# Generalized Linear Models (ST425A)
(Advanced level course, 7.5 hec, Aut. 2020)
# Home Re-Examnation (Parts 1 & 2)

Gebrenegus Ghilagaber (Professor)
Department of Statistics, Stockholm University

- **Date and time:** Wednesday 13 January 2021, 09:00 - 15:00
- **Permitted facilities:** All relevant facilities but **NOT** collaboration with (or support from) other person/s.
- **Return date of corrected exam:** Information will be sent via e-mail or Athena.
- **General Instructions:**
  - For questions about the content of the exam, contact the course coordinator on email Gebre@stat.su.se. Incoming e-mail questions will be answered continuously during the exam.
  - If the course coordinator needs to send out information to all students during the exam, it will be sent to your registered e-mail address. Therefore, check your e-mail during the exam.
  - Please note that practical help is only available during the first hour of the exam by e-mail to expedition@stat.su.se. Please read carefully the enclosed instructions for exam submission. There, you find all the necessary information about submission, anonymous code, etc. If you, despite the instructions have problems submitting the exam, e-mail the exam to tenta@stat.su.se. However, this is only done in exceptional cases.

# 1 Part I (Theoretical part)

- **Instructions for Part I:**
  - The exam consists of 4 questions
  - The total amount of points for this part of the exam is 30.
  - Minimum requirement to pass this part of examination is 20 points.
  - Solutions to each question should be detailed enough and well-motivated in order to get full points.

## 1.1 Question 1 (6 p)

- Suppose we have 4 observations from an exponential distribution with parameter $\theta$ and the realized values are $y_1 = 0.06$, $y_2 = 0.09$, $y_3 = 0.1$, and $y_4 = 0.2$.
- a) Use the observed values to compute a numerical value of the maximum likelihood estimator of $\theta$
- b) Use your result in (a) to compute the value of the corresponding likelihood function
- c) Does your result in (b) look strange (somehow)? If yes, in what way and what lesson can one learn from the result?

## 1.2 Question 2 (8)

Assume that $Y$ is a random variable with a distribution that belongs to the exponential family with a parameter $\theta$. Hence, the pdf can be written as

$$f(y;\theta) = \exp\left[a(y)\, b(\theta) + c(\theta) + d(y)\right]$$

for some functions $a(y)$, $b(\theta)$, $c(\theta)$, and $d(y)$. In a Generalized Linear Model, the expected value of $Y$, say $\mu$, satisfies the relation $g(\mu) = \eta$, where $g$ is a link function and $\eta = x^T\beta$ is a linear predictor of a vector of explanatory variables $x$.

- a) What is the canonical parameter?

- b) Use the properties of the exponential family to suggest an estimator for $\theta$
- c) Describe the difference between the method of Newton-Raphson and the Fisher scoring method when computing estimates of the parameters.
- d) Define deviance.

## 1.3 Question 3 (8 p)

- a) Use the properties of distributions in the exponential family to derive the expected values and variances of the Binomial, Poisson, and Negative Binomial distributions.
- b) How are the mean and variance in each of the above three distributions related?
- c) How do the relationships between the mean and variance in (b) affect your choice of a model for a given data set?

## 1.4 Question 4 (8 p)

The Tables below contains results from analyisis of effects of *Education* (with secondary-level education as a baseline) and *Age* on some demographic outcome among a sample of women.

| Parameter | Levels | Symbol | Estimate | Std. Eroor | t-ratio |
|---|---|---|---|---|---|
| Constant | - | $\mu$ | $-5.954$ | 7.166 | $-0.83$ |
| Education | None | $\alpha_1$ | 19.448 | 3.729 | 5.21 |
| | Primary | $\alpha_2$ | 4.144 | 3.191 | 1.30 |
| Age | (linear) | $\beta$ | 0.1693 | 0.1056 | 1.60 |

| Source of variation | Degrees of freedom | Sum of squares | Mean Square | F-Ratio |
|---|---|---|---|---|
| Age (linear) | 1 | 1201.1 | 1201.1 | 36.5 |
| Education\|Age | 2 | 923.4 | 461.7 | 14.1 |
| Residual | 16 | 525.7 | 32.9 | |
| Total | 19 | 2650.2 | | |

- a) What type of analyis is likely to have given rise to the above tables.
- b) Interpret the effect of *Age* on the outcome of interest.
- c) What portion of the total variation in the outcome variable is due to variability in *Age*?
- d) Do the results in the table support the hypothesis of no education-effect (i.e. $H_0 : \alpha_2 = \alpha_3 = 0$)?

# 2 Part II (Analyses of empirical data)

- **Instructions for Part II:**
  - The total amount of points for this part of the exam is 20.
  - The minimum requirement to pass this part of examination is 10 points.
  - Solutions to each question should be detailed enough and well-motivated in order to get full points.

---

The table below presents the results of a survey on ideal number of children (grouped into 3 increasing categories) among 3663 women cross classified by their education (3 levels) and residence (2 levels).

| | | **Ideal # children** | | |
|---|---|---|---|---|
| **Education** | **Residence** | $\leq 2$ | $3-5$ | $>5$ |
| None | Urban | 218 | 200 | 248 |
| | Rural | 731 | 643 | 532 |
| Primary | Urban | 220 | 143 | 121 |
| | Rural | 134 | 44 | 18 |
| Secondary | Urban | 250 | 94 | 35 |
| or above | Rural | 22 | 6 | 4 |

- a) Fit an appropriate model with *Ideal # children* as response variable and *Education* as explanatory variable.

- b) Repeat (a) with *Education* and *Residence* as explanatory variables.
- c) Use the model-deviances in (a) and (b) to suggest which model is more adequate.
- d) Do the results in (a) and (b) indicate you need to add an interaction term (between *Education* and *Residence*) to the model in (b)?
- e) Interpret your final results on the effect of *Education* and *Residence* on ideal children ever born

Summarize your results in a form of a report that includes choice of a model (with justification), the fitted model, and overall comments on your results (estimates and test statistics). Attach relevant SAS or R codes, tables and figures as appendices.