15 April 2010, 9:00 – 12:00

Rijksuniversiteit Groningen
Statistical Modelling

Exam

1. **Negative binomial regression.** In a Dutch study on the effectiveness of swimming lessons, data are recorded on the number of lessons it takes children to obtain both their A & B diploma, sequentially. For each child we record sex and age at the start of the swimming instruction. The aim is to find a relationship between the number of swimming lessons to obtain both diplomas (y) and the explanatory variables gender (x_1) and age (x_2). This is done via a negative binomial regression.

   The negative binomial regression model is defined via these ingredients:

   - The counts y are distributed as
   $$y \sim \text{NegBinom}(p, k),$$
   whereby $p(y) = \binom{y-1}{k-1} p^k (1-p)^{y-k}$.
   - We assume k is known and equal $k = 2$.
   - Let η be the linear predictor, i.e.
   $$\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2.$$
   - The parameters μ and η are linked via the *link function*
   $$\eta \quad = \quad g(\mu)$$
   - In this question consider y directly: you don't have to consider any transformation.
   - In your answer, use the following table with quantiles of the chi-squared distribution:

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

   (a) Why might a value $k = 2$ be *a priori* sensible in this problem?

   In what follows, you can assume that $k = 2$.

(b) Write the probability mass function of y as a function of the mean μ of y. What values can μ take?

(c) Show that the count y has a distribution from an exponential family. Determine the canonical parameter θ and the variance function $V(\mu)$.

(d) 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.

(e) 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.

(f) 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 the full data $\mathbf{y}$.

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

(h) 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 need to be used to derive the root $\hat{\beta}$ of these equations. Derive the Newton-Raphson algorithm used to determine β in this case. You can make use of matrix notation, but make sure that you explain what the entries in your matrices are.

(i) Explain, how *in this particular case* the Fisher Scoring algorithm differs from the Newton-Raphson algorithm?

(j) We perform the negative binomial regression in `R` and obtain the following results:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06713    0.15262  -0.440   0.6617
gendergirl   0.13350    0.07269   1.837   0.0715 .
age         -0.05013    0.02262  -2.217   0.0306 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 26.204  on 59  degrees of freedom
Residual deviance: 25.819  on 57  degrees of freedom
AIC: 131.50
```

(k) Interpret the coefficient $\hat{\beta}$ for gender.

(l) Formally tests how well the model fits.

(m) We also fit the model without gender and obtain

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02695    0.14321   0.188   0.8514
age         -0.05414    0.02249  -2.408   0.0193 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 26.204  on 59  degrees of freedom
Residual deviance: 25.961  on 58  degrees of freedom
AIC: 129.64
```

Use the AIC, the deviance test and the direct test for $\beta_{\text{gender}} = 0$ to check whether you can drop gender from the model.

2. **AIC.** The Akaike Information Criterion is tool to select a model. It is closely related to the concept of the *mean expected maximum log-likelihood*:

$$E_{Y|\theta_0} E_{X|\theta_0} l_X(\hat{\theta}(Y)),$$

where θ_0 is the true parameter, X and Y data generated according to the true model and $\hat{\theta}(Y)$ the maximum likelihood estimator based on the data Y.

(a) Explain heuristically the role of this double expectation in selecting a model, rather than for example the single expectation $E_{X|\theta_0} l_X(\hat{\theta}(X))$.

(b) Take a Taylor expansion of $l_X(\theta_0)$ around $\hat{\theta}(X)$ to prove that

$$E_{X|\theta_0} l_X(\theta_0) = E_{X|\theta_0} l_X(\hat{\theta}(X)) - k/2,$$

whereby k is the dimensionality of θ_0.

3. **Logistic regression.** In a toxicity experiment, different levels (linearly increasing from 1 to 5 in some unit) of a particular toxin were given to 5 batches of 10 plants. The results of the experiment are given below.

