15 April 2011, 9:00 – 12:00, zaal 5419.0013 (FEB)

Rijksuniversiteit Groningen
Statistical Modelling, Generalized Linear Models

Exam

RULES FOR THE EXAM:

- The use of a normal, non-graphical calculator is permitted.
- This is a CLOSED-BOOK exam.
- At the end of the exam you can find a chi-squared table.
- In this exam you can use the usual significance cut-off $\alpha = 0.05$.

1. **Gamma regression.** A study investigated the effectiveness of the "bonus-malus" (BM) system (system that adjust the premium according to the individual's claim history) and the height of the insurance excess (eigen risico) on reducing the height of the car insurance claims. Average car insurance claims per customer over a 1-year period were recorded from nine insurance companies, each with a different bonus-malus system (1= present, 0=absent) and with different excess levels.

Company	y	x1	x2
	(Av. claim in 1,000 euro)	BM system	Excess (euro)
1	0.11	1.00	50.00
2	0.18	1.00	0.00
3	0.26	0.00	0.00
4	0.32	1.00	200.00
5	0.46	1.00	50.00
6	0.93	0.00	50.00
7	1.16	0.00	100.00
8	2.50	0.00	100.00
9	3.09	0.00	100.00

The aim is to find a relationship between the height of the car insurance claim (y) and the explanatory variables presence/absence of bonus-malus system (x_1) and excess level (x_2). This is done via a Gamma regression.

The Gamma regression model is defined via these ingredients:

- The claims y are distributed as

$$y \sim \text{Gamma}(\alpha, \beta),$$

whereby $f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} (y > 0)$.

- Let η be the linear predictor, i.e.

$$\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2.$$

- You can use the fact that $\mu = E(Y) = \alpha/\beta$ and $V(Y) = \alpha/\beta^2$.
- The parameters μ and η are linked via the *link function*

$$\eta \;=\; g(\mu)$$

- In this question consider y directly; don't consider any transformation.

(a) Write the probability density function of y as a function of the mean μ of y (HINT: write β as a function of μ and α). What values can μ take?

(b) Show that the distribution of y comes from an exponential family. Determine the canonical parameter θ and the variance function $V(\mu)$.

(c) Determine the canonical link function g and its inverse g^{-1}. What is the problem with this link function?

In what follows, assume that you use the canonical link function.

(d) Show that $\frac{dl}{d\theta} = \frac{y-\mu}{a(\varphi)}$, where l is taken to be the log-likelihood of a single observation y.

(e) Use the fact that the likelihood for β is given as

$$l(\beta) = \log(f(\mathbf{y}; \eta(\beta)))$$

to derive an expression for $\frac{\delta l}{\delta \beta_j}$ for all the data $\mathbf{y}$.

(f) Derive the expression for the second derivative $\frac{\delta^2 l}{\delta \beta_j \delta \beta_k}$ for all the data.

(g) Unfortunately, there is typically no explicit solution for the system of p maximum likelihood equations

$$\frac{\delta l}{\delta \beta} = 0,$$

whereby l is the full log-likelihood and p is the number of columns of X. Therefore, numeric methods, such as the Newton-Raphson and Fisher Scoring algorithm, need to be used to derive the root $\hat{\beta}$ of these equations. Explain, how *in this particular case* the Fisher Scoring algorithm differs from the Newton-Raphson algorithm?

(h) We perform the Gamma regression in R and obtain the following results:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.003734   0.769193  -2.605   0.0404 *
x1          -3.379783   1.126448  -3.000   0.0240 *
x2           0.015513   0.007827   1.982   0.0948 .

(Dispersion parameter for Gamma family taken to be 0.3708685)
```

```
    Null deviance: 10.2367  on 8  degrees of freedom
Residual deviance:  2.0633  on 6  degrees of freedom
AIC: 10.395
```

Interpret the sign of the coefficient $\hat{\beta}$ for x1 (presence/absence of bonus-malus system). NOTE: as stated before, we use here the strict canonical link function (i.e. with minus sign) and this should be taken into account when interpreting $\hat{\beta}_{x1}$.

(i) Test formally whether the model fits the data.

(j) We also fit the model without excess (x2) and obtain

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6294     0.1906  -3.303   0.0131 *
x1           -3.1506     1.2938  -2.435   0.0451 *

(Dispersion parameter for Gamma family taken to be 0.458458)

    Null deviance: 10.2367  on 8  degrees of freedom
Residual deviance:  4.2246  on 7  degrees of freedom
AIC: 15.200
```

Use the AIC and the deviance test for β_{x2} to check whether you can drop insurance excess from the model.

2. **Quasi-likelihood.** The quasi-loglikelihood is defined by assuming that its derivative with respect to the mean $\mu = EY$ can be written as:

$$U(Y, \mu) = \frac{Y - \mu}{\varphi V(\mu)},$$

where φ is an arbitrary scale-factor and $V(\mu)$ the variance function. Assume that for strictly positive data Y, we use the variance function

$$V(\mu) = \frac{1}{\mu}.$$

(a) Derive the quasi-loglikelihood Q_y for a single observation in this case, where $Q_y(\mu) = \int_y^\mu U(y, t)\ dt$.

(b) We observe n data-points $y = (y_1, \ldots, y_n)$ and we would like to model the mean μ_i of y_i via a linear predictor

$$\eta_i = \sum_{j=0}^{p} x_{ij}\beta_j$$

and a link function

$$g(\mu_i) = \sqrt{\mu_i}.$$

Derive the following quantities

i. the full quasi-loglikelihood $Q^*(\beta) = \sum_{i=1}^n Q_{y_i}(\mu(x_i, \beta))$;

ii. the expression for its first derivative with respect to β_j, $\frac{\partial Q^*}{\partial \beta_j}$;

iii. the expression for its second derivative $\frac{\delta^2 Q^*}{\delta\beta_j \delta\beta_k}$.

