Generalized Linear Models (ST425A)
(Advanced level course, 7.5 hec, Aut. 2020)

Home-Examnation (Parts 1 & 2)

Gebrenegus Ghilagaber (Professor)
Department of Statistics, Stockholm University

- **Date and time:** Wednesday 2 December 2020, 10:00 - 16:00
- **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

Verify that each of the following distributions belongs to the exponential family and examine if it is in cannonical form:

- a) Exponential distribution: $f(y;\theta) = \theta e^{-\theta y}$
- b) Negative Binomial distribution:

$$f(y;\theta) = \begin{pmatrix} y-r+1 \\ r-1 \end{pmatrix} \theta^r (1-\theta)^y \text{ where } r \text{ is known}$$

- c) Pareto distribution: $f(y;\theta) = \theta y^{-(\theta+1)}$

1.2 Question 2

- 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 in a given situation?

1.3 Question 3

- a) Assume N observations are to be made on Y with common parameters (n and π) so that

$$Y_i \sim Bin\left(n, \pi\right), \quad i = 1, ..., N.$$

Derive the maximum likelihood estimator $\widehat{\pi}$ of π and compare the deviance of a model based on $\widehat{\pi}$ with that of a maximal model where π differs across the Y_i so that $Y_i \sim Bin\left(n, \pi_i\right)$.

- b) Let $Y_1, ..., Y_N$ be independent random variables with $Y_i \sim$Poisson(μ_i) and

$$\ln\left(\mu_i\right) = \beta_1 + \sum_{j=2}^{J} x_{ij}\beta_j, \qquad i = 1, ..., N.$$

Show that the score statistic for β_1 is $U_1 = \sum_{i=1}^{N}\left(y_i - \mu_i\right)$.

- c) Using the result in b), show that for the likelihood estimates $\widehat{\mu}_i$, we have that $\sum_{i=1}^{N}\widehat{\mu}_i = \sum_{i=1}^{N} y_i$.

1.4 Question 4

A generalized linear model is defined as

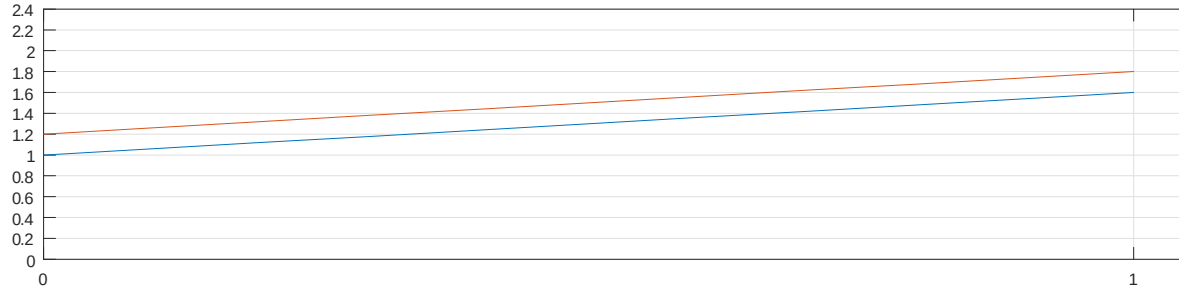
- (i) $Y \sim Bin\left(\pi\right)$ where $Y = 1$ if an observed individual has a particular disease and $Y = 0$ else.

- (ii) $\eta = \beta_0 + \beta_1 E + \beta_2 S$ where E inidcates whether the individual is exposed to a specific pollutant ($E = 1$) or not ($E = 0$), and S is indicator of sex (0 for men and 1 for women).

- (iii) a logit link:

$$g\left(\eta\right) = \frac{\exp\left(\eta\right)}{1 + \exp\left(\eta\right)}$$

The model was estimated using data from $N = 62$ individuals (35 women and 27 men). Below is a graph of the fitted model. The red line represents women ($S = 1$) and the blue line represents men ($S = 0$).



- a) Use the graph to calculate the estimates of β_0, β_1, and β_2
- b) Assume the variancs of the estimates are 0.0961, 0.0576, and 0.0009. Use Wald statistic to test whether the estimates of β_1 and β_2 are significantly different from zero at 5 % level.
- c) What is the log odds of getting the disease for individuals in the risk group (exposed to pollutant) as compared to those not exposed? Make calculations for men and women separately.

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 shows deaths and exposure months among children under five years old grouped by 5 birth cohorts (1 for the youngest cohort and 5 for the oldest) and 3 educational levels of their mothers.

Birth Cohort	Education	Deaths	Exposure months
	None	3	1694
1	Primary	24	6483
	Secondary+	23	12843
	None	21	6801
2	Primary	31	22454
	Secondary+	60	47516
	None	12	9599
3	Primary	21	17736
	Secondary+	47	38335
	None	11	10843
4	Primary	13	10480
	Secondary+	17	18331
	None	17	9167
5	Primary	10	6329
	Secondary+	9	6331
	Total	319	224936

- a) Fit an appropriate model with *Deaths* as response variable and *Birth Cohort* as explanatory variable (and taking due account of the *Exposure months*).
- b) Repeat (a) with *Birth Cohort* and *Education* 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 that effects of *Education* might have changed across the *Birth Cohorts* (suggesting to add an interaction term between *Education* and *Birth Cohort* in (b))?
- e) Interpret your final results on the effects of *Birth Cohort* and *Education* on the risk of under-five mortality among the children studied.

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 R/SAS codes, tables and figures as Appendices.