ST6103 - GLM - Assignment

Marc Henrion

15-19 July 2019

Assigment

Please email your typed or scanned solutions to `mhenrion@mlw.mw` and `biostat-unima@cc.ac.mw`, before 23:59 on Monday 2 September 2019. Please include `ST6103 - Assignment` in the subject line.

While we used R in the classroom, you can use any software package of your liking to fit models for this assignment. Please include both your fitting code and model output and graphs with your solutions.

However, for any code that you are submitting, please explain what each block of code does by including comment lines in your code.

Notation

Please try to use the following notation where possible.

- X, Y - random variables (here: X = predictor, Y = response)
- x, y - measured / observed values
- ϵ - random variable (here: error / residual)
- θ - a vector of parameters
- $\bar{X}$, $\bar{Y}$ - sample mean estimators for X, Y
- $\bar{x}$, $\bar{y}$ - sample mean estimates of X, Y
- $\hat{T}$, $\hat{t}$ - given a statistic T, estimator and estimate of T
- $P(A)$ - probability of an event A occuring
- $f_X(.)$, $f_Y(.)$ - distribution mass / density functions of X, Y
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[T]$ - the expectation of X, Y, T respectively
- GLM = generalised linear model

Exercise 1 [25 marks]

Let Y be a random variable, distributed according to the inverse Gaussian distribution with parameters μ, λ: $Y \sim IG(\mu, \lambda)$. This implies that the probability density function of Y is given by

$$f_Y(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$$

The mean of Y is $E[Y] = \mu$ and the variance of Y is $Var(Y) = \frac{\mu^3}{\lambda}$.

For this exercise:

1. Show that $IG(\lambda, \mu)$ is an exponential family distribution and state the canonical parameter θ as well as functions $a(\phi), b(\theta), c(y, \phi)$. [10 marks]
2. Derive the canonical link function for $IG(\lambda, \mu)$. [5 marks]
3. Given a dataset $\{\mathbf{y}, \mathbf{X}\}$ where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ an $n \times (p+1)$ matrix of predictor variables (with the first column being just a vector of 1s), derive the deviance function for an inverse Gaussian GLM with canonical link function. [5 marks]
4. Derive the deviance residuals for this model. [5 marks]

Exercise 2 [40 marks]

You are given the following data:

$$\mathbf{x} = (-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)^T$$

$$\mathbf{y} = (-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)^T$$

a. Calculate (by hand! - use the template in the appendix) the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for a linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$. [5 marks]

b. Describe the resulting regression line:

   - What is the relationship between variables X and Y?
   - How much (on average) does Y change when X changes by 1?
   - What value does Y take (on average) when $X = 0$? [3 marks]

c. Now re-fit the model with a software package (R, Stata, SAS, SPSS, . . .) of your choice and verify you get the same estimates. Show the model summary output from your software. [2 marks]

d. Compute the coefficient of determination R^2, the adjusted R^2, the likelihood and the AIC. Which of these tell you how good your model fits the data? [5 marks]

e. Compute the residuals $\epsilon_i = y_i - \hat{y}_i$ and do a normal distribution QQ plot. [5 marks]

f. What other diagnostic check(s) could you do? Do this and explain whether you think this is a good model. [8 marks]

g. Re-fit the model, but now including a term for X^2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Check and discuss the resulting model and compare it to the previous one. Which model would you recommend for this dataset? [12 marks]

Exercise 3 [35 marks]

Download the dataset `titanic.csv` from the course GitHub repository (`https://github.com/gitMarcH/Chanco_ST6103`).

This dataset contains data on survivors of the sinking of the Titanic ocean liner in 1912. It is well known that there were too few lifeboats, that different classes of passengers were housed in different areas of the ship (e.g. from some areas it might have been more difficult to get off the ship) and that there was a 'women and children first' policy for spaces on the rescue boats. So you could expect that this could have had an effect on survival rates.

