Chapter 4: Beyond the GLM

Noryanti Muhammad
Centre for Mathematical Sciences
College of Computing and Applied Sciences
Universiti Malaysia Pahang

Centre of Excellence (CoE) for Data Science & Artificial Intelligence
Research & Innovation Department
Universiti Malaysia Pahang

Expected Outcomes:

By the end of this chapter, students should be able:

- ✓ To investigate overdispersion models for in GLM.
- ✓ To develop GLM model considering mixed effect and random effects models.
- ✓ To combine fixed and random effects in the GLM models.

Content:

4.1 Overdispersion

4.1.1 Quasi-likelihood

4.1.2 Direct Models

4.1.3 Random effects

4.2 Smoothing and nonparametric regression

4.2.1 Kernel density estimation

4.2.2 Splines and penalised splines

4.2.3 Generalised additive models

4.2.4 Linear smoothing

4.3 Dependence

4.3.1 Random effects

4.3.2 Marginals models

4.3.3 Clustered Data

4.4 Random effect and Mixed Models

4.4.1 Linear mixed models (LMM)

4.4.2 LMM fitting

4.4.3 Restricted Maximum Likelihood (REML)

4.4.4 Estimating random effects

4.4.5 Bayesian LMMse

4.1 Overdispersion

- Overdispersion describes the observation that variation is higher than would be expected. Some distributions do not have a parameter to fit variability of the observation.
- For example, the normal distribution does that through the parameter σ (i.e. the standard deviation of the model), which is constant in a typical regression.
- In contrast, the Poisson distribution has no such parameter, and in fact the variance increases with the mean (i.e. the variance and the mean have the same value).
- Overdispersion is often mentioned together with zero-inflation, but it is distinct. Overdispersion also includes the case where none of your data points are actually 0.

NOTE: could also be lower, underdispersed. This is less often the case, and not all approaches below allow for modelling underdispersion, but some do.

4.1 Overdispersion: Causes

If the Residual Deviance >= Residual df?

1. We may simply have a badly fitting model for one of a number of reasons such as

   - omitted terms or variables in the linear predictor;
   - incorrect relationship between mean and explanatory variables, i.e. we may have the wrong link function or need to transform one or more explanatory variables;
   - outliers.

4.1 Overdispersion: Causes

If the Residual Deviance >= Residual df?

2. The variation may simply be greater than that predicted by model, and it is this phenomenon that is described as **overdispersion**. In essence our model is too restrictive for our data and for the models.

- Proportion data with $Var(Y_i) > n_i\pi_i(1-\pi_i)$.

- Count data with $Var(Y_i) > \mu_i$.

4.1 Overdispersion: Causes

- Variability of experimental material – this can be thought of as individual variability of the experimental units and may give an additional
- Component of variability which is not accounted for by the basic model;
- Correlation between individual responses – for example in cancer studies involving litters of rats we may expect to see some correlation between rats in the same litter;
- Cluster sampling;
- Aggregate level data – the aggregation process can lead to compound distributions;
- Omitted unobserved variables – in some sense the other categories are all special cases of this, but generally in a rather complex way

4.1 Overdispersion: Causes

Notes that:

- Some circumstances the cause of the overdispersion may be apparent from the **nature of the data collection process**,

- Although it should be noted that **different explanations of the overdispersion process** can **lead to the same model**

- In general, it is **difficult to infer the precise** cause, or underlying process, leading to the overdispersion.

4.1 Overdispersion: Consequences

- The standard errors obtained from the model will be incorrect and may be seriously underestimated and consequently we may incorrectly assess the significance of individual regression parameters.

- Changes in deviance associated with model terms will also be too large and this will lead to the selection of overly complex models.

- Interpretation of the model will be incorrect and any predictions will be too precise.

4.1 Overdispersion

- Many a time data admit more variability than expected under the assumed distribution.

- The greater variability than predicted by the generalized linear model random component reflects overdispersion.

- Overdispersion occurs because the mean and variance components of a GLM are related and depends on the same parameter that is being predicted through the independent vector.

- For the binomial response, if $Y_i \sim Bin(n_i, p_i)$, the mean is $\mu_i = n_i p_i$ and the variance $\sigma_i^2 = n_i p_i q_i$.

- Overdispersion means that the data show evidence that the variance of the response y_i is greater than $n_i p_i q_i$.

- Underdispersion is also theoretically possible, but rare in practice. McCullagh and Nelder (1989) say that overdispersion is the rule rather than the exception.

4.1 Overdispersion

- In the context of logistic regression, overdispersion occurs when the discrepancies between the observed responses y_i and their predicted values $\widehat{\mu}_i = n_i p_i$ are larger than what the binomial model would predict.

- Overdispersion arises when the n_i Bernoulli trials that are summarized in a line of the dataset are not identically distributed (i.e. the success probabilities vary from one trial to the next), or not independent (i.e. the outcome of one trial influences the outcomes of other trials).

- In practice, it is impossible to distinguish non-identically distributed trials from non-independence; the two phenomena are intertwined.

4.1 Overdispersion

Issue: If overdispersion is present in a dataset, the estimated standard errors and test statistics the overall goodness-of-fit will be distorted, and adjustments must be made. When a logistic model fitted to *n* binomial proportions is satisfactory, the residual deviance has an approximate χ^2 distribution with ($n - p$) degrees of freedom, where *p* is the number of unknown parameters in the fitted model. Since the expected value of a χ^2 distribution is equal to its degree of freedom, it follows that the residual deviance for a well-fitting model should be approximately equal to its degrees of freedom.

Equivalently, we may say that the **mean deviance** (deviance/df) should be **close to one**. Similarly, if the variance of the data is greater than that under binomial sampling, the residual mean deviance is likely to be greater than 1.

The problem of overdispersion may also be confounded with the problem of omitted covariates. If we have included all the available covariates related to y_i in our model and it still does not fit, it could be because our regression function $x_i^T\beta$ is incomplete. Or it could be due to overdispersion. Unless we collect more data, we cannot do anything about omitted covariates. *But we can **adjust** for overdispersion.*

4.1 Overdispersion

Recognising and testing for overdispersion in R

- Start with an example to get the point visualized using summary and plot (package lme4).
- Run the glm model (package glm).
- Check the residual deviance.
- Could consider overdispersion test by using package AER or DHARMa.

Example

Count Data

Number of ticks on the heads of red grouse chicks sampled in the field.

INDEX
(factor) chick number (observation level)
TICKS
number of ticks sampled
BROOD
(factor) brood number
HEIGHT
height above sea level (meters)

YEAR
year (-1900)
LOCATION
(factor) geographic location code
cHEIGHT
centered height, derived from HEIGHT

Reference

Elston, D. A., R. Moss, T. Boulinier, C. Arrowsmith, and X. Lambin. 2001. "Analysis of Aggregation, a Worked Example: Numbers of Ticks on Red Grouse Chicks." Parasitology 122 (05): 563-569. doi:10.1017/S0031182001007740.

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=82701.

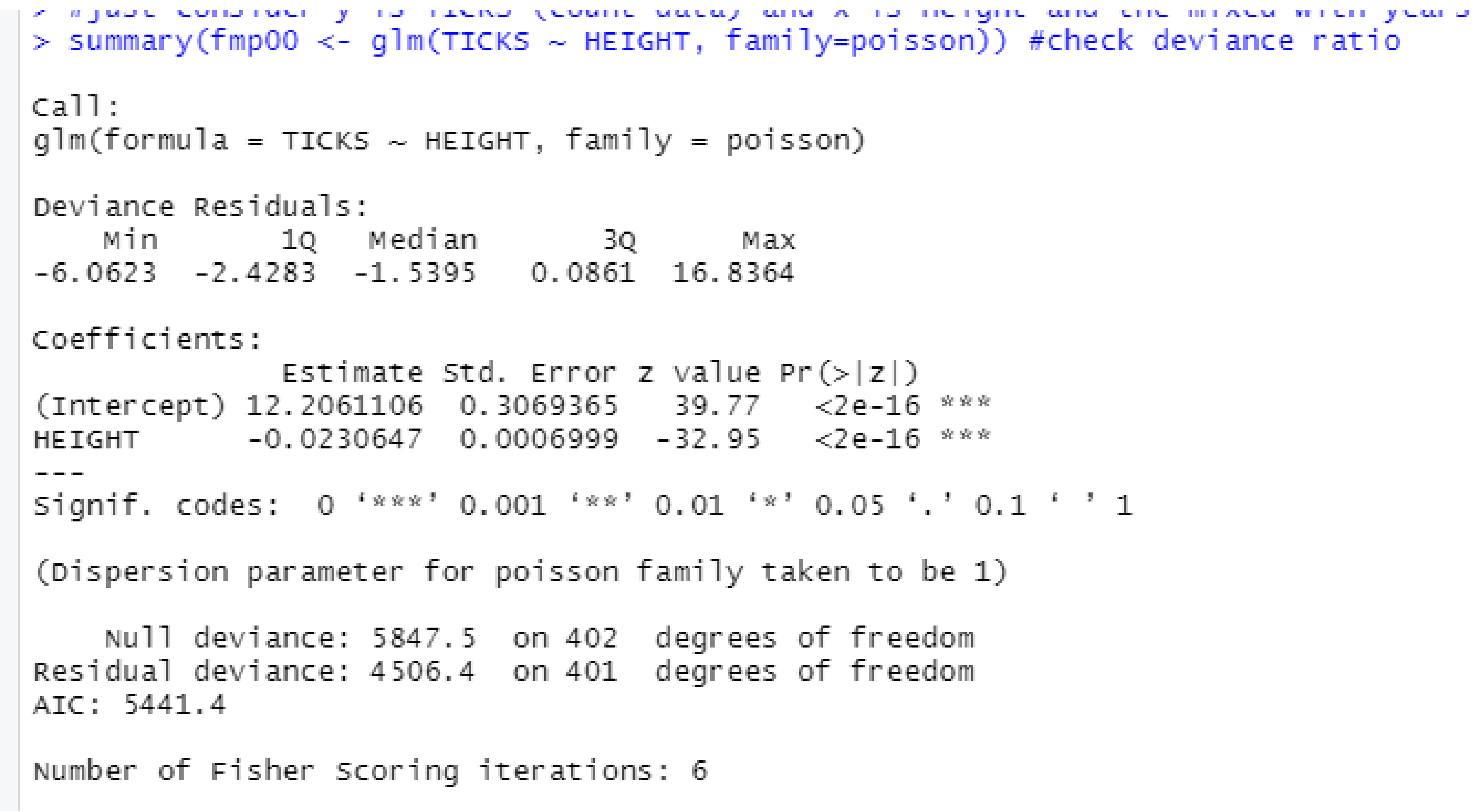
Example

```
library(lme4)
data(grouseticks)
summary(grouseticks)
# INDEX is individual
head(grouseticks)
attach(grouseticks)
hist(TICKS, col="grey", border=NA, las=1, breaks=0:90)
plot(TICKS ~ HEIGHT, las=1)
summary(fmp <- glm(TICKS ~ HEIGHT*YEAR,
family=poisson))
Call:
glm(formula = TICKS ~ HEIGHT * YEAR, family = poisson)
```

