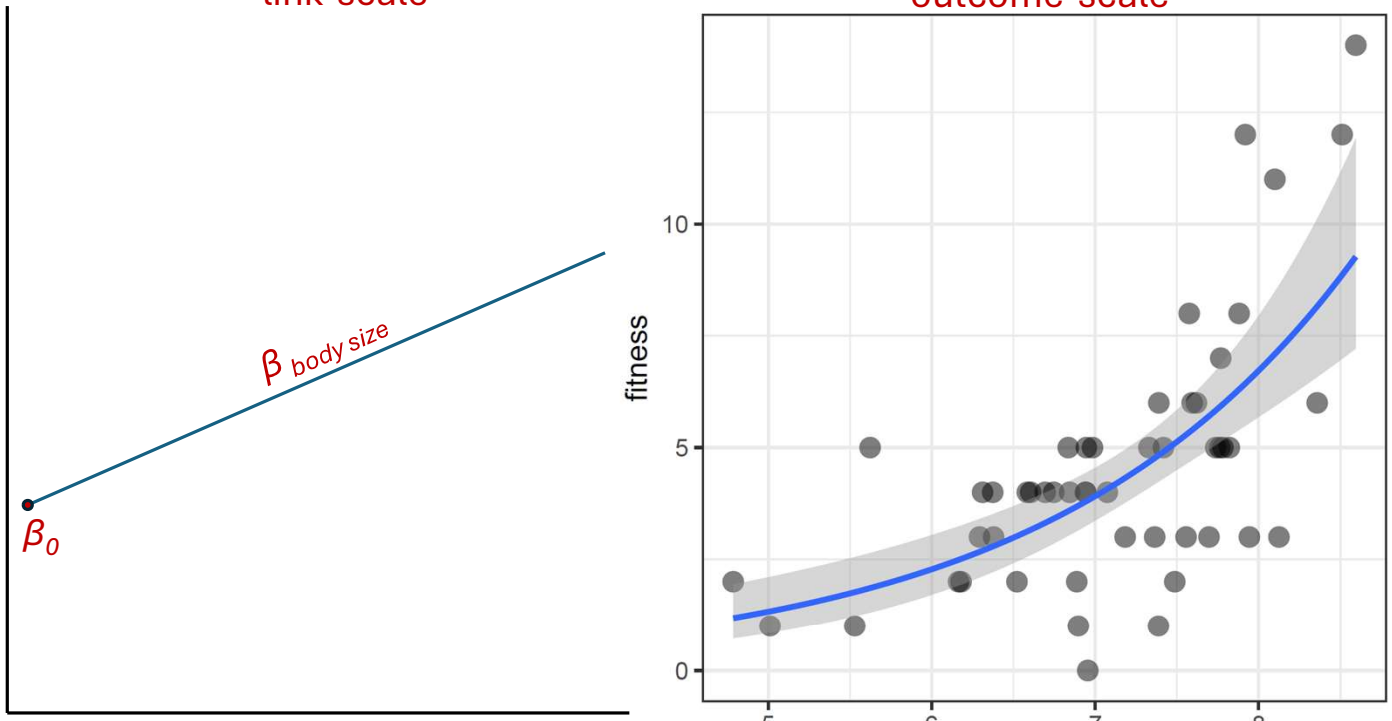
Making a nice graph

exp-transform

See section 7.4.5 in the GSWR book.

link-scale

outcome-scale



log-transform

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 to reduce the number of polyps in the colon (Dickdarm). The data are publicly available from the Hothorn/Everitt book package:

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

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

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

number (outcome)	treat (predictor)	age (confound)
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
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

Like an ANCOVA: number~ treat + age

Your task (in teams, if you like):

