TMA4315 Generalized linear models

Tentative solutions to exam 19.12.2018

Mette Langaas

12/31/2018

Report improvements to Mette.Langaas@ntnu.no

Problem 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)

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

where θ is called the *natural parameter* (or parameter of interest) and ϕ the *dispersion parameter*. In this problem we will let $w = 1$.

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

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

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 (θ, ϕ, $b(\theta)$, $c(y, \phi)$) are. (Remember we have set $w = 1$, so $c(y, \phi, w) = c(y, \phi)$ here.)

Answer:

$$f(y) = \frac{1}{\sqrt{2\pi y^3\sigma}} \exp[-\frac{1}{2y}(\frac{y-\mu}{\mu\sigma})^2] = \exp(-\frac{1}{2}\ln(2\pi y^3\sigma^2) \exp[-(\frac{y^2 - 2\mu y + \mu^2}{2y\mu^2\sigma^2})]$$

$$= \exp(-\frac{1}{2}\ln(2\pi y^3\sigma^2) \exp[-\frac{y\frac{1}{2\mu^2} - \frac{1}{\mu} + \frac{1}{2y}}{\sigma^2}] = \exp(-\frac{1}{2}\ln(2\pi y^3\sigma^2) \exp(-\frac{1}{2y\sigma^2}) \exp(\frac{-y\frac{1}{2\mu^2} + \frac{1}{\mu}}{\sigma^2})$$

This shows that we may define: $\theta = -\frac{1}{2\mu^2}$, $b(\theta) = -\frac{1}{\mu} = -\sqrt{-2\theta}$, $\phi = \sigma^2$, and $c(y, \phi) = \exp(-\frac{1}{2}\ln(2\pi y^3\sigma^2) - \frac{1}{2y\phi})$.

Other solutions: It is possible to put the minus and/or the 2 with σ^2 instead of in θ and $b(\theta)$.

(2) What is the connections between $\text{E}(Y)$ and $\text{Var}(Y)$ and elements of the exponential family?

Answer: For an univariate exponential family we have proven that $\text{E}(Y) = b'(\theta)$ and $\text{Var}(Y) = \frac{b''(\theta)\phi}{w}$.

(3) Use these connections to derive the mean and variance for the inverse Gaussian distribution.

Answer:

$$\text{E}(Y) = b'(\theta) = \frac{d}{d\theta}(-\sqrt{-2\theta}) = \frac{d}{d\theta}(-(-2\theta)^{-\frac{1}{2}}) = -\frac{1}{2}(-2\theta)^{-\frac{1}{2}} \cdot (-2) = \frac{1}{\sqrt{-2\theta}} = \mu$$

$$\text{Var}(Y) = \frac{b''(\theta)\phi}{w} = \frac{d}{d\theta}(\phi(-2\theta)^{-\frac{1}{2}}) = \phi(-\frac{1}{2}(-2\theta)^{-\frac{3}{2}} \cdot (-2)) = \phi(-2\theta)^{-\frac{3}{2}} = \sigma^2\mu^3$$

(4) What constrains should be put on the parameters involved?

Answer: We see that the pdf is defined for $y > 0$, and need to have $\text{E}(Y) \geq 0$. This means that since $\text{E}(Y) = \frac{1}{\sqrt{-2\theta}}$ this means that we need to assume that $\theta \leq 0$.

In addition we must have that $\text{Var}(Y) \geq 0$, which means that $\phi = \sigma^2 \geq 0$, since we have already assumed that $\theta \leq 0$.

(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?

Answer: The canonical link function is defined to be when $\eta = \theta$, so that the canonical link function is $g(\mu) = -\frac{1}{2\mu^2}$.

(6) What advantages are there in using a canonical link function.

Answer: When we have a canonical link the maths get easier, and the loglikelihood function is nice and concave. This will lead to that the expected and observed Fisher information are equal.

Problem 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:

1. The three model components for the GLM, and the connection to the exponential family of distributions.
2. 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.
3. 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.