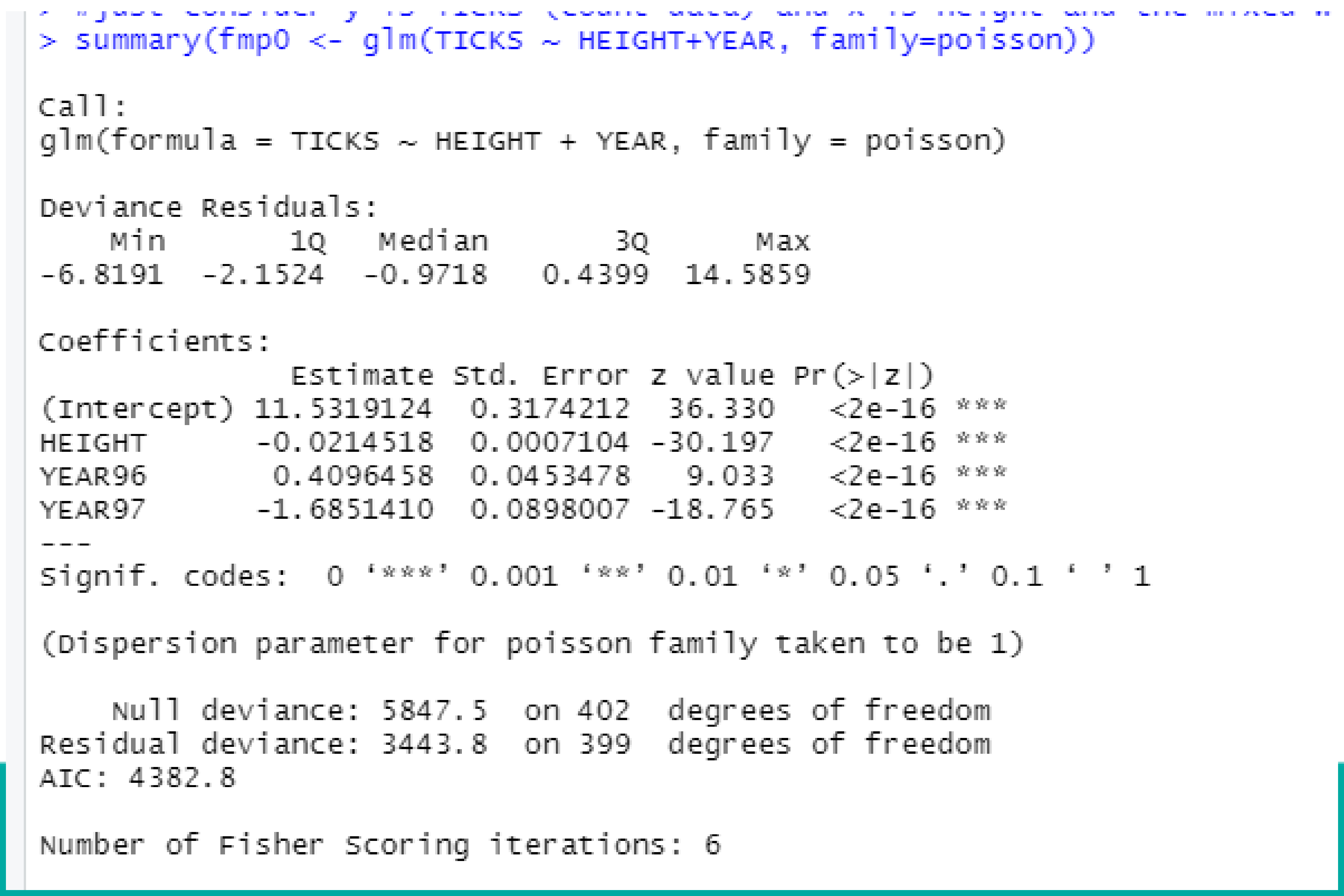
```
library(AER)
dispersiontest(fmp)

library(devtools) # assuming you have that
devtools::install_github(repo = "DHARMa",
username = "florianhartig", subdir = "DHARMa")

library(DHARMa)
sim_fmp <- simulateResiduals(fmp, refit=T)
testOverdispersion(sim_fmp)
plotSimulatedResiduals(sim_fmp)
```

In this case, our residual deviance is 3000 for 397 degrees of freedom. The rule of thumb is that the ratio of deviance to df should be 1, but it is 7.6, indicating severe overdispersion.

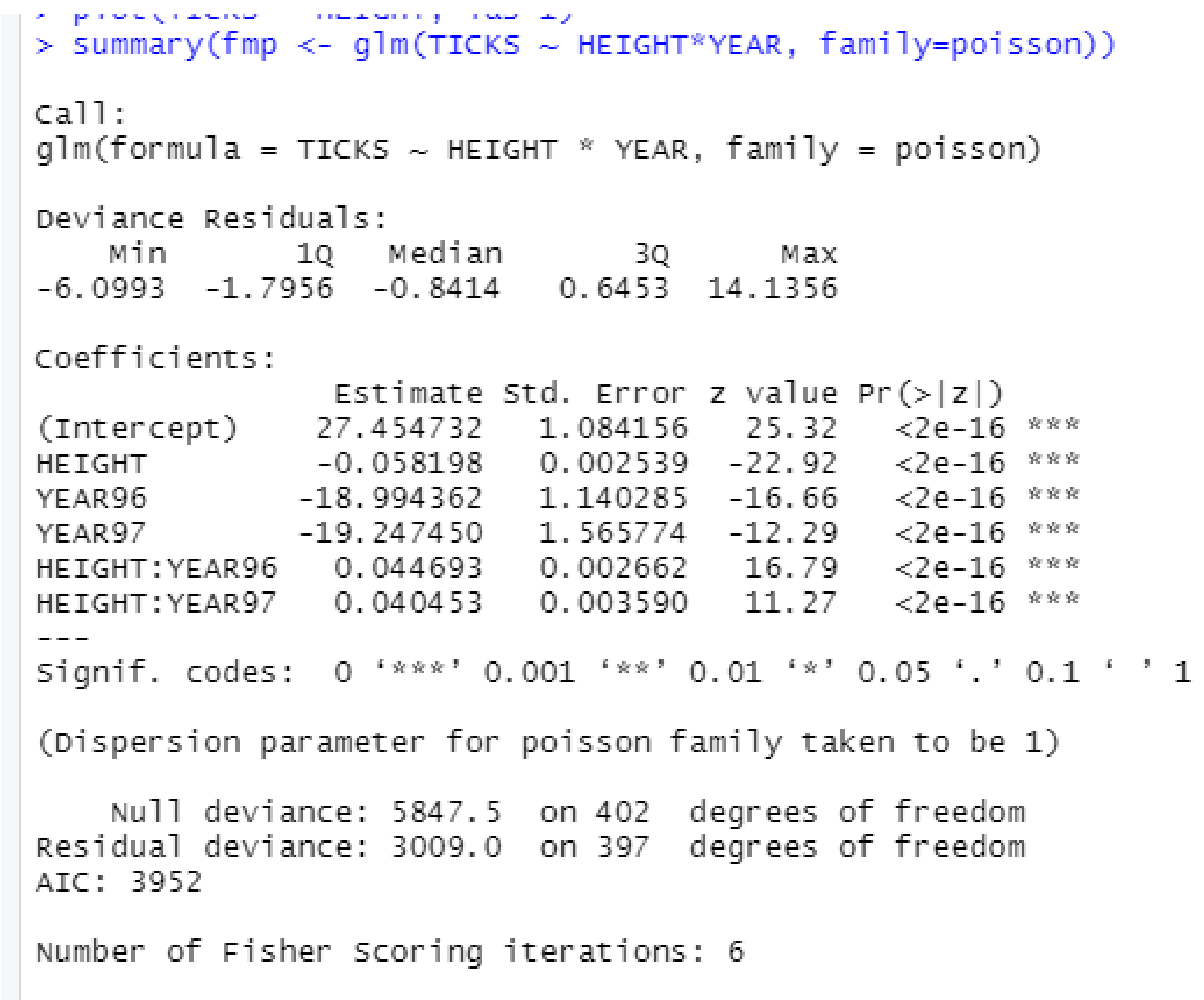
OUTPUT



```
> #just consider y is TICKS (count data) and x is height and the mixed with years
> summary(fmp00 <- glm(TICKS ~ HEIGHT, family=poisson)) #check deviance ratio

Call:
glm(formula = TICKS ~ HEIGHT, family = poisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.0623  -2.4283  -1.5395   0.0861  16.8364  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) 12.2061106  0.3069365   39.77   <2e-16 ***
HEIGHT      -0.0230647  0.0006999  -32.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5847.5  on 402  degrees of freedom
Residual deviance: 4506.4  on 401  degrees of freedom
AIC: 5441.4

Number of Fisher Scoring iterations: 6
```

