

i Front page

Department of Mathematical Sciences

Examination paper for **TMA4315 Generalized linear models**

Academic contact during examination: Mette Langaas

Phone: 988 47 649

Examination date: 19 December 2018

Examination time (from-to): 15:00 - 19:00

Permitted examination support material: C.

- *Tabeller og formler i statistikk* (Tapir forlag, Fagbokforlaget),
- one yellow A5 sheet with your own handwritten notes (stamped by the Department of Mathematical Sciences),
- specified calculator.

Other information:

- All answers must be justified, and relevant calculations provided.
- The exam questions are only available in English since this is a course at master's level given in English.
- For each problem the maximum possible score is noted.
- This exam is given in Inspira, and third party software (R and RStudio) is available for the exam.
- For each problem you may write your answer into Inspira, or use the provided paper sheets. Remember to correctly specify the code for each problem (available in lower right corner of the problem) on your paper sheet, so that the scanned sheets will be matched with the correct problem. It is advisable that you write down the code when starting to write on the sheet.

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

1 Inverse Gaussian distribution

[15 points]

Consider a random variable Y . In our course we have considered the univariate exponential family having distribution (probability density function for continuous variables and probability mass function for discrete variables)

$$f(\mathbf{y}) = \exp\left(\frac{\mathbf{y}\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi} \mathbf{w} + c(\mathbf{y}, \phi, \mathbf{w})\right)$$

where $\boldsymbol{\theta}$ is called the *natural parameter* (or parameter of interest) and ϕ the *dispersion parameter*. In this problem we will let $\mathbf{w} = \mathbf{1}$.

The inverse Gaussian distribution is a continuous distribution, and the probability density function can be written as

$$f(y) = \frac{1}{\sqrt{2\pi y^3 \sigma}} \exp\left[-\frac{1}{2y} \left(\frac{y - \mu}{\mu\sigma}\right)^2\right], \text{ for } y > 0.$$










Remark: our "Tabeller og formler i statistikk" uses a slightly different version with $\lambda = 1/\sigma^2$.

1. Show that the inverse Gaussian distribution is an univariate exponential family, and specify what the elements of the exponential family $(\boldsymbol{\theta}, \phi, b(\boldsymbol{\theta}), c(\mathbf{y}, \phi))$ are. (Remember we have set $\mathbf{w} = \mathbf{1}$, so $c(\mathbf{y}, \phi, \mathbf{w}) = c(\mathbf{y}, \phi)$ here.)
2. What are the connections between $E(\mathbf{Y})$ and $\text{Var}(\mathbf{Y})$ and elements of the exponential family?
3. Use these connections to derive the mean and variance for the inverse Gaussian distribution.
4. What constraints should be put on the parameters involved?
5. If the inverse Gaussian distribution is used as the distribution for the response in a generalized

linear model, what is the *canonical link* function?

6. What advantages is there in using a canonical link function?

Fill in your answers here, and/or use the paper sheets provided.

Format | **B** | *I* | U | x_2 | x^2 | \int_x |  |  |  |  |  |  | Ω |  |  | Σ | 

Words: 0

Maximum marks: 15

2 The generalized linear model

[20 points]

This course is called *generalized linear models* (GLM) and this model constitutes the core of the course. Write a short text describing and explaining the following.

- The three model components for the GLM, and the connection to the exponential family of distributions.
- The role and general formulas for the likelihood, score function and expected Fisher information matrix in parameter estimation for GLMs. Remark: You do not need to derive the formulas, but all notation needs to be well explained. Insightful explanations are rewarded.
- The role of the deviance for model assessment and model choice for GLM.

If you want you may use one of the data sets from this exam, or an example you come up with yourself, as a running thread in your short text.

Fill in your answers here, and/or use the paper sheets provided.

Format | **B** | *I* | U | x_2 | x^2 | \int_x | | | | | | | Ω | | | Σ | ABC |

Words: 0

Maximum marks: 20

3 Lung cancers in Danish cities - model1

[10 points]

Words underlined are names of objects/data sets/packages/functions in R (what would have been written with typewriter font if that had been available in Inspira).