This is a minimal solution:

1) The three model components of the GLM:

- the random component Y_i is assumed to come from an exponential family with mean μ_i
- the systematic component is defined using $\eta_i = \mathbf{x}_i^T\beta$, and we require that the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T)$ has full rank (which is p).
- finally, the link between the random and systematic component is via the *link function* $\eta_i = g(\mu_i)$ or using the inverse thereof, the response function $\mu_i = h(\eta_i)$. And, we require that the link function is one-to-one and twice differentiable.

We also require the Y_is to be independent.

The role and general formulas for the likelihood, score function and expected Fisher information matrix in parameter estimation for GLMs:

To estimate parameters in a GLM we find the maximum likelihood estimators. This is done by first setting up the likelihood and log likelihood.

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} \frac{1}{\phi}(y_i\theta_i - b(\theta_i))w_i + \sum_{i=1}^{n} c(y_i, \phi, w_i)$$

where the elements here are as specified in the start of Problem 1.

Then we differentiate the loglikelihood with respect to the unknown parameters, the θ and if present, also the ϕ. However, for an univariate exponential family it turns out that the parameters are orthogonal and we may handle the estimation of θ first, then impute parameter estimates and solve for ϕ. The derivative of the loglikelihood with respect to the regression parameters $\boldsymbol{\beta}$ is called the score function, and is a p-dimensional vector. The general form for the score function is:

$$s(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathbf{x}_i h'(\eta_i)}{\text{Var}(Y_i)} = \mathbf{X}^T \mathbf{D} \Sigma^{-1}(\mathbf{y} - \mu)$$

where $\Sigma = \text{diag}(\text{Var}(Y_i))$ and $\mathbf{D} = \text{diag}(h'(\eta_i))$ (derivative wrt η_i).

To find maximum likelihood estimators we solve the set of p equations $\mathbf{s}(\boldsymbol{\beta}) = \mathbf{0}$, and we usually do this using the Fisher scoring algorithm. At step $t+1$ the Fisher scoring algoritm is given as:

$$\beta^{(t+1)} = \beta^{(t)} + F(\beta^{(t)})^{-1} s(\beta^{(t)})$$

Insert formulas for expected Fisher information and score function.

This means that we need to compute the expected Fisher information matrix $\mathbf{F}(\boldsymbol{\beta})$. For element h, l in the $p \times p$ expected Fisher information matrix:

$$F_{[h,l]}(\beta) = \sum_{i=1}^{n} \frac{x_{ih} x_{il} (h'(\eta_i))^2}{\text{Var}(Y_i)}$$

or with matrix notation

$$F(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W} = \text{diag}(\frac{h'(\eta_i)^2}{\text{Var}(Y_i)})$. The expected Fisher information if often most easily found by the relationship

$$F(\boldsymbol{\beta}) = \text{Cov}(\mathbf{s}(\boldsymbol{\beta})).$$

In addition, since $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimator, it has an asymptotic multivariate distribution with mean $\boldsymbol{\beta}$ and covariance matrix $F^{-1}(\hat{\boldsymbol{\beta}})$. This can be used as the basis for performing inference about $\boldsymbol{\beta}$.

The role of the deviance for model assessment and model choice:

The deviance is based on the likelihood ratio statistics, and is also referred to as the scaled deviance. We use the following notation. A: the larger model (this is H_1) and B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A). The likelihood ratio statistic is defined as

$$-2\ln\lambda = -2(\ln L(\hat{\beta}_B, \tilde{\phi}_B) - \ln L(\hat{\beta}_A, \tilde{\phi}_A))$$

(so, -2 times small minus large), when a dispersion parameter is present. Under weak regularity conditions the test statistic is approximately χ^2-distributed with degrees of freedom equal the difference in the number of parameters in the large and the small model. P-values are calculated in the upper tail of the χ^2-distribution.

To compute the deviance we define model A as the saturated model and B as the candidate model. The saturated model is a model where we have a perfect fit to our data, so that the difference between observed values y_i and fitted values $\hat{y}_i$ is zero. For our candidate model assume that the estimator for μ_i is $\hat{\mu}_i$, and for the saturated model the estimator for μ_i is $\hat{y}_i$.

$$D = -2[\sum_{i=1}^{G}(l_i(\hat{\mu}_i) - l_i(\hat{y}_i))]$$

with approximate χ^2-distribution with $G - p$ degrees of freedom, if the saturated model has G parameters and the candidate model has p.

