UPPSALA UNIVERSITY
Department of Mathematics
Raazesh Sainudiin

Generalized Linear Models 1MS369
SAMPLE exam, autumn 2018

Time: 08.00-13.00. Total sum: 30p. Grades 3, 4 and 5 require 14p, 19p, 24p, respectively. Permitted aids: Any text books, notes and pocket calculator. Do not use red colour, only write on one side of each paper.

1. The following table gives the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1983.

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Collisions	3	6	4	7	6	2	2	4	1	7	3	5	6	1
Miles	281	276	268	269	281	271	265	264	267	265	267	260	231	249

   (a) Is it plausible that the collision counts are independent Poisson variates? Examine by considering the fitted model `m1` in Appendix (significance level 0.95). A second model (found as `m2`) was introduced, including log(miles) as an offset. Compare the models by checking estimates of the dispersion parameter in the exponential dispersion family in each case — is the fit improved by inclusion of an offset term? Judging from data, is the last conclusion likely? (3p)

   (b) Suppose the collissions can be modelled as independent and identically distributed Poisson random variates with probability mass function:

$$f(X = x; \lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, 2, \ldots\}, \quad \lambda > 0.$$

   Derive the Maximum Likelihood Estimator (MLE) and the $1-\alpha$ confidence interval based on the asymptotic normality of the MLE. Show each step of your derivation. (6p)

   (c) Find the MLE of the parameter λ in (b) based on the observed collissions in (a) above. (1p)

   (d) Produce the 95% confidence interval for the parameter λ in (b) based on the observed collissions in (a) above. (1p)

   (e) What is the relationship between MLE of λ obtained in (c) and the estimate of the intercept term in the fitted model `m1` given in (a) above. (1p)

   [Some quantiles of possible use: Normal quantile $\Phi^{-1}(1-0.025) = z_{0.05/2} = 1.96$ and χ^2 quantile $\chi^2_{0.05}(13) = 22.36$.]

2. Consider worldwide airline fatalities for the period 1976–1985. Data are given in the table below, where we find number of annual fatal accidents, number of annual passenger deaths and the annual number of recorded passenger miles (100 million). Note that the volume of air traffic nearly doubled over this 10-year period.

Year	Fatal accidents	Passenger deaths	Passenger miles (10^8)
1976	24	734	3863
1977	25	516	4300
1978	31	754	5027
1979	31	877	5481
1980	22	814	5814
1981	21	362	6033
1982	26	764	5877
1983	20	809	6223
1984	16	223	7433
1985	22	1066	7107

Models `m1` and `m2` were fitted by R, using a GLM with Poisson distributed response, see Appendix.

(a) Consider model `m1`. Introduce notation and formulate the corresponding formulae for the regression model (use the canonical link). (2p)

(b) Based on the estimates in each of the models, was air traffic becoming safer over the period in question? Motivate your answer. (2p)

(c) Investigate goodness of fit for the models. If a model is not working well, try to think of an explanation (in terms of Poisson processes and aircraft sizes). (2p)

[Some quantiles: $\chi^2_{0.05}(1) = 3.84$, $\chi^2_{0.05}(8) = 15.51$, $\chi^2_{0.05}(10) = 18.31$]

3. Consider the risk of infection from births by Caesarian section. The response variable of interest is the occurrence of infections following the operation: `infection`, (= 1 if infection, = 0 if not). Three dichotomous covariates that might affect the risk of infection were studied:

`planned`	Caesarian section planned (= 1) or not (= 0)
`risk`	Risk factors such as diabetes, excessive weight or others present (= 1) or absent (= 0)
`antibio`	Antibiotics given as prophylactic (= 1) or not (= 0)

Models for logistic regression were fitted in R, see output in Appendix.

(a) First, a model with only the three main effects was fitted (model `m1`, see summary in Appendix). Investigate goodness of fit by using a suitable deviance. (2p)

(b) A model consisting of `m1` plus an interaction `planned*antibio` was fitted (model `m2`). Perform a test to investigate whether this model is better. (2p)

(c) Consider again the output for model `m1`. In the computations performed, the dispersion parameter was taken to be one. Based on a suitably chosen deviance, give an estimate of the dispersion parameter. Would a quasi-binomial approach be a relevant idea to examine in further modelling? (2p)

[Useful quantiles: $\chi^2_{0.05}(8) = 15.51$, $\chi^2_{0.05}(10) = 18.31$, $\chi^2_{0.05}(12) = 21.02$]

4. A Poisson random variable Y is believed to depend on a covariate x and it is proposed that a log-linear model with a systematic component $a + bx$ is appropriate.