We will look at a data set giving the number of new cases of lung cancer in four Danish cities in the period 1968-1971. The data set has 24 observations of the following four variables.

- Cases: the number of lung cancer cases.
- Pop: the population of each age group in each city.
- Age: the age group, a factor with the six levels Age40-54, Age55-59, Age60-64, Age65-69, Age70-74 and Age>74. Dummy variable coding is used, with Age40-54 as reference category.
- City: the city, a factor with levels Fredericia, Horsens, Kolding and Vejle. Dummy variable coding is used, with Fredericia as reference category. The names CityH for Horsens, CityK for Kolding, and CityV for Vejle, will be used in print-outs.

Since the population for each combination of city and age group differ, we would *not* like to model the number of new lung cancer cases, but instead *the rate of new lung cancers*. The rate is given as the number of new lung cancer cases per unit of population. In Figure 1 (in the pdf-file) is plot of this lung cancer rate (that is, Cases/Pop) against the age group, for each of the cities.

We have fitted a Poisson regression with log link to Cases as response, with log(Pop) (natural log) as offset, and with Age, City and the interaction between Age and City as covariates, see the R print-out in the pdf-file for model summary. We refer to this as model1.

Let Y_i be the observed number of new lung cancer cases for observation i , for $i = 1, \dots, 24$ (all combinations of Age and City), and let the corresponding population be denoted t_i . Further, let $\lambda_i = \mathbf{E}(Y_i)$, and $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ be the linear predictor (for some choice of covariates \mathbf{x} and corresponding regression parameters $\boldsymbol{\beta}$).

Answer the following questions:

1. What is the connection between λ_i , t_i and η_i for our Poisson regression with log link and log(Pop) as offset?
2. In the print-out from R in the pdf-file the estimate for the intercept is -5.63 . Explain what this number means, and relate it to the rate of new lung cancer cases.

3. Perform a test to investigate if the regression coefficients for Age55-59 and Age60-64 are equal, and report the p -value, see R print-out in the pdf-file for needed numerical values.
4. Explain why the (residual) deviance in the print-out in the pdf-file is practically 0 and has 0 degrees of freedom.

Fill in your answers here, and/or use the paper sheets provided.

Format | **B** | *I* | U | x_2 | x^2 | \int_x | | | | | | | Ω | | | Σ | ABC |

Words: 0

Maximum marks: 10

4 Lung cancers in Danish cities - models2-4

[10 points]

This is a continuation of Problem 3. Based on the model fit in Problem 3 for model1 we proceed to look at models without City as covariate, but with different codings of age.

Let AgeNum be a numerical version for age, where the lower limit of the age group is used. This means that we have six unique values for AgeNum: 40, 55, 60, 65, 70 and 74.

We consider the following three models:

- model2: Age as factor
- model3: AgeNum as linear term
- model4: AgeNum as linear and quadratic term

See the pdf-file for R code and results.

Answer the following questions:

1. Using the AIC as a measure for model choice, which model would you prefer?
2. Based on likelihood ratio tests performed in the R print-out in the pdf-file, which model would you prefer? Specify which hypothesis test(s) you are investigating, and which of the results in the R print-out in the pdf-file you are using.
3. For simplicity we now use model2. For another Danish city we consider individuals in the Age55-59 year group and observe that the population size is 1000. Estimate the expected number of new lung cancer cases (for 1968-1971). Also give a 95% confidence interval for this quantity. Numerical values needed for these calculations are in the R print-out in the pdf-file.

Fill in your answers here, and/or use the paper sheets provided.

Format | **B** | *I* | U | x_2 | x^2 | \int_x | | | | | | | Ω | | | Σ | ABC |

Words: 0

Maximum marks: 10

5 GLM data analysis in R: low birth weight

[15 points]

If you experience problems with the R or RStudio installation, or the MASS package is not available on your computer, you may answer Problem 7 instead of Problem 5. Words underlined are names of objects/data sets/packages/functions in R (what would have been written with typewriter font if that had been available in Inspera).