For discrete response variables we group observations together in groups of maximal size (covariate patterns or interval versions thereof). Group i has n_i observations, and there are G groups. The asymptotic distribution approximation is good if all groups have large n_i. For individual discrete data asymptotic results can not be trusted. For continuous reponse variables the approximation should be good for individual data also, but the total number of observations need to be large.

When performing model selection this can (preferably) be done based on choosing model with low AIC. But, alternatively a likelihood ratio test (LRT) can be used to choose between two *nested* models. The LRT can be performed by looking at the deviance from the two nested models, and since the saturated model part then will cancel the difference in deviance statistics can be related to a χ^2-distribution with the difference in number of estimated parameters in the two models.

Problem 3: Lung cancers in Danish cities - model1

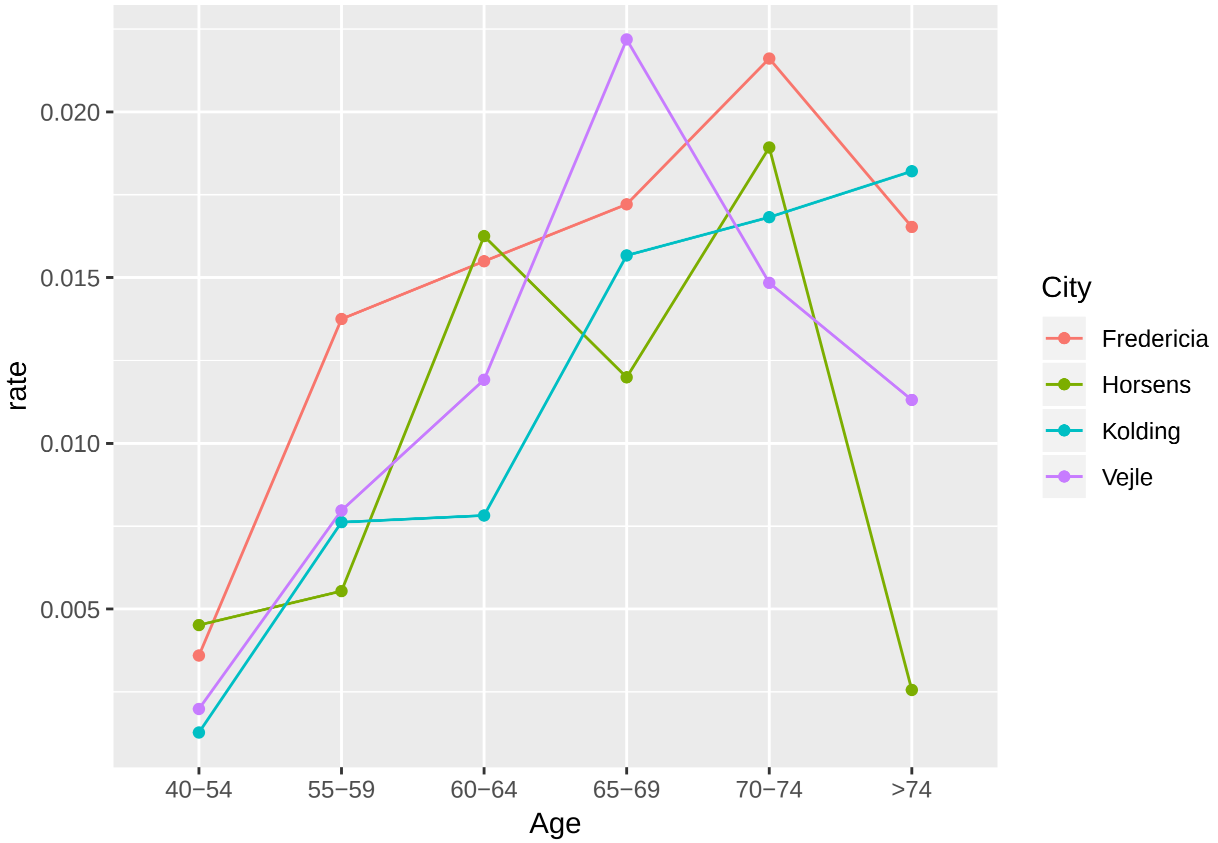
We will look at a data set giving the number of new cases of lung cancer in four Danish cities in the period 1968-1971. The data set has 24 observations of the following four variables.