In an experiment, the following values were observed:

$x_1 = -1$	$x_2 = 0$	$x_3 = 1$
$y_1 = 3$	$y_2 = 4$	$y_3 = 13$

(a) Write down the log-likelihood function for the proposed log-linear model for these data. (2p)

(b) Find the maximum likelihood estimates of the coefficients in this log-linear model. (4p)

Table of formulae.

Exponential dispersion family:

$$f(y;\theta) = c(y,\lambda)\exp(\lambda(\theta y - \kappa(\theta)))$$

Unit deviance:

$$d(y;\mu) = 2\int_{\mu}^{y} \frac{y-u}{V(u)}\,\mathrm{d}u.$$

Table of canonical links and variance functions:

Family	$V(\mu)$	Canonical link
Normal	1	μ
Poisson	μ	$\ln\mu$
Gamma	μ^2	$1/\mu$
Binomial	$\mu(1-\mu)$	$\ln(\mu/(1-\mu))$

Good luck!

Appendix.

Problem 1

```
summary(m1)

Call:
glm(formula = cols ~ 1, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8262  -0.9943  -0.0355   0.8922   1.3152

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4040     0.1325    10.6   <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: 15.937  on 13  degrees of freedom
Residual deviance: 15.937  on 13  degrees of freedom
AIC: 61.748

Number of Fisher Scoring iterations: 4

summary(m2)

Call:
glm(formula = cols ~ 1 + offset(log(miles)), family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.83701  -1.02075  -0.04088   0.79625   1.31757

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1768     0.1325  -31.54   <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: 16.06  on 13  degrees of freedom
Residual deviance: 16.06  on 13  degrees of freedom
AIC: 61.871

Number of Fisher Scoring iterations: 4
```

Problem 2.

```
> summary(m1)

Call:
glm(formula = fatals ~ time + offset(log(miles)), family = poisson,
    data = flights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2827  -0.5812  -0.1231   0.7251   1.0209

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 201.32999   45.62333   4.413 1.02e-05 ***
time         -0.10442    0.02304  -4.532 5.84e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26.1320  on 9  degrees of freedom
Residual deviance:  5.4551  on 8  degrees of freedom
AIC: 59.424

Number of Fisher Scoring iterations: 4

> summary(m2)

Call:
glm(formula = deaths ~ time + offset(log(miles)), family = poisson,
    data = flights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-22.342   -3.886    3.351    4.778   14.240

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 117.767873   8.420250   13.99   <2e-16 ***
time         -0.060522   0.004252  -14.23   <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: 1253.6  on 9  degrees of freedom
Residual deviance: 1051.4  on 8  degrees of freedom
AIC: 1138.3

Number of Fisher Scoring iterations: 4
```

Problem 3.

```
> summary(m1)

Call:
glm(formula = infection ~ antibio + planned + risk, family = binomial,
    data = caesar, weights = weight)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.771  -2.664   0.000   2.752   6.904

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8207     0.4947  -1.659   0.0971 .
antibio1     -3.2544     0.4813  -6.761 1.37e-11 ***
planned1     -1.0720     0.4254  -2.520   0.0117 *
risk1         2.0299     0.4553   4.459 8.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.01  on 11  degrees of freedom
Residual deviance: 226.52  on  8  degrees of freedom
AIC: 234.52

Number of Fisher Scoring iterations: 5

> summary(m2)

Call:
glm(formula = infection ~ antibio + planned + risk + planned *
    antibio, family = binomial, data = caesar, weights = weight)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.745  -2.692   0.000   2.661   6.936

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.7883     0.5071  -1.554   0.1201
antibio1           -3.3130     0.5256  -6.303 2.91e-10 ***
planned1           -1.1188     0.4568  -2.449   0.0143 *
risk1               2.0333     0.4564   4.455 8.37e-06 ***
antibio1:planned1   0.3375     1.1698   0.289   0.7729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 299.01  on 11  degrees of freedom
Residual deviance: 226.44  on  7  degrees of freedom
AIC: 236.44

Number of Fisher Scoring iterations: 5
```