We will study a data set on risk factors associated with low infant birth weight birthwt in the package MASS. You may access the data set by first loading the MASS package, library(MASS) and then writing birthwt. The data set consists of 189 observations of 10 variables, and we use the following four of the variables:

- low: indicator of low birth weight for infant ('1'=birth weight below 2.5 kg, '0'=birth weight of 2.5 kg or above)
- smoke: mother's smoking status during pregnancy ('0'=non-smoker, '1'=smoker)
- ht: history of hypertension for mother, ('1'=history of hypertension, '0'=no hypertension)
- lwt: weight of mother at last menstrual period, numerical value in lbs.

Our aim is to model the probability of low birthweight for the infant. Fit a generalized linear model with canonical link to the data. Use low as the response and smoke, ht and lwt as covariates. You may also need to perform additional commands (in addition to fitting the model) on your fitted GLM-object.

Based on your data analyses answer the following questions.

i) Which type of GLM did you fit?

- normal
- binomial
- Poisson
- gamma
- Other

ii) Which link function did you use?

- identity
- log
- logit
- exp
- inverse
- negative inverse
- other

iii) How would you explain the meaning of the estimated effect of smoke β_{smoke} ?

When we compare two women where one smoked during pregnancy and one did not smoke during pregnancy (but their other covariates were the same), then

- the probability of getting a low weight infant increased by β_{smoke} for the woman who smoked
- the odds of getting a low weight infant was multiplied by β_{smoke} for the woman who smoked
- the odds of getting a low weight infant was multiplied by $\exp(\beta_{smoke})$ for the woman who smoked
- the probability of getting a low weight infant increased by $\exp(\beta_{smoke})$ for the woman who smoked

iv) Consider a mother who smoked during pregnancy, had a history of hypertension and had weight 130 (lbs) at the last menstrual period. What is the predicted probability that the infant will be low weight? (Write your answer with two significant digits, e.g. 0.0045 or 0.34.)

v) Perform a likelihood ratio test to compare the fitted model to the model where lwt is not included as a covariate (so, only ht and smoke as covariates). Report the p-value from the test. (Write your answer with two significant digits, e.g. 0.0045 or 0.34.)

vi) The conclusion from the likelihood ratio test is that the model with lwd is the best. True or false?

- True
- False

Maximum marks: 15

6 Rcode for Problem 5 (optional)

If you want you may copy your R-code from the analyses of the low infant birth weight problem here. This is optional.

Fill in your answer here

Format | **B** | *I* | U | x_2 | x^2 | \int_x | | | | | | | Ω | | | Σ | ABC |

Words: 0

Maximum marks: 0

7 GLM analysis in R back-up question: lime trees

[15 points]

If you experience problems with the R or RStudio installation, or the MASS package is not available on your computer, you may answer Problem 7 *instead of* Problem 5. If you answer Problem 5 you shall not answer Problem 7 and vice versa. Words underlined are names of objects/data sets/packages/functions in R (what would have been written with typewriter font if that had been available in Inspira).

We will analyse a data set with measurements on small-leaved lime trees grown in Russia, and the aim is to model the foliage biomass. The data set consists of 385 observations, and we will look at the following variables:

- Foliage: the foliage biomass, in kg (oven dried matter).
- DBH: the tree diameter, at breast height, in cm.
- Origin: the origin of the tree; one of Coppice, Natural, Planted. Dummy variable coding is used with Coppice as reference category.

Print-out from fitting Foliage as response and $\log(\text{DBH})$ (natural log), Origin and the interaction thereof as covariates are presented in the pdf-file shown, along with additional analyses. Based on the print-out, answer the following questions.

i) Why are you answering Problem 7 and not Problem 5?

- R failed
- RStudio failed
- MASS not installed
- Other

Which type of GLM was fitted?

- normal
- binomial
- Poisson
- gamma
- other

iii) What is the canonical link for this type of response?

- log
- logit
- exp
- inverse
- negative inverse
- other

iv) We look at a tree from Coppice. If the $\log(\text{DBH})$ of the tree increases by 1 cm, what is the multiplicative change in the estimated mean foliage biomass? (Write your answer with two significant digits, e.g. 0.0045 or 0.34.)

v) A deviance test is performed in the print-out, and a p-value of 0.99996 is reported. The conclusion from the deviance test is that the model is good. True or false.