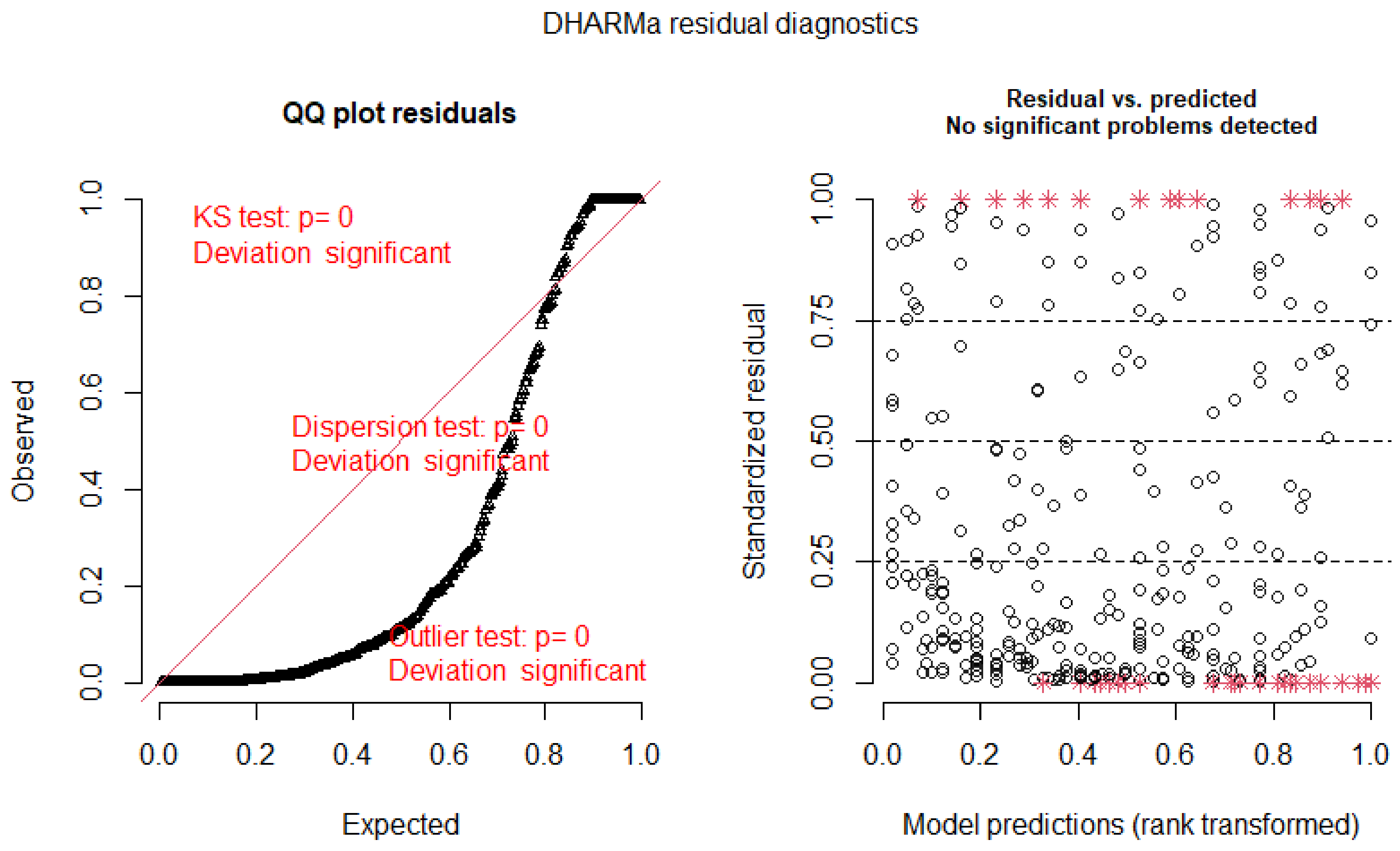


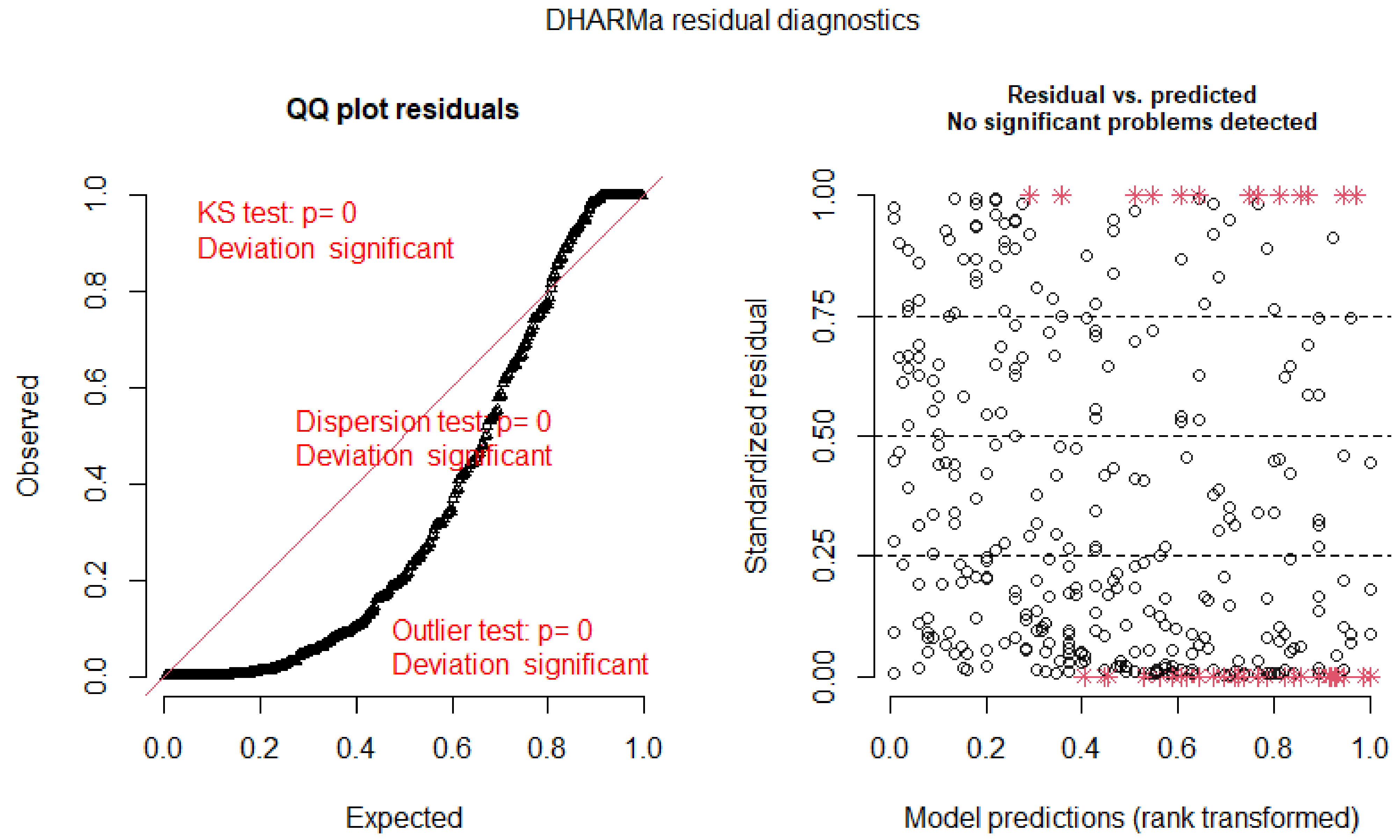
```
> #just consider y is TICKS (count data) and x is height and the mixed w
> summary(fmp0 <- glm(TICKS ~ HEIGHT+YEAR, family=poisson))

Call:
glm(formula = TICKS ~ HEIGHT + YEAR, family = poisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.8191  -2.1524  -0.9718   0.4399  14.5859  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) 11.5319124  0.3174212  36.330   <2e-16 ***
HEIGHT      -0.0214518  0.0007104 -30.197   <2e-16 ***
YEAR96       0.4096458  0.0453478   9.033   <2e-16 ***
YEAR97      -1.6851410  0.0898007 -18.765   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5847.5  on 402  degrees of freedom
Residual deviance: 3443.8  on 399  degrees of freedom
AIC: 4382.8

Number of Fisher Scoring iterations: 6
```

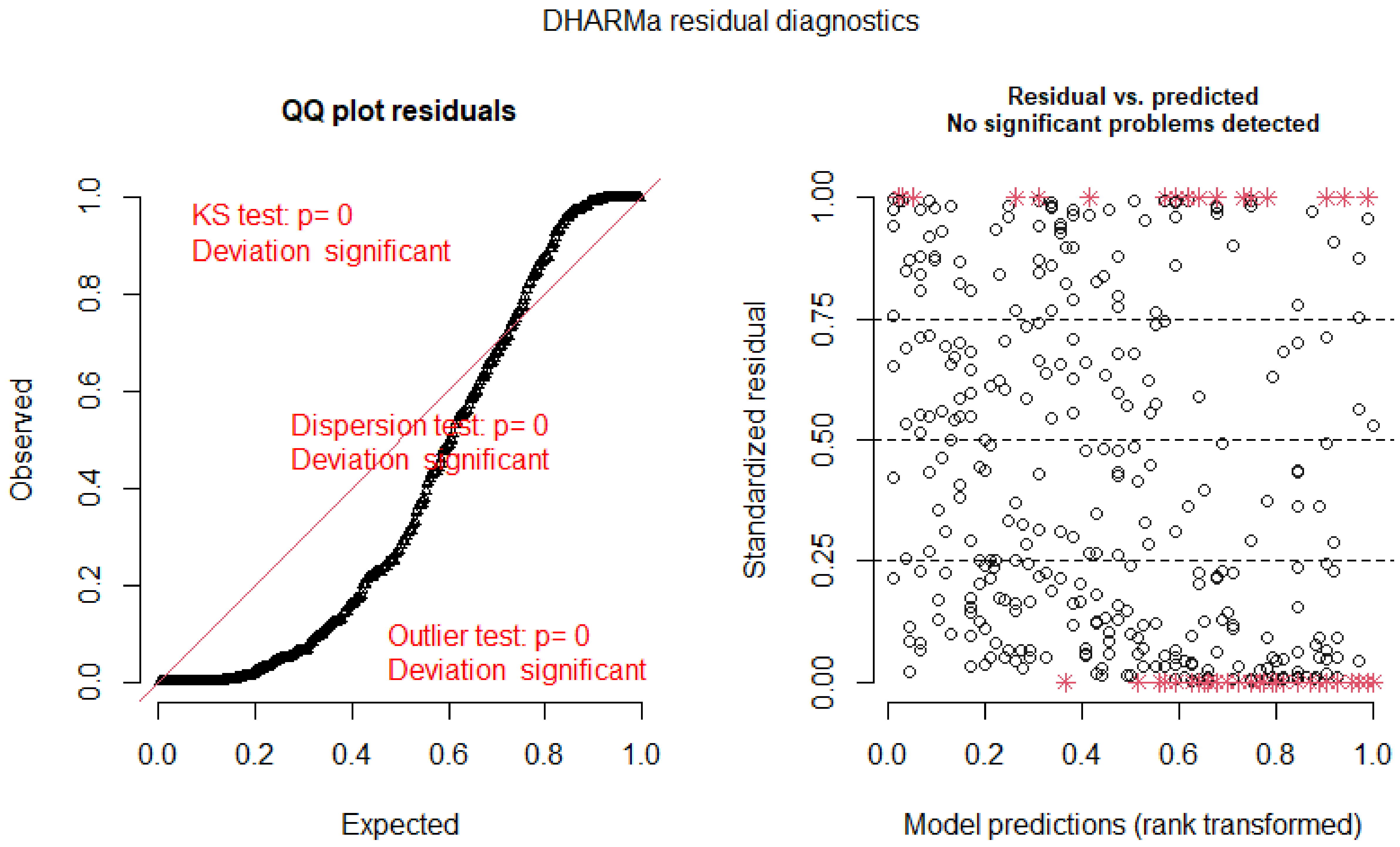


```
> plot(TICKS ~ HEIGHT, las=1)
> summary(fmp <- glm(TICKS ~ HEIGHT*YEAR, family=poisson))

Call:
glm(formula = TICKS ~ HEIGHT * YEAR, family = poisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.0993  -1.7956  -0.8414   0.6453  14.1356  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)    27.454732   1.084156   25.32   <2e-16 ***
HEIGHT         -0.058198   0.002539  -22.92   <2e-16 ***
YEAR96        -18.994362   1.140285  -16.66   <2e-16 ***
YEAR97        -19.247450   1.565774  -12.29   <2e-16 ***
HEIGHT:YEAR96   0.044693   0.002662   16.79   <2e-16 ***
HEIGHT:YEAR97   0.040453   0.003590   11.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5847.5  on 402  degrees of freedom
Residual deviance: 3009.0  on 397  degrees of freedom
AIC: 3952

Number of Fisher Scoring iterations: 6
```

OUTPUT


DHARMa residual diagnostics
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
0.0
0.2
0.4
0.6
0.8
1.0
Residual vs. predicted
No significant problems detected
Standardized residual
Model predictions (rank transformed)
0.00
0.25
0.50
0.75
1.00


DHARMa residual diagnostics
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
0.0
0.2
0.4
0.6
0.8
1.0
Residual vs. predicted
No significant problems detected
Standardized residual
Model predictions (rank transformed)
0.00
0.25
0.50
0.75
1.00


DHARMa residual diagnostics
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
0.0
0.2
0.4
0.6
0.8
1.0
Residual vs. predicted
No significant problems detected
Standardized residual
Model predictions (rank transformed)
0.00
0.25
0.50
0.75
1.00

4.1.1 Quasi-likelihood

“Fixing” overdispersion

- The quasi-families augment the normal families by adding a dispersion parameter.

- Constant overdispersion models (as in Slide 6) can be fitted into the class of simple quasi-likelihood models.

$$Q = -\tfrac{1}{2}\sum_{i=1}^{n}\left\{\frac{D(y_i,\mu_i)}{\phi}\right\}, \quad \Longrightarrow \quad D(y,\mu) = -2\int_{y}^{\mu}\frac{(y-t)}{V(t)}dt.$$

- For overdispersed Binomial and Poisson models, can be consider as

$$\tilde{\phi} = \frac{1}{(n-p)}\sum_{i=1}^{n}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}, \qquad \tilde{\phi} = \frac{1}{(n-p)}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