- `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 $1 = 1, \ldots, 24$ (all combinations of `Age` and `City`), and let the corresponding population be denoted t_i. Further, let $\lambda_i = \mathrm{E}(Y_i)$ and $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ be the linear predictor (for some choice of covariates 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?

Answer: We have that $Y_i \sim \text{Poisson}(\lambda_i)$, with $\text{E}(Y_i) = \lambda_i$. We would not link λ_i to η_i, but instead λ_i/t_i to η_i. With log-link this is

$$\log(\frac{\lambda_i}{t_i}) = \eta_i \Leftrightarrow \log(\lambda_i) = \log(t_i) + \eta_i$$

The latter formulation is then why we talk about `log(Pop)`$= \log(t_i)$ (natural log) as offset. In addition this also gives

$$\lambda_i = t_i \exp(\eta_i)$$

(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.

Answer: $\eta_i = \beta_0$ for `city=Fredericia` and `Age=Age40-54`. This means that if we look at persons in Fredericia of age 40-54 then the estimated rate of new lung cancer cases in 1968-1971 is $\hat{\lambda}_i/t_i = \exp(\hat{\beta}_0) = \exp(-5.63) = 0.0036$.

Remark: since we have fitted a saturated model (see below) this is also equal to the ratio y_i/t_i for `city=Fredericia` and `Age=Age40-54` in the data set. This is not given in the problem, but the data is 11 new cases and population 3059, so $11/3059 = 0.0036$.

```
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  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
```

```r
exp(model1$coefficients[1])
```

```
## (Intercept)
## 0.003595946
```

(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.

Answer: Let β_5 denote the coefficient for `Age55-59` and β_6 the coefficient for `Age60-64`. The hypothesis test we want to investigate is

$$H_0 : \beta_5 = \beta_6 \text{ vs. } H_1 : \beta_5 \neq \beta_6$$

which is equivalent to testing

$$H_0 : \beta_5 - \beta_6 = 0 \text{ vs. } H_1 : \beta_5 - \beta_6 \neq 0$$

Under the null hypothesis the test statistic

$$\frac{\hat{\beta}_5 - \hat{\beta}_6}{\widehat{\text{SD}}(\hat{\beta}_5 - \hat{\beta}_6)} \approx N(0,1)$$

The $\widehat{SD}(\hat{\beta}_5 - \hat{\beta}_6)$ is found from the covariance matrix of $\hat{\boldsymbol{\beta}}$, and the R print-out gives that

$$\widehat{\text{Cov}}(\hat{\beta}_5, \hat{\beta}_6) = \begin{pmatrix} \frac{2}{11} & \frac{1}{11} \\ \frac{1}{11} & \frac{2}{11} \end{pmatrix}.$$

```
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
```

From this matrix we calculate $\widehat{\text{SD}}(\hat{\beta}_5 - \hat{\beta}_6)$ as

$$\widehat{\text{SD}}(\hat{\beta}_5 - \hat{\beta}_6) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_5) + \widehat{\text{Var}}(\hat{\beta}_6) - 2\widehat{\text{Cov}}(\hat{\beta}_5, \hat{\beta}_6)} = \sqrt{\frac{2}{11} + \frac{2}{11} - 2 \cdot \frac{1}{11}} = \sqrt{\frac{2}{11}} = \sqrt{0.18}$$

```
#alternatively using linear comb
length(model1$coefficients)#24
```

```
## [1] 24
```

```
cvec=matrix(c(rep(0,4),1,-1,rep(0,18)),nrow=1)
cvec%*%vcov(model1)%*%t(cvec)
```

```
##           [,1]
## [1,] 0.1818182
```

This gives observed test statistic:

$$z_0 = \frac{\hat{\beta}_5 - \hat{\beta}_6}{\widehat{\text{SD}}(\hat{\beta}_5 - \hat{\beta}_6)} = \frac{1.34 - 1.46}{\sqrt{0.18}} = -0.28$$

and a p-value of $2 \cdot P(Z \leq -0.28) = 0.78$.

```
z0=(model1$coefficients[5]-model1$coefficients[6])/sqrt(2/11)
z0
```

```
##   Age55-59
## -0.279893
```

```r
2*pnorm(abs(z0),lower.tail=FALSE)
```

```
##  Age55-59
## 0.7795596
```

(4): Explain why the (residual) deviance in the print-out in the pdf-file is practically 0 and has 0 degrees of freedom.