- True
- False

vi) A likelihood ratio test is performed to compare this model to the model where the interaction between $\log(\text{DBH})$ and Origin is not included in the linear predictor. Report the p-value from the test. (Write your answer with two significant digits, e.g. 0.0045 or 0.34.)

vii) For the likelihood ratio test, which of the two model are preferred? Use a significance level of 0.05.

- Origin*log(DHB)
- Origin+log(DHB)

Maximum marks: 15

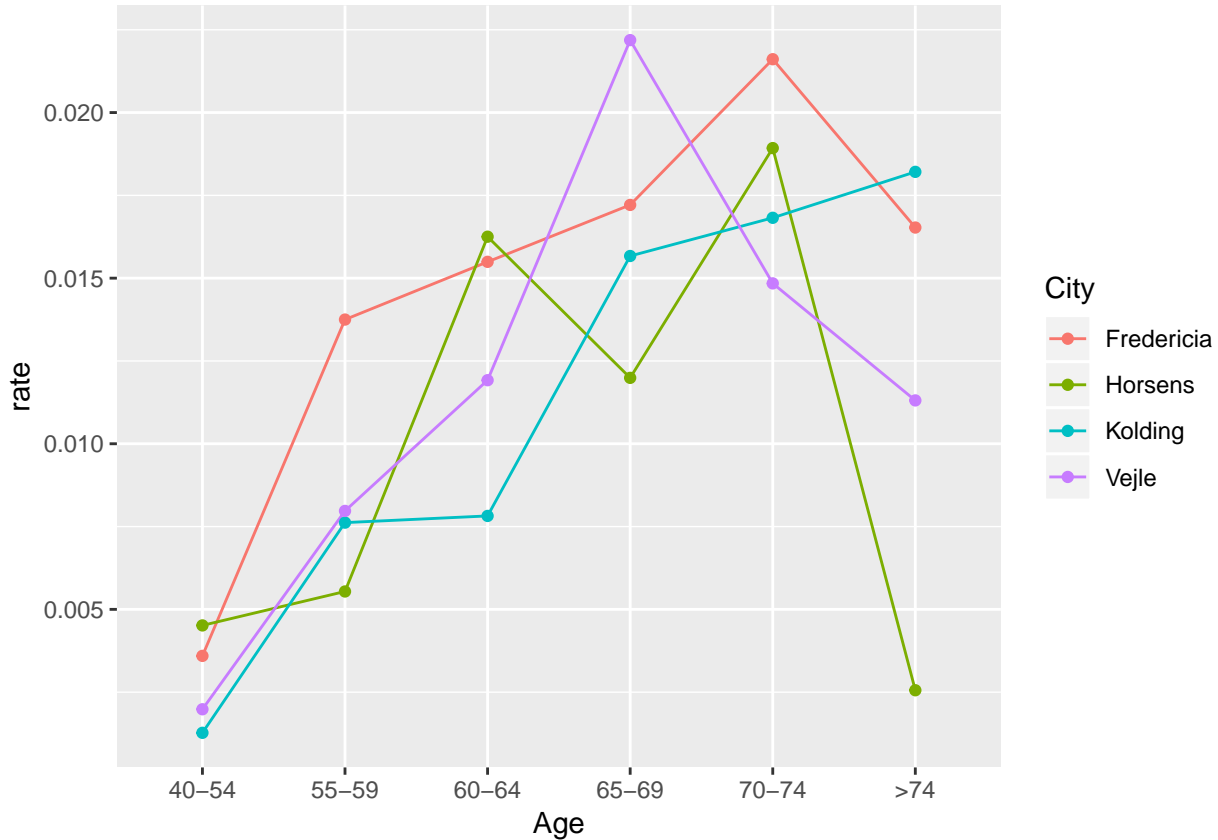
Question 3
Attached



TMA4315 GLM - Exam December 2018: Problem 3

Figure 1

Below you see a plot of the lung cancer rate (that is, **Cases/Pop**) against the age group, for each of the four cities.