4.1.1 Quasi-likelihood

“Fixing” overdispersion

- Extended quasi-likelihood models can be used for complex overdispersion model and can be generally described with a variance of the form

$$\mathrm{Var}(Y_i) = \phi_i(\gamma)V_i(\mu_i, \boldsymbol{\lambda})$$

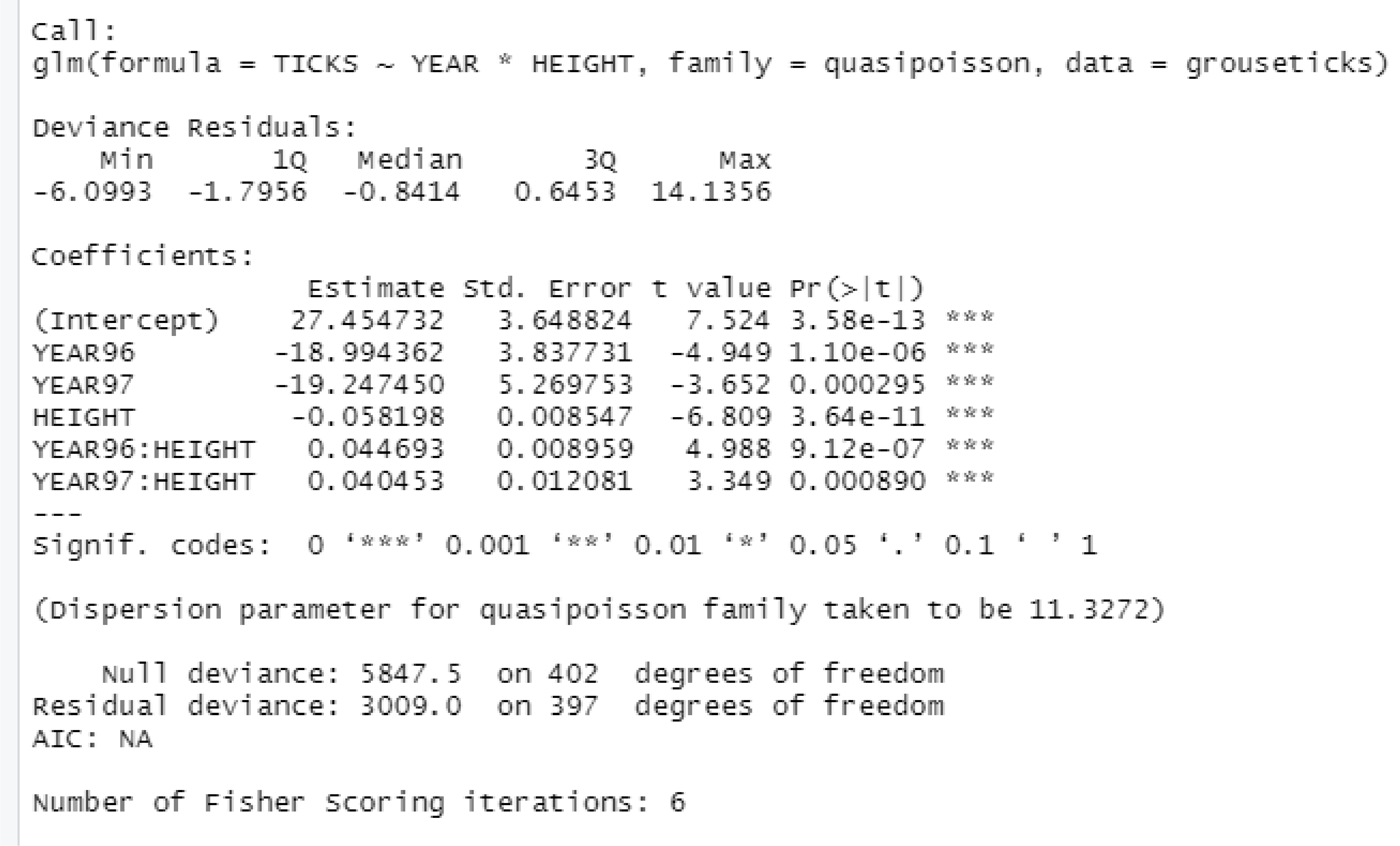
- Where the scale parameter ϕ_i and the variance function $V_i(.)$ may depend upon additional parameters.

- Estimate by maximizing the extended quasi likelihood function

$$Q^+ = -\tfrac{1}{2}\sum_{i=1}^{n}\left\{\frac{D(y_i,\mu_i)}{\phi_i} + \log\left(2\pi\phi_i V_i(y_i)\right)\right\}, \quad \Longrightarrow \quad D(y,\mu) = -2\int_y^{\mu}\frac{(y-t)}{V_i(t)}dt.$$

Example

- Based on previous example.
- τ is estimated as 11.3, a value similar to those in the overdispersion tests in previous diagnostic check.
- The main effect is the substantially larger errors for the estimates (the point estimates do not change), and hence potentially changed significances (though not here).
- Can manually compute the corrected standard errors as Poisson-standard errors*$\sqrt{\tau}$



```
Call:
glm(formula = TICKS ~ YEAR * HEIGHT, family = quasipoisson, data = grouseticks)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.0993  -1.7956  -0.8414   0.6453  14.1356  

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    27.454732   3.648824   7.524 3.58e-13 ***
YEAR96        -18.994362   3.837731  -4.949 1.10e-06 ***
YEAR97        -19.247450   5.269753  -3.652 0.000295 ***
HEIGHT         -0.058198   0.008547  -6.809 3.64e-11 ***
YEAR96:HEIGHT   0.044693   0.008959   4.988 9.12e-07 ***
YEAR97:HEIGHT   0.040453   0.012081   3.349 0.000890 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 11.3272)

    Null deviance: 5847.5  on 402  degrees of freedom
Residual deviance: 3009.0  on 397  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

NOTES

Note that because this is no maximum likelihood method (but a quasi-likelihood method), no likelihood and hence no AIC are available..

4.1.1 Quasi-likelihood

- Most popular method for adjusting for overdispersion comes from the theory of quasi-likelihood.

- Quasi-likelihood has come to play a very important role in modern statistics.

- Foundation of many methods that are thought to be "robust" (e.g. Generalized Estimating Equations (GEE) for longitudinal data) because they do not require specification of a full parametric model. For more details see Agresti (2007, Sec 9.2) or Agresti (2013, Sec 12.2).

4.1.1 Quasi-likelihood

- First specify the "mean function" which determines how $\mu_i = E(Y_i)$ is related to the covariates. In the context of logistic regression, the mean function is $\mu_i = n_i expit(x_i^T \beta)$ which implies
  - $$log\left(\frac{p}{1-p}\right) = x_i^T \beta.$$
- To account for overdispersion, we will include another factor σ^2 called the "scale parameter," so that $V(Y) = \sigma^2 npq$.
  - If $\sigma^2 \neq 1$ then the model is not binomial; $\sigma^2 > 1$ is called "overdispersion" and $\sigma^2 < 1$ is called "**underdispersion**."
  - If σ^2 were known, we could obtain a consistent, asymptotically normal and efficient estimate for β by a quasi-scoring procedure, sometimes called "estimating equations."

NOTES

For the variance function shown above, the quasi-scoring procedure reduces to the Fisher scoring algorithm that we mentioned as a way to iteratively find ML estimates.

4.1.1 Quasi-likelihood

Note that no matter what σ^2 is assumed to be, we get the same estimate for β .

Therefore, this method for overdispersion does not change the estimate for β at all. However, the estimated covariance for $\hat{\beta}$ changes from

$$\hat{V}\left(\hat{\beta}\right) = (X^T WX)^{-1} \quad \text{to} \quad \hat{V}\left(\hat{\beta}\right) = \sigma^2 (X^T WX)^{-1}$$

That is, the estimated standard errors must be multiplied by the factor $\sigma = \sqrt{\sigma^2}$.

How do we estimate σ^2?

McCullagh and Nelder (1989) recommend

$$\sigma^2 = {X^2}/{(N-P)}$$

where X^2 is the usual Pearson goodness-of-fit statistic, N is the number of sample cases (number of rows in the dataset we are modeling), and P is the number of parameters.

If the model holds, then ${X^2}/{(N-p)}$ is a consistent estimate for σ^2 in the asymptotic sequence $N \to \infty$ for fixed n_i ‘s.

The deviance-based estimate ${G^2}/{(N-p)}$ does not have this consistency property and should not be used.

4.1.1 Quasi-likelihood

- This is a reasonable way to estimate $\boldsymbol{\sigma}$^2 if the mean model μ_i=g(X_i^T β) holds. But if important covariates are omitted, then X^2 tends to grow and the estimate for σ^2 can be too large. For this reason, we will estimate σ^2 under a maximal model, a model that includes all the covariates we wish to consider.

- The best way to estimate σ^2 is to identify a rich model for μ_i and designate it to be the most complicated one that we are willing to consider.

- For example, if we have a large pool of potential covariates, we may take the maximal model to be the model that has every covariate included as a main effect. Or, if we have a smaller number of potential covariates, we decide to include all main effects along with two-way and perhaps even three-way interactions. But we must omit at least a few higher-order interactions, otherwise we will end up with a model that is saturated.

4.1.1 Quasi-likelihood

In an overdispersed model,

- We must also adjust our test statistics. The statistics X^2 and G^2 are adjusted by dividing them by σ^2 . That is, tests of nested models are carried out by comparing differences in the scaled Pearson statistic, ${}^{\Delta X^2}/_{\sigma^2}$, or the scaled deviance, ${}^{\Delta G^2}/_{\sigma^2}$ to a chi-square distribution with Δdf (degrees of freedom).

If the **data are overdispersed** — that is, if

$$V(y_i) \approx \sigma^2 n_i \pi_i (1 - \pi_i)$$

- For a scale factor $\sigma^2 > 1$, then the residual plot may still resemble a horizontal band, but many of the residuals will tend to fall outside the ± 3 limits. In this case, the denominator of the Pearson residual will tend to understate the true variance of the y_i, making the residuals larger. If the plot looks like a horizontal band but X^2 and G^2 indicate lack of fit, an adjustment for overdispersion might be warranted.

- A warning about this, however: If the residuals tend to be too large, it doesn't necessarily mean that overdispersion is the cause. Large residuals may also be caused by omitted covariates. If some important covariates are omitted from x_i, then the true π_i 's will depart from what your model predicts, causing the numerator of the Pearson residual to be larger than usual. That is, apparent overdispersion could also be an indication that your mean model needs additional covariates. If these additional covariates are not available in the dataset, however, then there's not much we can do about it; we may need to attribute it to overdispersion.

4.1.1 Quasi-likelihood

NOTE

- There is **no overdispersion for ungrouped data**.
- McCullagh and Nelder (1989) point out that overdispersion is not possible if $n_i = 1$. If y_i only takes values 0 and 1, the it must be distributed as Bernoulli (π_i) and its variance must be $\pi_i(1 - \pi_i)$. There is no other distribution with support $\{0,1\}$.
- Therefore, with ungrouped data, we should always assume *scale*=1 and not try to estimate a scale parameter and adjust for overdispersion.

4.1.1 Quasi-likelihood

Summary of Adjusting for Overdispersion in the Binary Logistic Regression

- The usual way to correct for overdispersion in a logit model is to assume that:

$$E(Y_i) = n_i\pi_i \quad \text{and} \quad \text{Var}(Y_i) = \sigma^2 n_i \pi_i (1 - \pi_i)$$

  where σ^2 is a scale parameter.

- Under this modification, the Fisher-scoring procedure for estimating β does not change, but its estimated covariance matrix becomes $\sigma^2(X^T WX)^{-1}$ −that is, the usual standard errors are multiplied by the square root of σ^2.

- This will make the confidence intervals wider.

4.1.2 Direct Model

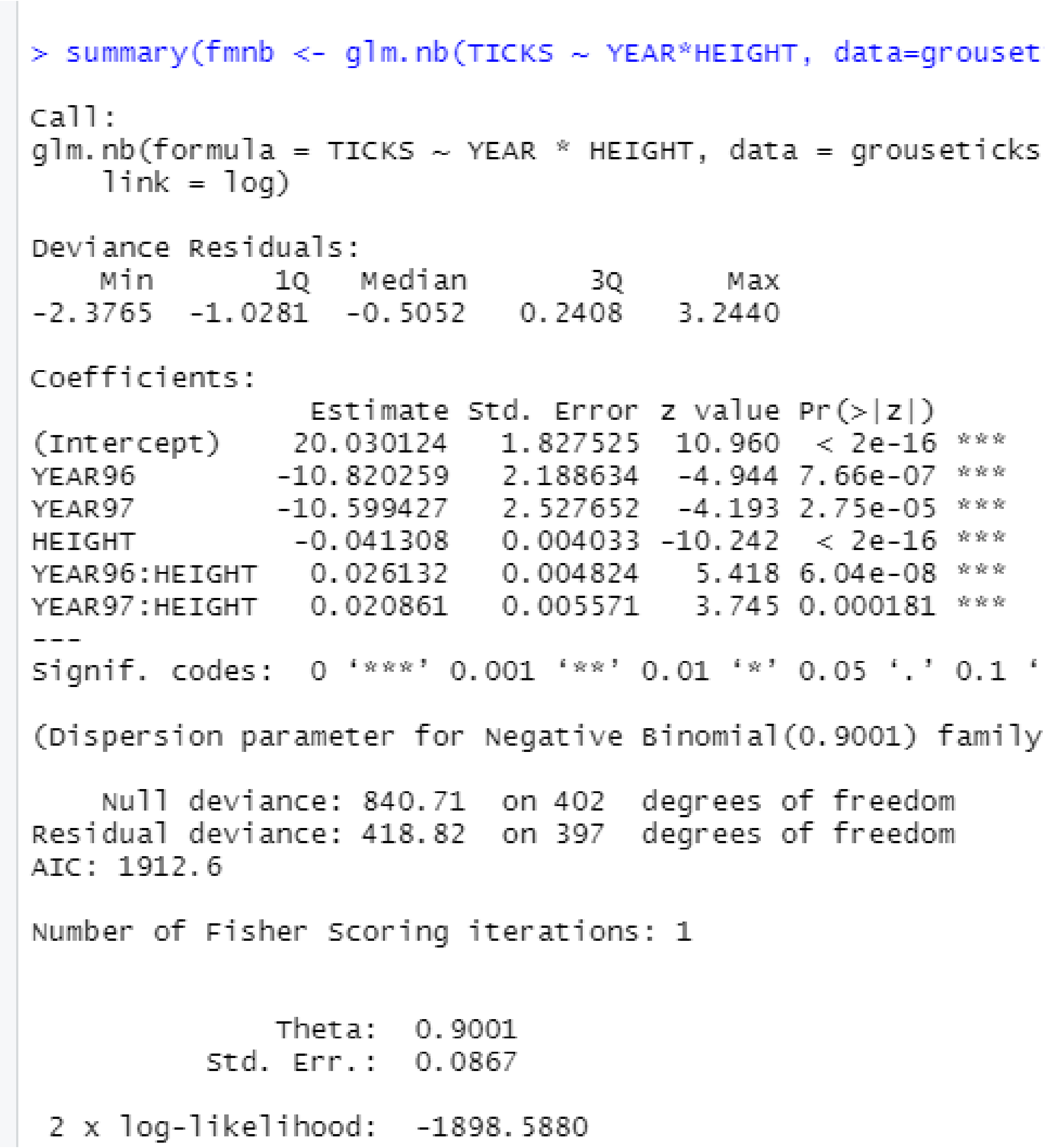
- Poisson-lognormal model for counts or binomial-logit-Normal model for proportions (see above, “observation-level random effects”).
- Negative binomial for counts or beta-binomial for proportions.

Example

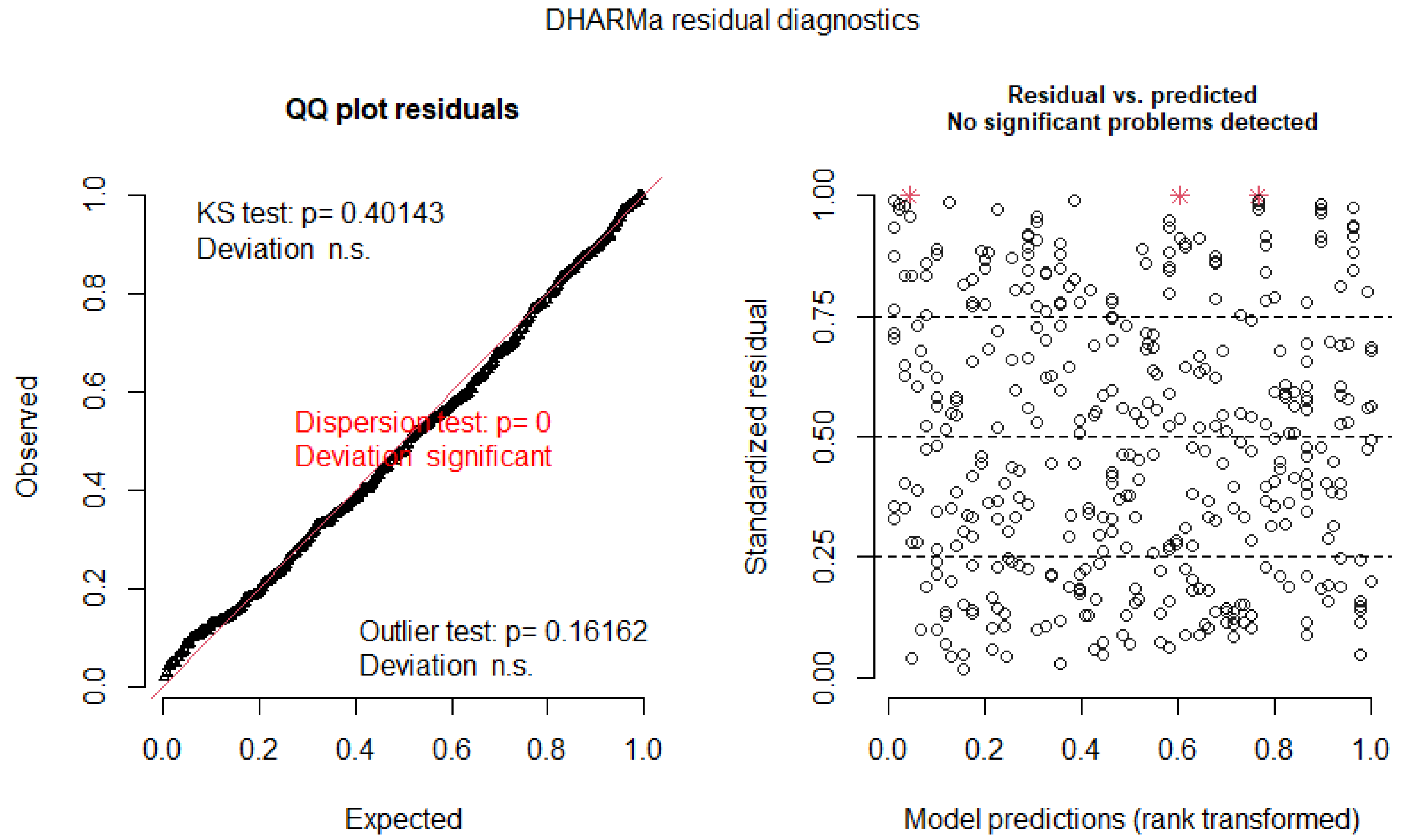
Different distribution