Answer: The fitted model is the saturated model, and we have a perfect fit to all observations. Thus the deviance is 0 and we have 0 degrees of freedom.

Problem 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

Answer the following questions:

(1) Using the AIC as a measure for model choice, which model would you prefer?

```r
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
```

Answer: The AIC for `model4` is the lowest, followed by `model2`. Since `model4` is a smaller model (less parameters estimated) than `model2` and the AIC is smaller, we choose `model4`.

(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.

```r
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
```

```r
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
```

```r
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
```

Answer: The likelihood ratio test (LRT) can only be performed for nested models. For our three models only `model3` is nested within `model4`, so these two can be compared. We may not use the LRT to compare `model2` with either of `model3` and `model4`. (However, the LRT could have been used to compare `model1` and `model2`.)

Comparing `model3` to `model4` we test

$$H_0 : \beta_{age^2} = 0 \text{ vs. } H_1 : \beta_{age^2} \neq 0$$

That is, "model 3 and 4 are equally good" vs. "model 4 is better than model 3".

The difference in twice the loglikelihood is reported in the print-out to be 16.47, and the difference in number of parameters is 1. The p-value is calculated as the tail-probability in the chi-squared distribution with one degree of freedom, and is reported to be $5 \cdot 10^{-5}$. Thus, we reject the null hypothesis that the models are equally good, and prefer the larger `model4` to the smaller `model3`.

(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.

Answer:

```r
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
```

```r
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
```

As an estimate for the rate we have $\hat{\lambda} = t\exp(\hat{\eta})$, and we have been informed to use $t = 1000$. From the print-out from `model2` we find the estimated linear predictor for the `Age55-59` year group (call coefficient for this year group β_1) to be $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 = -5.86 + 1.08 = -4.78$. Then the estimate for the expected number of new lung cancer cases is $t \cdot \exp(\hat{\eta}) = 1000 \cdot \exp(-4.78) = 8.4$.

To proceed to find confidence interval we may calculate the estimated standard deviation of $\hat{\eta}$ in a similar way as in Problem 3 (for the difference between the estimated betas).

$$\widehat{\mathrm{Var}}(\hat{\eta}) = \widehat{\mathrm{Var}}(\hat{\beta}_0) + \widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{33} + \frac{2.031}{33} + 2 \cdot \frac{-1}{33} = \frac{1.031}{33} = 0.031$$

```
#alternatively using linear comb
length(model2$coefficients)#6
```

```
## [1] 6
```

```
cvec=matrix(c(1,1,0,0,0,0),nrow=1)
cvec%*%vcov(model2)%*%t(cvec)
```

```
##          [,1]
## [1,] 0.03125
```

To find a confidence interval for λ we start by finding a confidence interval for the linear predictor, because since the estimated regression parameters are asymptotically normal, the estimated linear predictor will also be asymptotic normal.

```
etaL=sum(coef(model2)[1:2])-qnorm(0.975)*sqrt(cvec%*%vcov(model2)%*%t(cvec))
etaU=sum(coef(model2)[1:2])+qnorm(0.975)*sqrt(cvec%*%vcov(model2)%*%t(cvec))
c(etaL,etaU)
```

```
## [1] -5.126387 -4.433435
```

95% CI for η:

$$[\hat{\eta}_L, \hat{\eta}_U] = \hat{\eta} \pm 1.96 \cdot \widehat{\mathrm{SD}}(\hat{\eta}) = -4.78 \pm 1.96 \cdot \sqrt{0.031} = [-5.125, -4.435]$$

```
1000*exp(c(etaL,etaU))
```

```
## [1]  5.937976 11.873633
```

After finding the confidence interval for the linear predictor we then transform the upper and lower limits. 95% CI for λ:

$$[\hat{\lambda}_L, \hat{\lambda}_U] = [t \cdot \exp\hat{\eta}_L, t \cdot \exp\hat{\eta}_U] = [1000 \cdot \exp(-5.125), 1000 \cdot \exp(-4.435)] = [5.95, 11.86]$$

Problem 5: GLM data analysis in R: low birth weight

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.