R print-out

Print-out from fitting `model1`, and subsequent calculations to be used in Problem 4.

```
model1 <- glm( Cases ~ offset(log(Pop)) + City * Age, family=poisson, data=danishlc)
summary(model1)
```

```
##
## Call:
## glm(formula = Cases ~ offset(log(Pop)) + City * Age, family = poisson,
##      data = danishlc)
##
## Deviance Residuals:
## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [24] 0
##
## Coefficients:
```

```

##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.62795    0.30151 -18.666 < 2e-16 ***
## CityH           0.22770    0.40967   0.556 0.578343
## CityK          -1.03837    0.58387  -1.778 0.075335 .
## CityV          -0.59463    0.53936  -1.102 0.270257
## Age55-59       1.34123    0.42640   3.145 0.001658 **
## Age60-64       1.46058    0.42640   3.425 0.000614 ***
## Age65-69       1.56578    0.43693   3.584 0.000339 ***
## Age70-74       1.79340    0.42640   4.206 2.6e-05 ***
## Age>74         1.52530    0.43693   3.491 0.000481 ***
## CityH:Age55-59 -1.13671    0.65223  -1.743 0.081368 .
## CityK:Age55-59  0.44798    0.74620   0.600 0.548271
## CityV:Age55-59  0.04961    0.72434   0.068 0.945398
## CityH:Age60-64 -0.17991    0.57045  -0.315 0.752471
## CityK:Age60-64  0.35483    0.75807   0.468 0.639737
## CityV:Age60-64  0.33237    0.69413   0.479 0.632058
## CityH:Age65-69 -0.58918    0.60649  -0.971 0.331320
## CityK:Age65-69  0.94450    0.72926   1.295 0.195268
## CityV:Age65-69  0.84855    0.67995   1.248 0.212051
## CityH:Age70-74 -0.36029    0.58487  -0.616 0.537886
## CityK:Age70-74  0.78788    0.73684   1.069 0.284945
## CityV:Age70-74  0.21891    0.71191   0.307 0.758469
## CityH:Age>74   -2.09376    0.87626  -2.389 0.016874 *
## CityK:Age>74    1.13520    0.72405   1.568 0.116915
## CityV:Age>74    0.21508    0.73059   0.294 0.768463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance:  1.2991e+02 on 23 degrees of freedom
## Residual deviance: -2.6645e-15 on 0 degrees of freedom
## AIC: 144.39
##
## Number of Fisher Scoring iterations: 3
exp(model1$coefficients[1])

## (Intercept)
## 0.003595946
model1$coefficients[5:6]

## Age55-59 Age60-64
## 1.341232 1.460578
vcov(model1)[5:6,5:6]

##           Age55-59  Age60-64
## Age55-59 0.18181818 0.09090909
## Age60-64 0.09090909 0.18181818
11*vcov(model1)[5:6,5:6]

##           Age55-59 Age60-64
## Age55-59         2         1
## Age60-64         1         2