3. **Logistic regression.** In a toxicity dose-response experiment, different levels of a particular toxin were given to batches of plants. The aim is to model the probability of survival π as a function of the amount of toxin x. Consider the dose-response model

$$g(\pi) = \beta_0 + \beta_1 x,$$

where g is a link-function.

(a) Under the hypothesis that the response probability at dose x_0 is equal to π_0, show that the model reduces to

$$g(\pi) = \beta_0(1 - x/x_0) + g(\pi_0)x/x_0.$$

(b) How would you fit such a model using your favorite computer programme? In your answer focus only on the intercept, the linear term and the off-set.

(c) According to the literature there is a $\pi_0 = 73\%$ survival rate at dose $x_0 = 2$. We want to test this hypothesis and therefore perform an experiment. We measure survival in 5 batches of 10 plants subjected to 5 dose-levels of toxin and fit two logistic regressions in R, with the following results

```
glm(formula = y ~ dose, family = binomial(link = "logit"))

  Coefficients:
              Estimate Exp(Estimate) Std. Error z value Pr(>|z|)
  (Intercept)  -2.6182          0.07     0.9817  -2.667 0.007652
  dose          1.4442          4.24     0.4267   3.384 0.000713

  Residual deviance:  1.5886  on 3  degrees of freedom


T1.dose<-1-dose/2
T2.dose<-log(.73/.27)*dose/2

glm(formula=y~-1+T1.dose, family=binomial(link="logit"), offset=T2.dose)

  Coefficients:
          Estimate Std. Error z value Pr(>|z|)
  T1.dose  -2.0100     0.9854   -2.04   0.0414 *

  Residual deviance:  4.6368  on 4  degrees of freedom
```

Use this output to test the hypothesis $H_0 : \pi(x = 2) = 0.73$.

4. **Poisson regression and contingency tables.** In a study 91 couples in Tucson, Arizona, answered the question:

   > "Sex is fun for me and my partner."

   The possible answers were "never or occasionally", "fairly often", "very often" and "almost always". The data are summarized in the table below. We perform two Poisson regression on the frequencies and find the following results when fitting two models.

	never	fairly	very	always
never	7.00	7.00	2.00	3.00
fairly	2.00	8.00	3.00	7.00
very	1.00	5.00	4.00	9.00
always	2.00	8.00	9.00	14.00

NOTE: The factor **symfac** is the factor that models a symmetric response, i.e., if α_{ij} is the coefficient corresponding to the **symfac** factor (i=husband, j=wife), then $\alpha_{ij} = \alpha_{ji}$.

```
glm(formula = y ~ symfac, family = poisson, data = sexfun)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.6391     0.2673   9.874  < 2e-16 ***
symfac:always-fairly  -0.6242     0.3716  -1.680 0.093038 .
symfac:always-never   -1.7228     0.5210  -3.307 0.000944 ***
symfac:always-very    -0.4418     0.3563  -1.240 0.215016
symfac:fairly-fairly  -0.5596     0.4432  -1.263 0.206710
symfac:fairly-never   -1.1350     0.4272  -2.657 0.007894 **
symfac:fairly-very    -1.2528     0.4432  -2.827 0.004704 **
symfac:never-never    -0.6931     0.4629  -1.497 0.134297
symfac:never-very     -2.2336     0.6362  -3.511 0.000447 ***
symfac:very-very      -1.2528     0.5669  -2.210 0.027128 *

    Null deviance: 33.5846  on 15  degrees of freedom
Residual deviance:  4.0552  on  6  degrees of freedom
```

```
glm(formula = y ~ symfac + husband + wife, family = poisson, data = sexfun)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.3686     0.6449   5.223 1.76e-07 ***
symfac:always-fairly  -0.4829     0.4116  -1.173 0.240723
symfac:always-never   -2.1526     0.6544  -3.289 0.001004 **
symfac:always-very    -0.4716     0.4121  -1.144 0.252474
symfac:fairly-fairly  -0.2539     0.6088  -0.417 0.676684
```

```
symfac:fairly-never   -1.4753      0.6479  -2.277 0.022778 *
symfac:fairly-very    -1.1458      0.5563  -2.060 0.039436 *
symfac:never-never    -1.4227      0.7475  -1.903 0.057012 .
symfac:never-very     -2.6829      0.7750  -3.462 0.000536 ***
symfac:very-very      -1.3115      0.6952  -1.887 0.059216 .
husband:fairly        -1.0353      0.5694  -1.818 0.069037 .
husband:very          -0.6708      0.6230  -1.077 0.281608
husband:always        -0.7295      0.5869  -1.243 0.213881

    Null deviance: 33.58461  on 15  degrees of freedom
Residual deviance:  0.36600  on  3  degrees of freedom
```

Assuming that these two models are the only relevant models in this example (i.e. no simpler model is better than these two models), then answer the following questions.

(a) What are the two *types* of models that have been fitted above?

(b) Test whether marginal homogeneity holds for these data.

df	$\chi^2_{0.05}$	$\chi^2_{0.95}$
1	0.00	3.84
2	0.10	5.99
3	0.35	7.81
4	0.71	9.49
5	1.15	11.07
6	1.64	12.59

Table 1: Quantile of a chi-squared χ^2_{df} with df degrees of freedom.