The variables in the dataset are:

- `class` - What class of fare the individual had paid. Possible values: 1^{st} class (most expensive ticket), 2^{nd} class, 3^{rd} class and crew.
- `age` - Lists whether passengers were adults or children.
- `sex` - Lists whether passengers were male or female.
- `survivors` - Lists the number of survivors for the group defined by `class`, `age`, `sex`.
- `dead` - Lists the number of fatalities in the group defined by `class`, `age`, `sex`.

Build a GLM model where the outcome is `survivor/dead` and predictors are `class`, `age`, `sex`.

You need to decide yourself what is an appropriate distribution and link function, which predictors (and possible interaction terms) to include in the model or not.

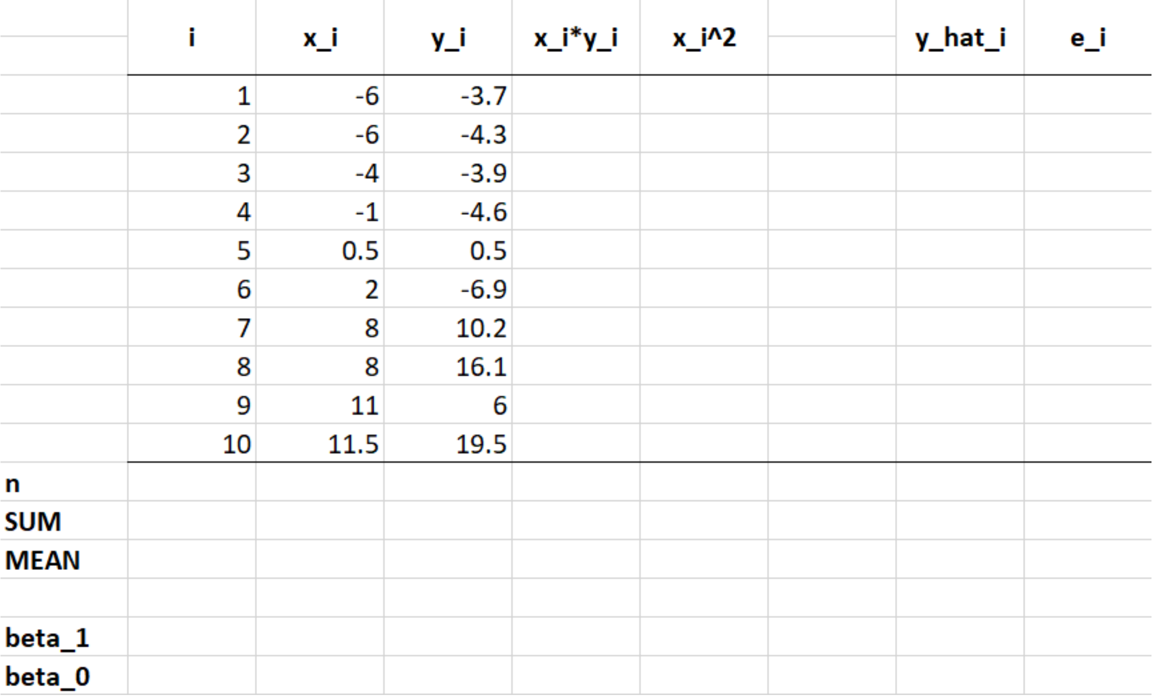
Please report goodness of fit statistics and perform diagnostic checks on the model.

[25 marks]

Once you have a satisfactory model:

- Is `class` predictive of survival? [1 mark]
- Is `age` predictive of survival? [1 mark]
- Is `sex` predictive of survival? [1 mark]
- What is the probability (according to your model) of a female child travelling in second class to have survived the disaster? [3 marks]
- What is the risk ratio of death for crew members compared to 1^{st} class passengers? [4 marks]

Appendix



	i	x_i	y_i	x_i*y_i	x_i^2		y_hat_i	e_i
	1	-6	-3.7					
	2	-6	-4.3					
	3	-4	-3.9					
	4	-1	-4.6					
	5	0.5	0.5					
	6	2	-6.9					
	7	8	10.2					
	8	8	16.1					
	9	11	6					
	10	11.5	19.5					
n								
SUM								
MEAN								
beta_1								
beta_0								