```
library(MASS)
summary(fmnb <- glm.nb(TICKS ~ YEAR*HEIGHT, data=grouseticks))
#check
sim_fmnb <- simulateResiduals(fmnb, refit=T, n=99)
plotSimulatedResiduals(sim_fmnb)
```

- Maybe our distributional assumption was simply wrong, and we choose a different distribution.
- For Poisson, the most obvious “upgrade” is the **negative binomial**, which includes in fact a dispersion parameter similar to τ .



```
> summary(fmnb <- glm.nb(TICKS ~ YEAR*HEIGHT, data=grouseticks))

Call:
glm.nb(formula = TICKS ~ YEAR * HEIGHT, data = grouseticks, init.theta = 0.9000852793,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3765  -1.0281  -0.5052   0.2408   3.2440

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   20.030124   1.827525  10.960  < 2e-16 ***
YEAR96       -10.820259   2.188634  -4.944 7.66e-07 ***
YEAR97       -10.599427   2.527652  -4.193 2.75e-05 ***
HEIGHT        -0.041308   0.004033 -10.242  < 2e-16 ***
YEAR96:HEIGHT  0.026132   0.004824   5.418 6.04e-08 ***
YEAR97:HEIGHT  0.020861   0.005571   3.745 0.000181 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.9001) family taken to be 1)

    Null deviance: 840.71  on 402  degrees of freedom
Residual deviance: 418.82  on 397  degrees of freedom
AIC: 1912.6

Number of Fisher Scoring iterations: 1


              Theta:  0.9001
          Std. Err.:  0.0867

 2 x log-likelihood:  -1898.5880
```



4.1.3 Random effects

- Assume that the linear predictor, η_i has some continuous distribution and this distribution is taken to be the location-scale family then the corresponds to including ad additive random effect in the linear predictor.

$$\eta_i = \mathbf{x}_i^T \beta + \sigma Z_i$$

- Where Z_i is assumed to be from the standardised form of the distribution.

- Most commonly Z_i is taken to be normally distributed leading to the logistic normal and probit-normal models for proportion data and the Poisson-normal model for count data.

- Example:

$$Y_i \sim \text{Bin}(m_i, \pi_i), \quad \text{logit}(\pi_i) = \eta_i = \mathbf{x}_i^T \beta + \sigma Z_i \quad \text{and} \quad Z_i \sim N(0,1).$$

4.1.3 Random effects

- Another example: General idea is to allow the expectation to vary more than a Poisson distribution would suggest. Therefore, multiply the Poisson-expectation with an overdispersion parameter (larger 1) as follows

$$Y \sim Pois(\lambda = e^{\tau} \cdot E(Y)) = Pois(\lambda = e^{\tau} \cdot e^{aX+b}),$$

- Where $E(Y)$ is the prediction from regression model. Without dispersion, $\tau = 0$. here, use e^{τ} to force this factor to be positive.

- Recall: the Poisson-regression uses a log-link, so we can reformulate the above formulation to

$$Y \sim Pois(\lambda = e^{\tau} \cdot e^{aX+b}) = Pois(\lambda = e^{aX+b+\tau}).$$

- So, the overdispersion multiplier at the response-scale becomes an overdispersion summand at the log-scale.

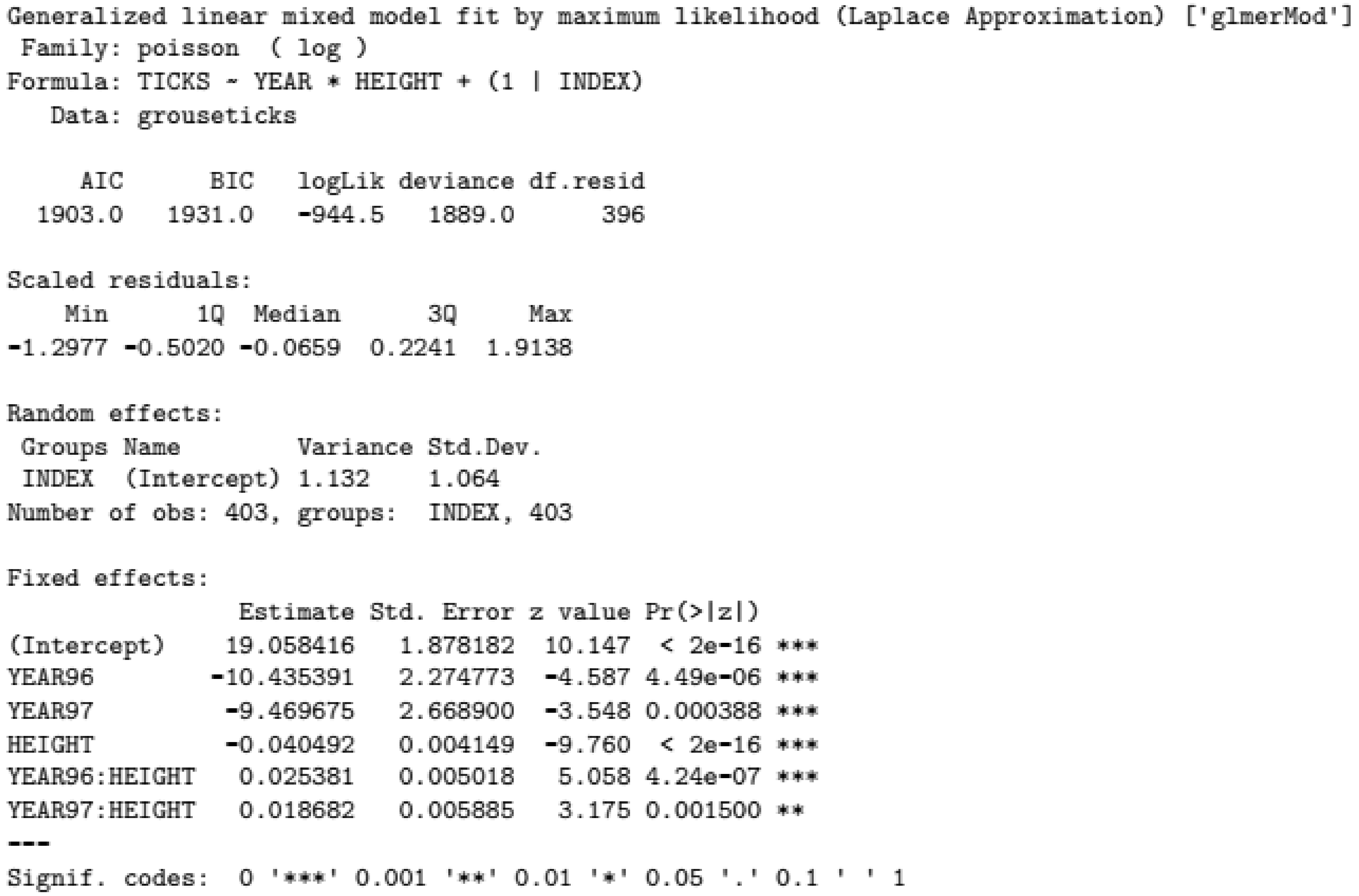
Example

Random Effect

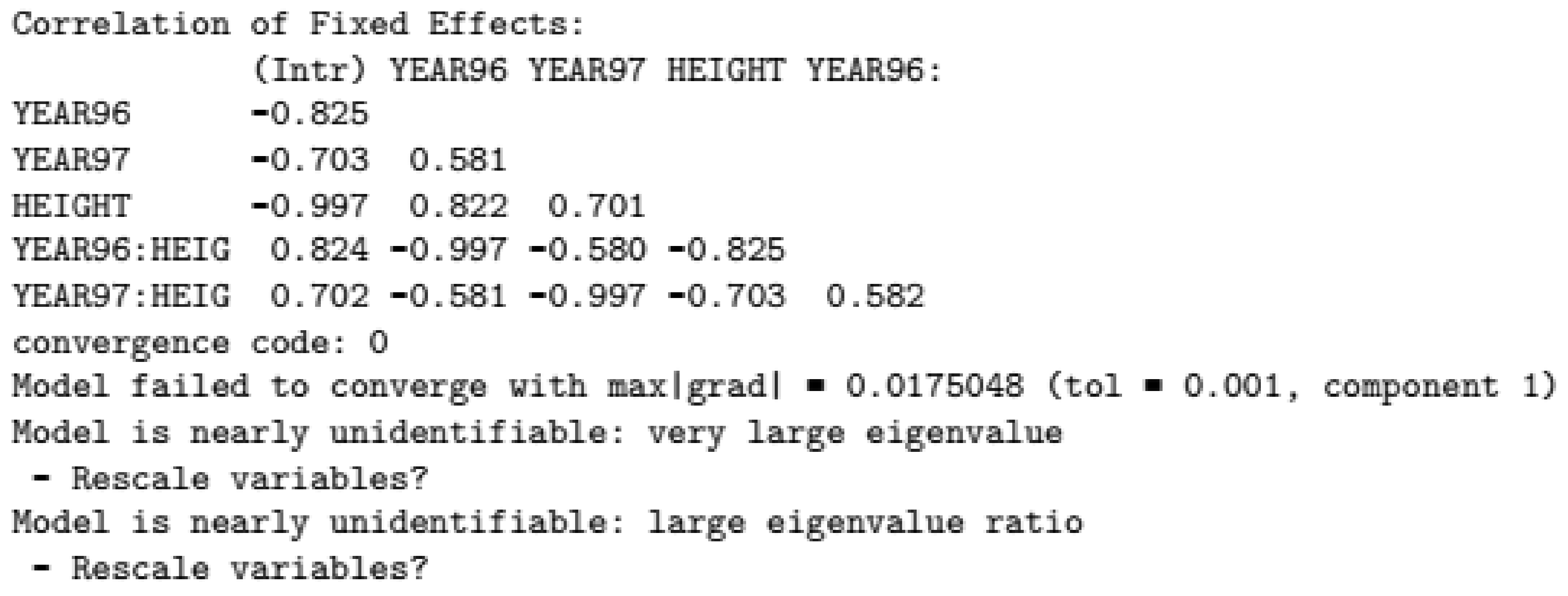
```
llibrary(lme4)
summary(fmOLRE <- glmer(TICKS ~ YEAR*HEIGHT + (1|INDEX),
family=poisson, data=grouseticks))
```

- Modelling overdispersion as Observation-level random effects (OLRE) is very simple: **just add a random effect which is different for each observation**.

- In our data set, the column INDEX is just a continuously varying value from 1 to N, which we use as random effect.