The commands in R to fit the GLM:

```r
library(MASS)
fit1=glm(low~smoke+ht+lwt,family=binomial,data=birthwt)
summary(fit1)
```

```
##
## Call:
## glm(formula = low ~ smoke + ht + lwt, family = binomial, data = birthwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7067  -0.8312  -0.6892   1.1550   2.2815
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.083538   0.834219   1.299  0.19399
## smoke        0.683910   0.330954   2.066  0.03878 *
## ht           1.822025   0.686039   2.656  0.00791 **
## lwt         -0.018046   0.006565  -2.749  0.00598 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 216.86  on 185  degrees of freedom
## AIC: 224.86
##
## Number of Fisher Scoring iterations: 4
```

Based on your data analyses answer the following questions.

(i) Which type of GLM did you fit? **Correct answer: Binomial**

Possible choices: normal, binomial, Poisson, gamma, Other

(ii) Which link function did you use? **Correct answer: Logit**

Possible choices: identity, log, logit, exp, inverse, negative inverse, other

(iii) How would you explain the meaning of the estimated effect of smoke?

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

Correct answer: the odds of getting a low weight infant was multiplied 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.)

```
predict(fit1,newdata=data.frame(smoke=1,ht=1,lwt=130),type="response")
```

```
##         1
## 0.7761613
```

```
#or alternatively
eta=sum(fit1$coefficients*c(1,1,1,130))
exp(eta)/(1+exp(eta))
```

```
## [1] 0.7761613
```

(v) Perform a likelihood ratio test to compare the fittted 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.)

```
fit2=glm(low~smoke+ht,family=binomial,data=birthwt)
anova(fit1,fit2,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: low ~ smoke + ht + lwt
## Model 2: low ~ smoke + ht
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       185     216.86
## 2       186     225.79 -1  -8.9342 0.002799 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Correct answer: Read off the anova that the p-value is 0.0028.

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

Correct answer: True, since we reject the null hypothesis that both models are equally good and go for the larger model. Formally, the null hypothesis is

$$H_0 : \beta_{lwt} = 0 \text{ vs. } H_1 : \beta_{lwt} \neq 0$$

.

Problem 6: R-code for Problem 5 (optional)

Only to copy in the R-code in Problem 5.

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

```r
library(GLMsData)
data(lime)
model1 <- glm(Foliage ~ Origin * log(DBH),
      family=Gamma(link="log"), data=lime)
summary(model1)
```

```
## 
## 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
```

```r
exp(model1$coefficients)
```

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

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

```
## [1] 0.9999559
```

```r
model2=glm(Foliage ~ Origin+log(DBH),
           family=Gamma(link="log"), data=lime)
anova(model1,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
```

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(DBH)` (natural log), `Origin` and the interaction thereof as covariates are presented in the print-out 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?

Possible choices: R failed, RStudio failed, MASS not installed, Other

(ii) Which type of GLM was fitted? **Correct answer:** Gamma.

Possible choices: normal, binomial, Poisson, gamma, other

(iii) What is the canonical link for this type of response? **Correct answer:** negative inverse.

Possible choices: log, logit, exp, inverse, negative inverse, other

(iv) We look at a tree from `Coppice`. If the `log(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.)

Correct answer: $\exp(1.84) = 6.31$.

```
exp(model1$coeff)
```

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

(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.

Correct answer: True.

The null hypotesis is not rejected, which means that the saturated model and the candidate model is not significantly different.

(vi) A likelihood ratio test is performed to compare this model to the model where the interaction between `log(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.)

Correct answer 0.00071 directly from print-out ANOVA.

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

Correct answer: Origin*log(DHB). The hypothesis that the models are equally good is rejected and we keep the largest model.