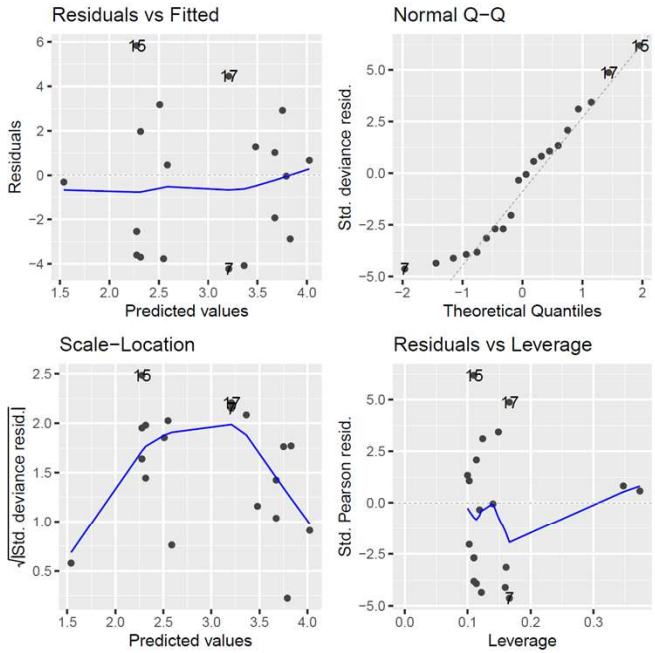
Look at the analysis on the next three slides, and answer the following questions:

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 in older patients?

Fit the model:

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

Look at diagnostics:



Perform an Analysis of Deviance:

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

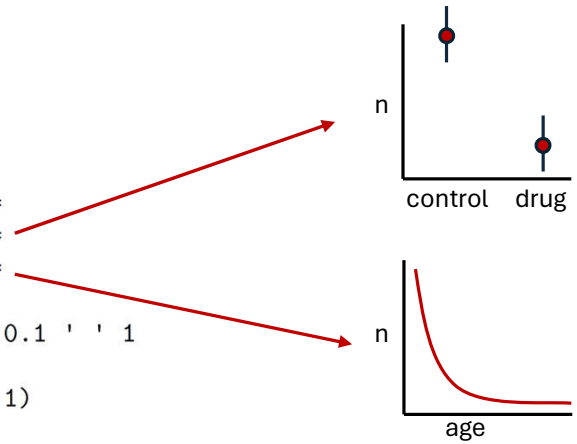
```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: number
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     19     378.66
## treat  1  150.101        18     228.56 < 2.2e-16 ***
## age  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

Inspect the summary table:

```
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)

Actually: λ = mean = $\phi \cdot$ variance, and Poisson assumes $\phi = 1$

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 that cannot be captured by the explanatory variables.

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

reality is more "messy" than mathematics!

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 48 / 48 ≈ 1

Polyps data: Res. deviance: 178.54, df= 17 179/ 17 ≈ 10

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

res. deviance / df ≈ 1

`Residual deviance` $\approx$ `df` .

The sheep model seems fine, but the polyps model does not:

`Residual deviance>> df`

$\rightarrow$ overdispersion

What is the problem with overdispersion?

When there is unaccounted overdispersion, reported p-values are too small!

Possible solutions:

λ = mean = $\phi \cdot$ variance

assumes $\phi = 1$

- Use the quasipoisson-, instead of poisson-family.

This estimates the variance parameter (denoted as dispersion parameter) from the data.

- Use a negative binomial regression (the `glm.nb()` function in the `MASS` package)

Reanalyzing the polyps data with a quasipoisson family

```r
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
```

same

different → multiplied by $\sqrt{10.72805}$

- The dispersion parameter is estimated as 10.73, which is close to that calculated by

  `Residual deviance / df.`

  (See the relevant FAQ on OLAT for further details.)

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

  $\rightarrow$ The `poisson` model gave p-values that were too optimistic

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

(Anti-conservative results are **a big problem**. Why?)

Inflated risk of Type I error (*i.e.* false positives)

Underdispersion?

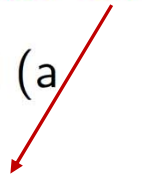
Can it happen that the observations are less variable than expected?

Yes: Especially when observations are dependent.

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

In that case, your p-values may be too large, that is, the results are overly conservative.

The `quasipoisson regression` is a pragmatic solution in that scenario as well (a negative binomial regression is not!)



Inflated risk of Type II error
(false negatives)

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 scenario.

A note on interpretation and model selection

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

- Use Poisson regression to model count outcomes.
- Pretending that count outcomes are continuous may lead to wrong results.
- The main ingredients of a GLM are
  - The linear predictor
  - The family
  - The link function.
- Model diagnostics are similar as in normal regression (`autoplot()`).
- Always check the dispersion of your data, and correct for possible issues
- Interpret the coefficients by back-transforming to the original scale.
- Analysis of Deviance is the MLE analogue to OLS ANOVA.