```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: TICKS ~ YEAR * HEIGHT + (1 | INDEX)
   Data: grouseticks

     AIC      BIC   logLik deviance df.resid
  1903.0   1931.0   -944.5   1889.0      396

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2977 -0.5020 -0.0659  0.2241  1.9138

Random effects:
 Groups Name        Variance Std.Dev.
 INDEX  (Intercept) 1.132    1.064
Number of obs: 403, groups:  INDEX, 403

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    19.058416   1.878182  10.147  < 2e-16 ***
YEAR96        -10.435391   2.274773  -4.587 4.49e-06 ***
YEAR97         -9.469675   2.668900  -3.548 0.000388 ***
HEIGHT         -0.040492   0.004149  -9.760  < 2e-16 ***
YEAR96:HEIGHT   0.025381   0.005018   5.058 4.24e-07 ***
YEAR97:HEIGHT   0.018682   0.005885   3.175 0.001500 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



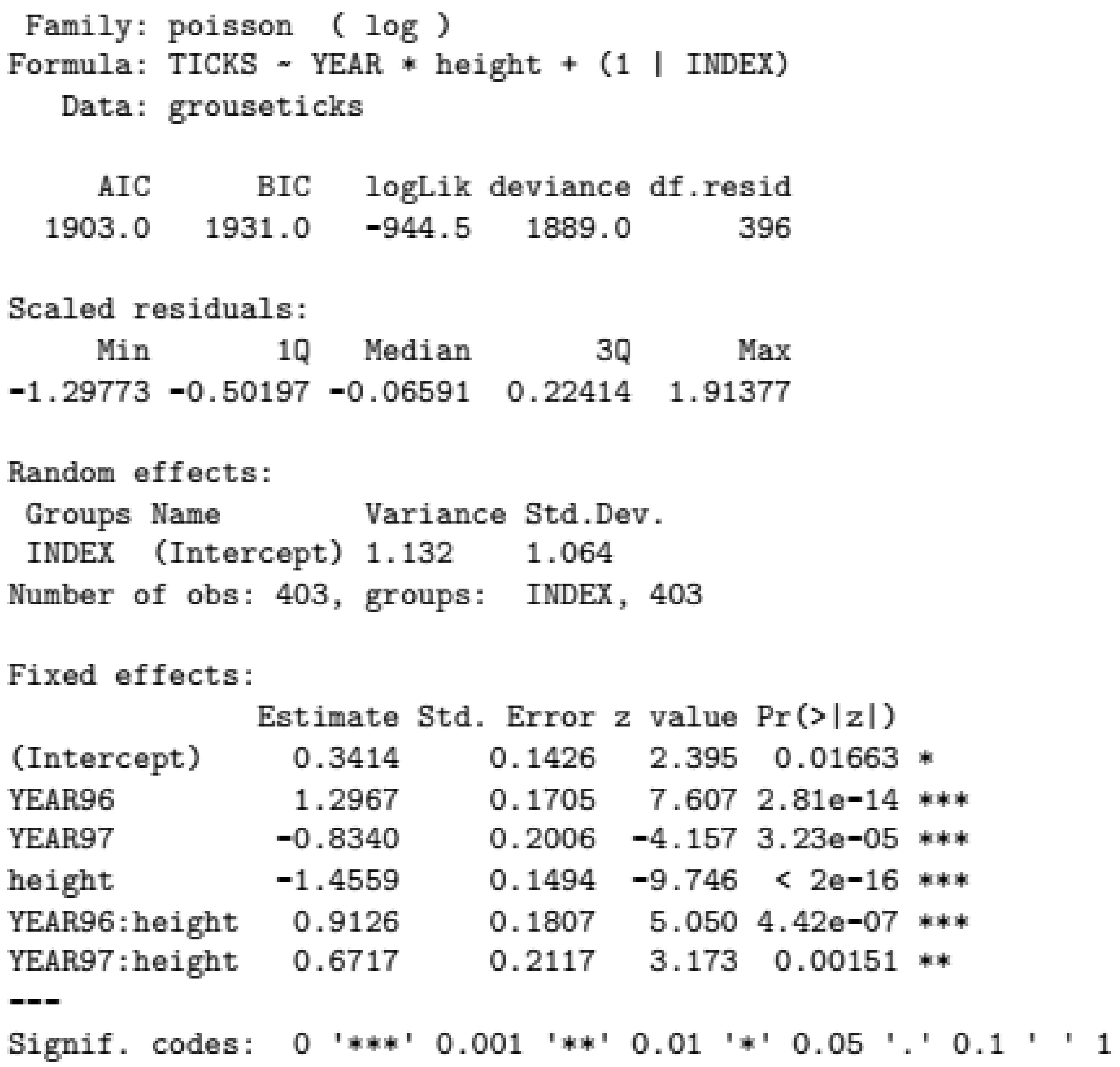
```
Correlation of Fixed Effects:
            (Intr) YEAR96 YEAR97 HEIGHT YEAR96:
YEAR96      -0.825
YEAR97      -0.703  0.581
HEIGHT      -0.997  0.822  0.701
YEAR96:HEIG  0.824 -0.997 -0.580 -0.825
YEAR97:HEIG  0.702 -0.581 -0.997 -0.703  0.582
convergence code: 0
Model failed to converge with max|grad| = 0.0175048 (tol = 0.001, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
```

Example

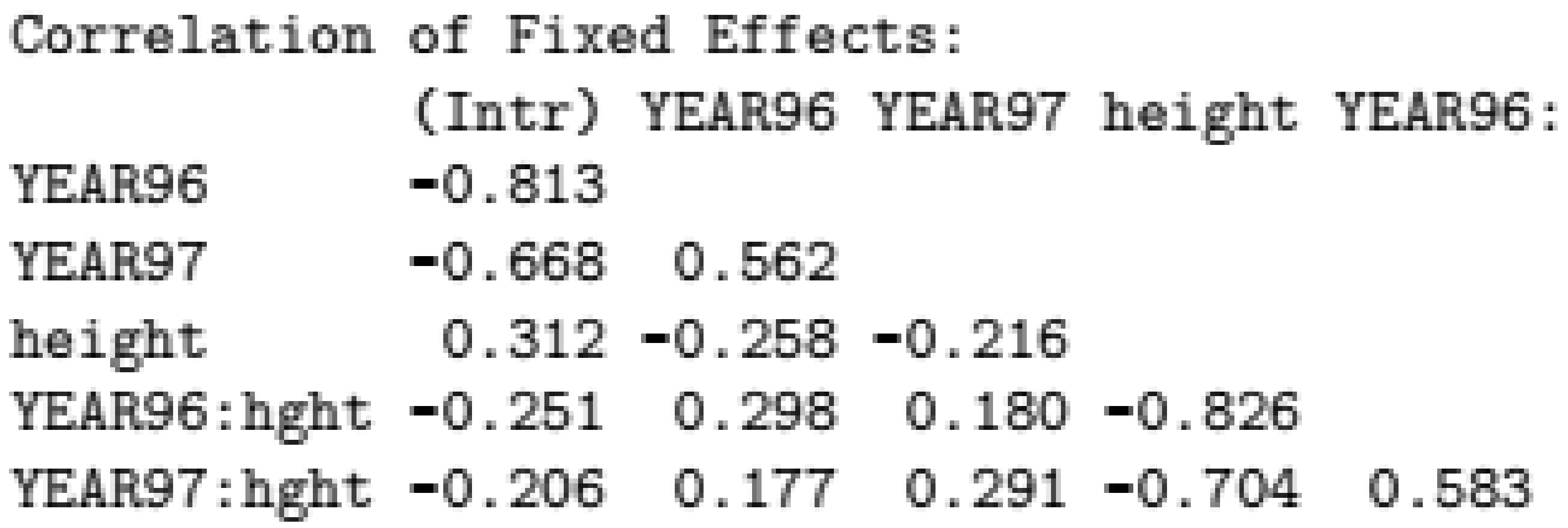
Random Effect

```
height <- scale(grouseticks$HEIGHT)
summary(fmOLRE <- glmer(TICKS ~ YEAR*height + (1|INDEX),
family=poisson, data=grouseticks))
```