```

Question 4
Attached



TMA4315 GLM - Exam December 2018: Problem 4

```
model2 <- glm(Cases~offset(log(Pop))+Age, family=poisson, data=danishlc)
model3 <- glm(Cases~offset(log(Pop))+AgeNum, family=poisson, data=danishlc)
model4 <- glm(Cases~offset(log(Pop))+AgeNum+I(AgeNum^2), family=poisson, data=danishlc)
```

```
AIC(model2,model3,model4)
```

```
##          df          AIC
## model2   6 136.6946
## model3   2 149.3556
## model4   3 134.8876
```

```
anova(model2,model3,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(log(Pop)) + Age
## Model 2: Cases ~ offset(log(Pop)) + AgeNum
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         18     28.307
## 2         22     48.968 -4  -20.661 0.0003696 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2,model4,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(log(Pop)) + Age
## Model 2: Cases ~ offset(log(Pop)) + AgeNum + I(AgeNum^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         18     28.307
## 2         21     32.500 -3  -4.1931  0.2414
```

```
anova(model3,model4,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(log(Pop)) + AgeNum
## Model 2: Cases ~ offset(log(Pop)) + AgeNum + I(AgeNum^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         22     48.968
## 2         21     32.500  1  16.468 4.948e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coefficients(model2)
```

```
## (Intercept)   Age55-59   Age60-64   Age65-69   Age70-74   Age>74
## -5.862253    1.082342    1.501676    1.750287    1.847222    1.408281
```

```
vcov(model2)
```

```
##          (Intercept)   Age55-59   Age60-64   Age65-69   Age70-74
## (Intercept)  0.03030303 -0.03030303 -0.03030303 -0.03030303 -0.03030303
## Age55-59     -0.03030303  0.06155303  0.03030303  0.03030303  0.03030303
```

```

## Age60-64 -0.03030303 0.03030303 0.05355884 0.03030303 0.03030303
## Age65-69 -0.03030303 0.03030303 0.03030303 0.05252525 0.03030303
## Age70-74 -0.03030303 0.03030303 0.03030303 0.03030303 0.05530303
## Age>74 -0.03030303 0.03030303 0.03030303 0.03030303 0.03030303
## Age>74
## (Intercept) -0.03030303
## Age55-59 0.03030303
## Age60-64 0.03030303
## Age65-69 0.03030303
## Age70-74 0.03030303
## Age>74 0.06256109

```

33*vcov(model2)

```

## (Intercept) Age55-59 Age60-64 Age65-69 Age70-74 Age>74
## (Intercept) 1 -1.00000 -1.000000 -1.000000 -1.000 -1.000000
## Age55-59 -1 2.03125 1.000000 1.000000 1.000 1.000000
## Age60-64 -1 1.00000 1.767442 1.000000 1.000 1.000000
## Age65-69 -1 1.00000 1.000000 1.733333 1.000 1.000000
## Age70-74 -1 1.00000 1.000000 1.000000 1.825 1.000000
## Age>74 -1 1.00000 1.000000 1.000000 1.000 2.064516

```

Question 7
Attached



TMA4315 GLM - Exam December 2018: Problem 7

```
library(GLMsData)
data(lime)
modell1 <- glm(Foliage ~ Origin * log(DBH),
              family=Gamma(link="log"), data=lime)
summary(modell1)

##
## Call:
## glm(formula = Foliage ~ Origin * log(DBH), family = Gamma(link = "log"),
##      data = lime)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7480  -0.5354  -0.1509   0.2528   3.2938
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.6289     0.2756 -16.793 < 2e-16 ***
## OriginNatural     0.3245     0.3882   0.836  0.40371
## OriginPlanted   -1.5285     0.5727  -2.669  0.00793 **
## log(DBH)         1.8432     0.1016  18.149 < 2e-16 ***
## OriginNatural:log(DBH) -0.2040     0.1433  -1.424  0.15536
## OriginPlanted:log(DBH)  0.5768     0.2093   2.755  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.5443774)
##
##      Null deviance: 508.48  on 384  degrees of freedom
## Residual deviance: 152.69  on 379  degrees of freedom
## AIC: 750.33
##
## Number of Fisher Scoring iterations: 6
exp(modell1$coefficients)

##              (Intercept)              OriginNatural              OriginPlanted
##      0.009765246              1.383368697              0.216862659
##              log(DBH) OriginNatural:log(DBH) OriginPlanted:log(DBH)
##      6.316642306              0.815435562              1.780301334

nu1 = 1/summary(modell1)$dispersion
dev=deviance(modell1) * nu1
1-pchisq(dev,model1$df.residual)

## [1] 0.9999559

modell2=glm(Foliage ~ Origin+log(DBH),
           family=Gamma(link="log"), data=lime)
anova(modell1,model2,test="LRT")

## Analysis of Deviance Table
##
```



```

## Model 1: Foliage ~ Origin * log(DBH)
## Model 2: Foliage ~ Origin + log(DBH)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      379      152.69
## 2      381      160.58 -2   -7.8903 0.0007122 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(model1,model2)

##           df           AIC
## model1    7 750.3267
## model2    5 767.0195

```