Generalized Linear Models (ST425A)
(Advanced level course, 7.5 hec, Aut. 2019)

Examnation (Part 1)

Gebrenegus Ghilagaber (Professor & Head)
Department of Statistics, Stockholm University

- **Date and time:** Monday 30 September 2019, 16:00 - 19:00
- **Permitted facilities:** Pocket calculator and sample of formula attached at the end of this exam.
- **Return of exam:** Not yet decided (information will be sent via E-mail or Athena).
- **Instructions:**
  - The exam consists of 4 questions and a sample of formula is given at the end
  - The total amount of points for this part of the exam is 30.
  - The 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 Question 1

Define a Generalized Linear Model and describe its "components".

2 Question 2

Consider the density function

$$f(y;\theta) = \theta \exp(-\theta y), \quad 0 < y < \infty, \ \theta > 0$$

- a) Show that the function belongs to the exponential family and indicate the canonical parameter
- b) Derive the maximum likelihood estimator of θ
- c) Suppose now we have 4 observations from the above density function with values $y_1 = 0.06$, $y_2 = 0.09$, $y_3 = 0.1$, and $y_4 = 0.2$. Use these values to compute a numerical value of the maximum likelihood estimator of θ
- d) Use your result in (c) to compute the value of the corresponding likelihood function
- e) Does your result in (d) look strange (somehow)? If yes, in what way and what lesson can one learn from the result?

3 Question 3

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	-	μ	-5.954	7.166	-0.83
Education	None	α_1	19.448	3.729	5.21
	Primary	α_2	4.144	3.191	1.30
Age	(linear)	β	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) How large of the total variation in the outcoem variable is due to variability in Age?
- d) Do the results in the table support the hypothesis of no education-effect (i.e. $H_0 : \alpha_2 = \alpha_3 = 0$)?

4 Question 4

The Table below presents results of analysis of count data across some models:

	Fitted models			
Estimates of interest	Poisson	Overdispersed Poisson	Negative Binomial	Zero-inflated
Coefficient ($\widehat{\beta}$)	-0.187	-0.187	-0.306	-0.187
Z-value	-5.12	-1.44	-3.10	-5.00
# param. estimated	p	$p+1$	$p+1$	$2p$
Log-likelihood	-18373	-18373	-9829	-15957

- a) Based on the results above, which of the fitted models provides the best description of the data? Don't forget to justify your answer
- b) Does the table indicate the original data is overdispersed?
- c) Does the table indicate the original data contains excess zeroes than would be expected?
- d) What is your conclusion concerning the explanatory variable of interest (whose coefficient is denoted by β)

A Formulas for Generalized Linear Models (A sample)

- Exponential family

$$\begin{aligned} f(y;\theta) &= \exp\left[a(y)\,b(\theta)+c(\theta)+d(y)\right] \\ E\left[a(Y)\right] &= -\frac{c'(\theta)}{b'(\theta)} \\ var\left[a(Y)\right] &= -\frac{b''(\theta)\,c'(\theta)-c''(\theta)\,b'(\theta)}{\left[b'(\theta)\right]^3} \end{aligned}$$

- Score statistic

$$\begin{aligned} U &= a(y)\,b'(\theta)+c'(\theta) \\ U_j &= \sum_{i=1}^{N}\left[\left(\frac{y_i-\mu_i}{var(Y_i)}\right)x_{ij}\left(\frac{\partial\mu_i}{\eta_i}\right)\right] \end{aligned}$$

- GLM weights

$$w_{ii}=\frac{1}{var(Y_i)}\left(\frac{\partial\mu_i}{\eta_i}\right)^2$$

- Deviance

$$2\left[\log L\left(\widehat{\beta}_{\max};y\right)-\log L\left(\widehat{\beta};y\right)\right]$$

- **Binomial distribution**
  - Probability function (using π in place of $\theta=p$) :

$$f(y;n,\pi)=\binom{n}{y}\pi^y(1-\pi)^{n-y},\qquad y=0,1,\ldots,n$$

  - Probit model

$$\pi_i=\Phi\left(\mathbf{x}_i^T\beta\right)\Longrightarrow\Phi(\pi_i)=\mathbf{x}_i^T\beta$$

  - Logit model

$$\pi_i=\frac{\exp\left(\mathbf{x}_i^T\beta\right)}{1+\mathbf{x}_i^T\beta}\Longrightarrow\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\mathbf{x}_i^T\beta$$

- Complementary log-log link

$$\pi_i = 1 - \exp\left[-\exp\left(\mathbf{x}_i^T\beta\right)\right] \Longrightarrow \ln\left[-\ln\left(1-\pi_i\right)\right] = \mathbf{x}_i^T\beta$$

- **Multinomial distribution**
  - Probability function (again using π_i in place of $\theta_i = p_i$) :

$$f\left(y_1, ..., y_J ; n_1, ..., n_J ; \pi_1, ..., \pi_J ;\right) = \begin{pmatrix} n \\ y_1...y_J \end{pmatrix} \pi_1^{y_1}\pi_2^{y_2}....\pi_J^{y_J},$$

for $y_j = 0, ...n$; $j = 1...., J$; and $\sum_{j=1}^{J} y_j = n$.

  - Normal logistic model

$$\ln\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j^T\beta_j, \quad j = 2, ..., J.$$

  - Ordinal response:
    - ∗ Cummulative logit model

$$\ln\left(\frac{\pi_1 + ... + \pi_j}{\pi_{j+1} + ... + \pi_J}\right) = \mathbf{x}_j^T\beta_j, \quad j = 1, ..., J-1.$$

    - ∗ Proportional odds model

$$\ln\left(\frac{\pi_1 + ... + \pi_j}{\pi_{j+1} + ... + \pi_J}\right) = \beta_{0j} + \beta_1 x_1 + ... + \beta_{p-1}x_{p-1}, \quad j = 1, ..., J.$$

    - ∗ Adjacent categories model

$$\ln\left(\frac{\pi_j}{\pi_{j+1}}\right) = \mathbf{x}_j^T\beta_j, \quad j = 1, ..., J-1.$$

    - ∗ Continuation ratio logit model

$$\ln\left(\frac{\pi_j}{\pi_{j+1} + ... + \pi_J}\right) = \mathbf{x}_j^T\beta_j, \quad j = 1, ..., J-1.$$

- **Poisson distribution**
  - Probability function (using λ in place of θ) :

$$f\,(y\,;\lambda\,) = \frac{\lambda^y}{y!}e^{-\lambda}, \quad y = 0, 1, ...$$

  - log link function

$$\ln\,(y_i) = \ln\,(n_i) + \mathbf{x}_j^T\beta_j$$

- **Normal distribution**
  - Probability function (using μ and σ^2 in place of θ_1 and θ_2) :

$$f\left(y\,;\mu;\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right], \quad -\infty < y < \infty.$$