- Scale the numeric predictor.



```
 Family: poisson  ( log )
Formula: TICKS ~ YEAR * height + (1 | INDEX)
   Data: grouseticks

     AIC      BIC   logLik deviance df.resid
  1903.0   1931.0   -944.5   1889.0      396

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.29773 -0.50197 -0.06591  0.22414  1.91377

Random effects:
 Groups Name        Variance Std.Dev.
 INDEX  (Intercept) 1.132    1.064
Number of obs: 403, groups:  INDEX, 403

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.3414     0.1426   2.395  0.01663 *
YEAR96          1.2967     0.1705   7.607 2.81e-14 ***
YEAR97         -0.8340     0.2006  -4.157 3.23e-05 ***
height         -1.4559     0.1494  -9.746  < 2e-16 ***
YEAR96:height   0.9126     0.1807   5.050 4.42e-07 ***
YEAR97:height   0.6717     0.2117   3.173  0.00151 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
Correlation of Fixed Effects:
            (Intr) YEAR96 YEAR97 height YEAR96:
YEAR96      -0.813
YEAR97      -0.668  0.562
height       0.312 -0.258 -0.216
YEAR96:hght -0.251  0.298  0.180 -0.826
YEAR97:hght -0.206  0.177  0.291 -0.704  0.583
```

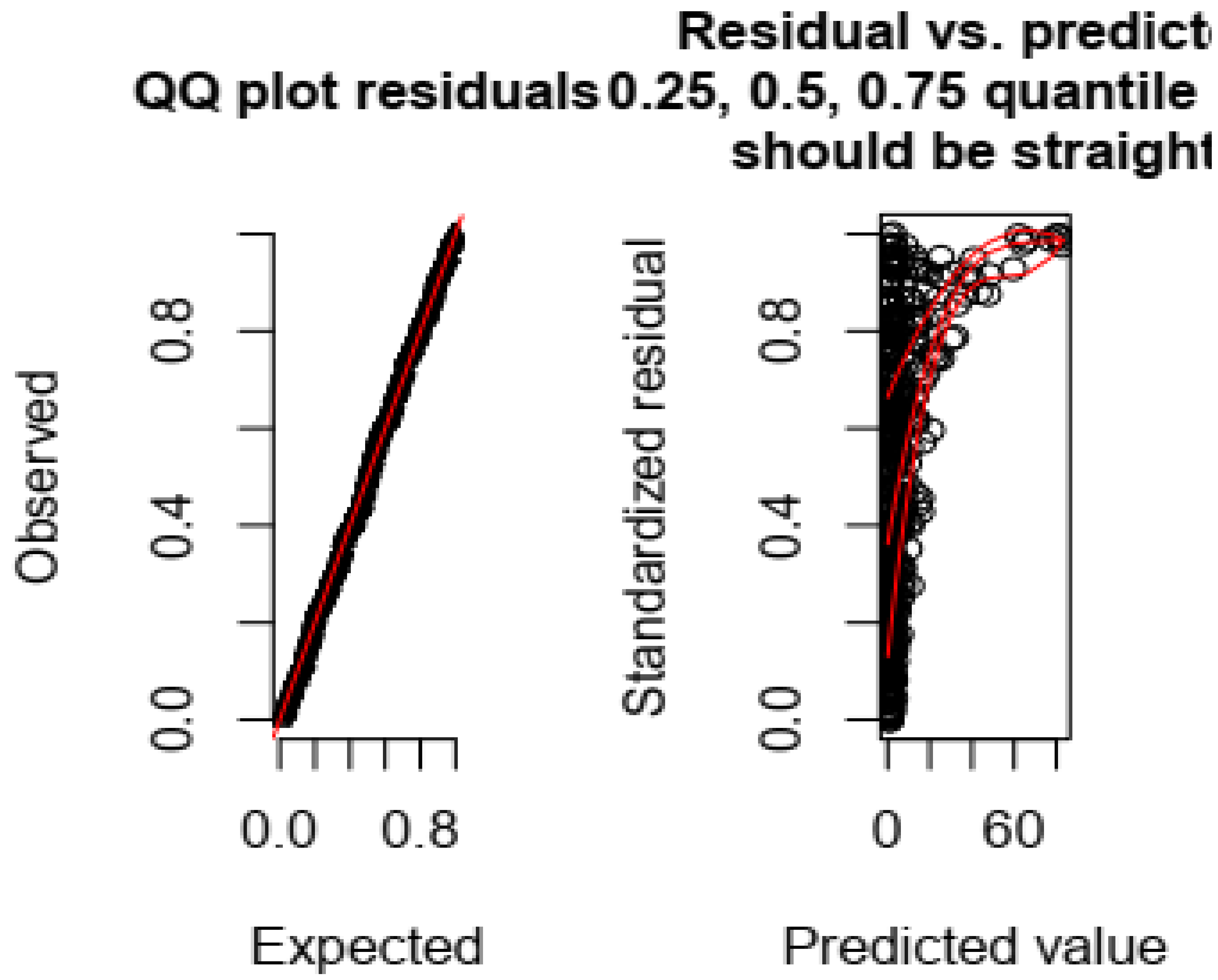
- The standard deviation for the random effect is around 1.06, i.e. similar to what we have simulated (in R coding).
- The overdispersion is thus substantial.
- Note that the estimates or intercept, YEAR96 and YEAR97 are *substantially* different (as is height, but then that has been rescaled.

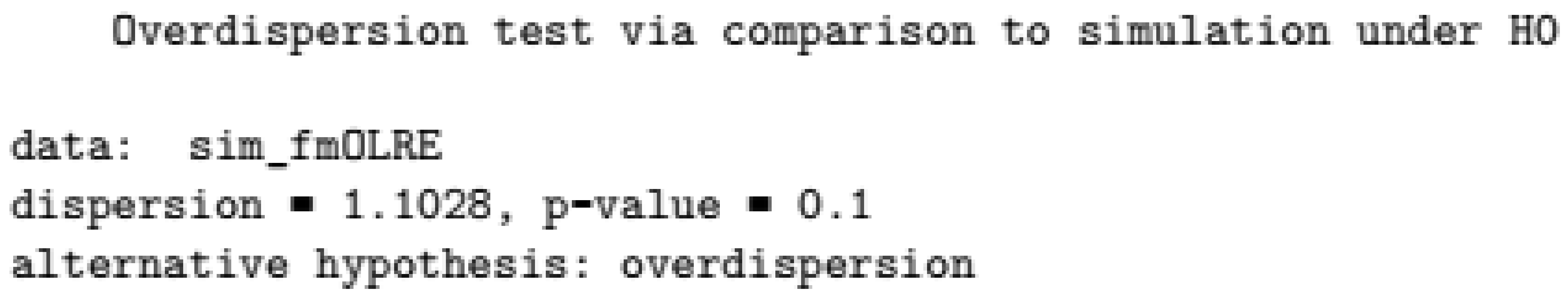
Example

Random Effect

```
#check diagnostic plot
sim_fmOLRE <- simulateResiduals(fmOLRE, refit=T, n=250) # takes a while, about 10 minutes or so
plotSimulatedResiduals(sim_fmOLRE)
testOverdispersion(sim_fmOLRE) # requires refit=T
#Compare the actual fit
AIC(fmp, fmnb, fmOLRE)
```

- Check the diagnostic plot

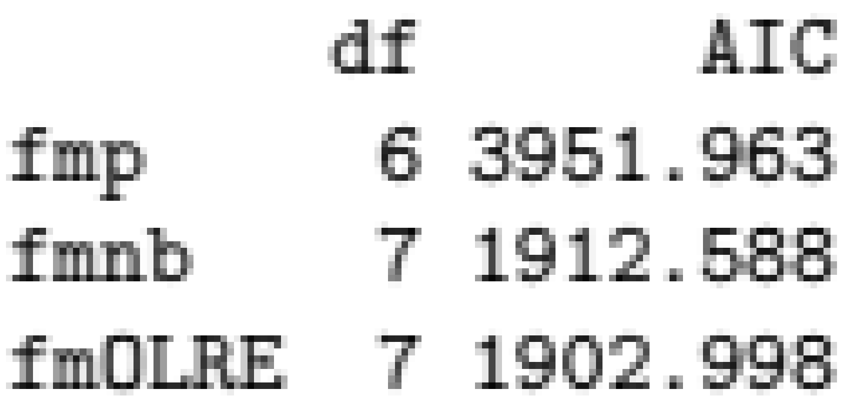




```
        Overdispersion test via comparison to simulation under H0

data:  sim_fmOLRE
dispersion = 1.1028, p-value = 0.1
alternative hypothesis: overdispersion
```

- The QQ-plot looks great,
- But the residual-predicted-plot is miserable. This may be due to a misspecified model (e.g. missing important predictors),
- Leading to underfitting of the high values (all large values have high quantiles, indicating that these residuals (O-E) are all positive and large.

Comparison



```
       df      AIC
fmp     6 3951.963
fmnb    7 1912.588
fmOLRE  7 1902.998
```

4.2 Smoothing and nonparametric regression

4.2.1 Kernel density estimation

4.2.2 Splines and penalised splines

4.2.3 Generalised additive models

4.2.4 Linear smoothing

References:
http://halweb.uc3m.es/esp/Personal/personas/durban/esp/web/notes/nonparam.pdf

4.3 Dependence

The standard GLM assumes that the observations are uncorrelated. Extensions have been developed to allow for **correlation between observations**, as occurs for example in **longitudinal studies** and **clustered** designs:

- **Generalized estimating equations (GEEs)** allow for the correlation between observations **without the use of an explicit probability model** for the origin of the correlations, so there is **no explicit likelihood**.

- **Generalized linear mixed models (GLMMs)** are an extension to GLMs that **includes random effects in the linear predictor, giving an explicit probability** model that **explains the origin of the correlations**.

4.3 Dependence

Generalized estimating equations (GEEs)

- Suitable when the random effects and their variances are not of inherent interest, as they allow for the correlation without explaining its origin.
- The focus is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more components of X on a given individual.
- GEEs are usually used in conjunction with Huber–White standard errors.[7][8]

4.3 Dependence

Generalized estimating equations (GEEs)

- Models for repeated categorical response data, and thus generalize models for matched pairs.
- GEE approach is an extension of GLMs. It provides a semi-parametric approach to longitudinal analysis of categorical response; it can be also used for continuous measurements.
- GEE's were first introduced by Liang and Zeger (1986); see also Diggle, Liang and Zeger, (1994).
- The gist of GEE is instead of attempting to model the within-subject covariance structure, to treat it as a nuisance and simply model the mean response.
- The covariance structure doesn't need to be specified correctly for us to get reasonable estimates of regression coefficients and standard errors.

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Variables:

- A response variable Y can be either continuous or categorical. We will focus on categorical.

  $Y = (Y_{ij})$ Each y_i can be, for example, a binomial or multinomial response.

  $X = (X_1, X_2, \ldots, X_k)$ be a set of explanatory variables which can be discrete, continuous, or a combination. X_i is $n_i \times k$ matrix of covariates.

Model:

- Its form is like GLM, but full specification of the joint distribution not required, and thus no likelihood function: $g(\mu_i) = x_i^T \beta$

Random component:

- Any distribution of the response that we can use for GLM, e.g., binomial, multinomial, normal, etc.

Systematic component:

- A linear predictor of any combination of continuous and discrete variables.

Link function:

- Can be any $g(\mu_i)$, e.g., identity, log, logit, etc.

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Covariance structure:

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as in the case of independent responses.

The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated responses must also be specified and modeled.