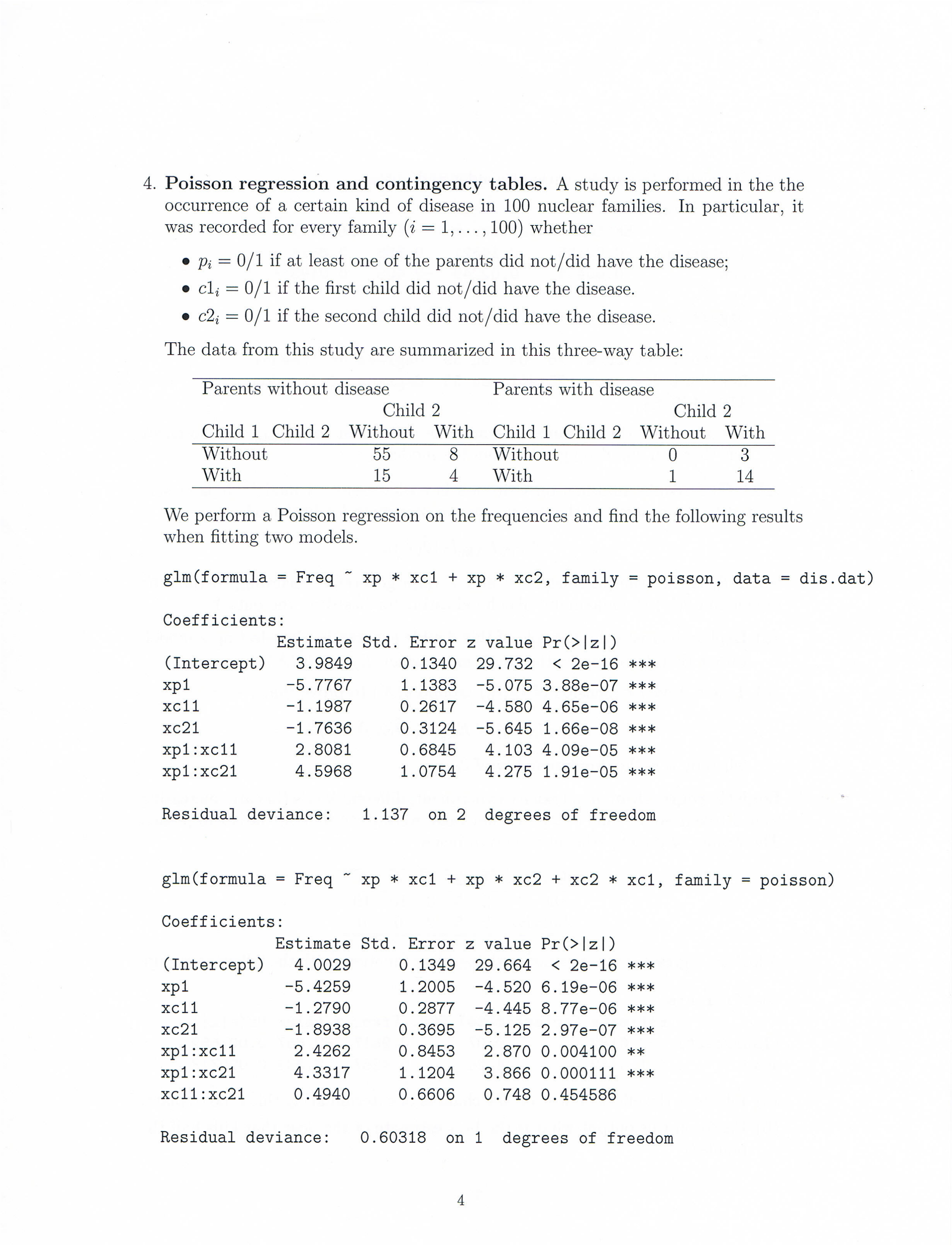
Dose	1	2	3	4	5
Die	3	5	8	10	10
Survive	7	5	2	0	0

A logistic regression is performed with logit link function, with the following results

```
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
```

(a) Interpret the effect of dose on the chances of exterminating this type of plant.

(b) Based on this output what is the best estimate of the dose that kills half the plants?



4. **Poisson regression and contingency tables.** A study is performed in the the occurrence of a certain kind of disease in 100 nuclear families. In particular, it was recorded for every family ($i = 1, \ldots, 100$) whether

- $p_i = 0/1$ if at least one of the parents did not/did have the disease;
- $c1_i = 0/1$ if the first child did not/did have the disease.
- $c2_i = 0/1$ if the second child did not/did have the disease.

The data from this study are summarized in this three-way table:

Parents without disease				Parents with disease			
		Child 2				Child 2	
Child 1	Child 2	Without	With	Child 1	Child 2	Without	With
Without		55	8	Without		0	3
With		15	4	With		1	14

We perform a Poisson regression on the frequencies and find the following results when fitting two models.

```
glm(formula = Freq ~ xp * xc1 + xp * xc2, family = poisson, data = dis.dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9849     0.1340  29.732  < 2e-16 ***
xp1          -5.7767     1.1383  -5.075 3.88e-07 ***
xc11         -1.1987     0.2617  -4.580 4.65e-06 ***
xc21         -1.7636     0.3124  -5.645 1.66e-08 ***
xp1:xc11      2.8081     0.6845   4.103 4.09e-05 ***
xp1:xc21      4.5968     1.0754   4.275 1.91e-05 ***

Residual deviance:   1.137  on 2  degrees of freedom


glm(formula = Freq ~ xp * xc1 + xp * xc2 + xc2 * xc1, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.0029     0.1349  29.664  < 2e-16 ***
xp1          -5.4259     1.2005  -4.520 6.19e-06 ***
xc11         -1.2790     0.2877  -4.445 8.77e-06 ***
xc21         -1.8938     0.3695  -5.125 2.97e-07 ***
xp1:xc11      2.4262     0.8453   2.870 0.004100 **
xp1:xc21      4.3317     1.1204   3.866 0.000111 ***
xc11:xc21     0.4940     0.6606   0.748 0.454586

Residual deviance:   0.60318  on 1  degrees of freedom
```

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

(a) Interpret both models that have been fitted.

(b) Is there evidence that the full dependence model is required to model these data?

(c) Test which of the two models is better.