Assumptions:

- The responses are are correlated or clustered, i.e., cases are NOT independent.
- Covariates can be the power terms or some other nonlinear transformations of the original independent variables, can have interaction terms.
- The homogeneity of variance does NOT need to be satisfied
- Errors are correlated
- It uses quasi-likelhood estimation rather than maximum likelihood estimation (MLE) or ordinary least squares(OLS) to estimate the parameters, but at times these will coincide.
- Covariance specification. These are typically four or more correlation structures that we assume a priori.

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Covariance structure:

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as in the case of independent responses.

The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated responses must also be specified and modeled.

Assumptions:

- The responses are are correlated or clustered, i.e., cases are NOT independent.
- Covariates can be the power terms or some other nonlinear transformations of the original independent variables, can have interaction terms.
- The homogeneity of variance does NOT need to be satisfied
- Errors are correlated
- It uses quasi-likelhood estimation rather than maximum likelihood estimation (MLE) or ordinary least squares(OLS) to estimate the parameters, but at times these will coincide.
- Covariance specification. These are typically four or more correlation structures that we assume a priori.

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

There are basically four **correlation** structures:

- Independence –

  (correlation between time points is independent)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Exchangable (or Compund Symmetry)

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

- AutoRegressive Order 1 (AR 1)

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

- Unstructured

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

$\rho_{ij} = corr\left(Y_{ij}, Y_{ik}\right)$ for the *i*th subject at times *j* and *k*.

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Parameter Estimation:

The quasi-likelihood estimators are estimates of quasi-likelihood equations which are called generalized estimating equations.

A quasi-likelihood estimate of β arise from maximization of normality-based loglikelihood **without assuming** that the response is normally distributed.

Recall, that we briefly discussed quasi-likelihood when we introduced overdispersion.

In general, there are no closed-form solutions, so the GEE estimates are obtained by using an **iterative algorithm**, that is **iterative quasi-scoring procedure**.

GEE estimates of model parameters are valid even if the covariance is mis-specified (because they depend on the first moment, e.g., mean).

However, if the **correlation structure is mis-specified**, the standard errors are not good, and some adjustments based on the data(empirical adjustment) are needed to get more appropriate standard errors.

Agresti (2013) points out that a chosen model in practice is never exactly correct, but choosing carefully a working correlation(covariance structure) can help with efficiency of the estimates.

(*Lenz, S.T. Alan Agresti (2013): Categorical data analysis. Stat Papers 57, 849–850 (2016). https://doi.org/10.1007/s00362-015-0733-8*)

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Interpretation of Parameter Estimates:

The interpretation will depend on the chosen link function.

For example, if it is logit, $\exp(\beta_0) =$ the odds that the characteristic is present in an observation *i* when $X_i = 0$, but if it is identity, $\exp(\beta_0) =$ is the value of the response when $X_i = 0$.

Inferences:

- Wald statistics based confidence intervals and hypothesis testing for parameters.
- Recall they rely on asymptotic normality of estimator and their estimated covariance matrix.

Model Fit:

- We do not test for the model fit of the GEE, because this is really an estimating procedure; there is no likelihood function.
- We look at the **empirical estimates** of the standard errors and the covariance.
- **Compare** the empirical estimates with the model-based estimates.
- For model based output, we can still use overall goodness-of-fit statistics: Pearson chi-square statistic, Deviance, Likelihood ratio test, Hosmer-Lemeshow test and statistic, and Residual analysis: Pearson, deviance, adjusted residuals, etc...
- Overdispersion

4.3 Dependence

Objective:
Fit a model to repeated categorical responses, that is correlated and clustered responses, by GEE methodology.

Advantages:

- Computationally simpler than MLE for categorical data
- Does not require multivariate distribution
- Consistent estimation even with mis-specified correlation structure

Limitations:

- There is no likelihood function since the GEE does not specify completely the joint distribution; thus some do not consider it a model but just a method of estimation.
- Likelihood-based methods are NOT available for testing fit, comparing models, and conducting inferences about parameters.
- Empirical based standard errors underestimate the true ones, unless very large sample size.
- Empirical estimators more variable than the parametric ones.

Example

Longitudinal data analysis

Longitudinal data,

- also called panel data,
- is data that is collected **through a series of repeated observations** of the **same subjects over some extended time** frame.

For example,

- in social-personality and clinical psychology, to study rapid fluctuations in behaviors, thoughts, and emotions from moment to moment or day to day;
- in developmental psychology, to study developmental trends across the life span; and
- in sociology, to study life events throughout lifetimes or generations; and
- in consumer research and political polling to study consumer trends.

Longitudinal studies can be **retrospective** (looking back in time, thus using existing data such as medical records or claims database) or **prospective** (requiring the collection of new data).

Cohort studies are one type of longitudinal study which sample a cohort (a group of people who share a defining characteristic, typically who experienced a common event in a selected period, such as birth or graduation) and perform cross-section observations at intervals through time.

Example 1

Longitudinal data analysis

Data analyzed by Hedeker and Gibbons (1997). A randomized trial for schizophrenia where:

- 312 patients received drug therapy; 101 received placebo
- measurements at weeks 0, 1, 3, 6, but some subjects have missing data due to dropout
- outcome: severity of illness (1 = normal, ... , 7 = extremely ill)

At baseline (week0), the two groups have very similar averages. This makes sense. In a randomized trial, the groups are initially just a random division of the subjects; there should be no "treatment" effect because the treatment hasn't yet started. If there were a difference at baseline, it would lead us to believe that the randomization was not carried out properly.

Example 1

Longitudinal data analysis

Fit a model for mean response with

- an intercept,
- a main effect for group,
- a main effect for square root(week), and
- an interaction between group and square root(week).

This allows the two groups to have different intercepts and slopes. Because the intercepts are defined as the average responses at week 0, we expect that the main effect for group (i.e., the difference in intercepts will be small).

How can we fit this model, considering the fact that the multiple observations for a subject are **correlated**?

Data: The data from the schizophrenia trial (schiz.dat).

Column 1: subject ID

Column 2: group (0 = placebo, 1 = drug)

Column 3: week (0, 1, 3, 6)

Column 4: severity of illness (1, ... ,7)

Example 2

Growth curves of pigs-The dietox data frame has 861 rows and 7 columns.

- Weight Weight in Kg
- Feed Cumulated feed intake in Kg
- Time Time (in weeks) in the experiment
- Pig Factor; id of each pig
- Evit Factor; vitamin E dose; see ’details’.
- Cu Factor, copper dose; see ’details’
- Start weight in experiment, i.e. weight at week 1.
- Litter Factor, id of litter of each pig

Data contains weight of slaughter pigs measured weekly for 12 weeks. Data also contains the startweight (i.e. the weight at week 1). The treatments are 3 different levels of Evit = vitamin E (dose: 0, 100, 200 mg dl-alpha-tocopheryl acetat /kg feed) in combination with 3 different levels of Cu=copper (dose: 0, 35, 175 mg/kg feed) in the feed. The cumulated feed intake is also recorded. The pigs are littermates.

Example 2

Growth curves of pigs -The dietox data frame has 861 rows and 7 columns.

```
Install.package(“geepack)
library(geepack)
#Example 2
data(dietox)
head(dietox)
## Not run:
if (require(ggplot2)){
  qplot(Time, Weight, data=dietox, col=Pig) +
geom_line() +
    theme(legend.position = "none") +
facet_grid(Evit~Cu)
} else {
  coplot(Weight ~ Time | Evit * Cu, data=dietox)
}
## End(Not run)
```

```
data(dietox)
dietox$Cu <- as.factor(dietox$Cu)
mf <- formula(Weight ~ Cu * (Time + I(Time^2) + I(Time^3)))
gee1 <- geeglm(mf, data=dietox, id=Pig,
family=poisson("identity"), corstr="ar1")
gee1
coef(gee1)
vcov(gee1)
summary(gee1)
coef(summary(gee1))
mf2 <- formula(Weight ~ Cu * Time + I(Time^2) + I(Time^3))
gee2 <- geeglm(mf2, data=dietox, id=Pig,
family=poisson("identity"), corstr="ar1")
anova(gee2)
```

Example 3

Epiliptic Seizures data

- Description - The seizure data frame has 59 rows and 7 columns. The dataset has the number of epiliptic seizures in each of four two-week intervals, and in a baseline eight-week inverval, for treatment and control groups with a total of 59 individuals.
- Usage – seizure
- Format -This data frame contains the following columns:

  y1 - the number of epiliptic seizures in the 1st 2-week interval

  y2 - the number of epiliptic seizures in the 2nd 2-week interval

  y3 - the number of epiliptic seizures in the 3rd 2-week interval

  y4 - the number of epiliptic seizures in the 4th 2-week interval

  trt - an indicator of treatment

  base - the number of epilitic seizures in a baseline 8-week interval

  age - a numeric vector of subject age

Example 4

Clustered Data

- The respdis data frame has 111 rows and 3 columns.
- The study described in Miller et. al. (1993) is a randomized clinical trial of a new treatment of respiratory disorder.
- The study was conducted in 111 patients who were randomly assigned to one of two treatments (active, placebo).
- At each of four visits during the follow-up period, the response status of each patients was classified on an ordinal scale.

```
data(respdis)
resp.l <- reshape(respdis, varying =list(c("y1", "y2", "y3",
"y4")),
v.names = "resp", direction = "long")
resp.l <- resp.l[order(resp.l$id, resp.l$time),]
fit <- ordgee(ordered(resp) ~ trt, id=id, data=resp.l,
int.const=FALSE)
summary(fit)
data(ohio)
ohio$resp <- ordered(as.factor(ohio$resp))
fit <- ordgee(resp ~ age + smoke + age:smoke, id = id,
data=ohio)
summary(fit)
```

4.4 Random effect and Mixed Models

Generalized linear mixed models (GLMMs)

- The resulting "subject-specific" parameter estimates are suitable when the focus is on estimating the effect of changing one or more components of X on a given individual.
- GLMMs are also referred to as multilevel models and as mixed model.
- In general, fitting GLMMs is more computationally complex and intensive than fitting GEEs.

4.4 Random effect and Mixed Models

Generalized linear mixed models (GLMMs)

- Mixed effects models are useful when we have data with more than one source of random variability.

- For example, an outcome may be measured more than once on the same person (repeated measures taken over time).

- We have to account for both **within-person** and **across-person variability**. A single measure of residual variance cannot account for both.

- Or maybe multiple fields each contain multiple plots with different varieties of produce to be compared, but the fields themselves are treated differently—we would need to account for the field-to-field (cluster) variability as well as the plot-to-plot variability.

4.4 Random effect and Mixed Models

Goal of GLMMs

- What is the relationship between a particular independent variable and the expected outcome?
- You can adjust for relationships of other variables with the outcome if they are important (covariates)
- Does one particular independent variable change the relationship of another particular independent variable with the expected outcome? (Or phrased slightly differently, is there an interaction of two variables?)
- What is the "best" combination of independent variables for estimating the expected outcome?
- For a given set of values of independent variables, what is the estimated expected outcome?

  A variant of this: what is the most likely predicted value of the outcome variable itself, and how likely is it to be equal or close to that value?

Example 1

- Comparing the economic growth over 5 decades between Rural and Metropolitan counties.
- Economic growth is the outcome, measured in thousands of jobs (JobsK). JobsK is continuous.
- County indicates from which county the observations come. Each county has up to 5 measurements, and this is why we need the mixed model—to account for the inherent correlation among the multiple observations from the same county. County is categorical.
- Time indicates number of decades since 1960, and ranges from 0 to 4. Treated as continuous.
- And Rural is an indicator (aka dummy) variable for whether the county is rural. Rural is categorical.

Summary

✓Understand the overdispersion and solve the problem.

Other modeling relevant to categorical data are

✓Latent Class Models

✓Structural Equation Modeling

✓General Estimating Equations (GEE) – semiparametric methods for modeling longitudinal data

✓Nonlinear Mixed Effects Model (NLME) – a parametric alternative to GEE

✓Bayesian Modelling

✓etc... (there are many more types of models!)

Thank You!