i	y_i	n_i	Age_i	Sex_i	$Class_i$
1					
⋮					

Table 2: Titanic dataset.

(b) Which distribution from the exponential family of distributions would you use to model the responses $y_1, y_2, \cdots$? Explain your answer.

[2 marks]

(c) What is the expression of the likelihood for the saturated model assuming independence of the responses? Justify your answer.

[2 marks]

(d) Give the definition of an indicator variable.

[1 marks]

(e) Explain what mathematical explanatory variables can be defined to represent the information Sex, Age and $Class$.

[4 marks]

(f) Propose three different GLMs to model the mean of the response y_i w.r.t. the explanatory variables that you have defined. Indicate the link function(s) that can be used.

[4 marks]

(g) What criterion can be used to select the best model? Explain your answer.

[3 marks]

(h) What criterion can be used to check that the selected best model is a good model to explain the dataset? Explain your answer.

[3 marks]

(25 marks)

3. According to McCulloch & Searle, *'building a generalized linear model involves three decisions:*

(a) What is the distribution of the data (for fixed values of the predictors and possibly after a transformation)?

(b) What function of the mean will be modeled as linear in the predictors?

(c) What will the predictors be?'

Comment and explain each statement (a,b,c) of this definition of Generalized Linear models (GLMs). In particular define the terms that are underlined in the context of GLMs and give a concrete example of each.

(25 marks)

Exam 2014 DW 29/03/16

1. The Weibull distribution is defined for a continuous random variable y as

$$p_{WB}(y|\lambda,\theta) = \lambda\,\theta\,y^{\lambda-1}\exp\left[-\theta\,y^{\lambda}\right], \quad \text{with } y \geq 0,\ \theta > 0,\ \lambda > 0$$

(a) Explain the connection between the Exponential distribution and the Weibull distribution.

[2 marks]

(b) Show that the expectation $\mathbb{E}[y]$ of the random variable y that follows a Weibull distribution, is

$$\mathbb{E}[y] = \left(\frac{1}{\theta}\right)^{1/\lambda}\Gamma\left(1+\frac{1}{\lambda}\right)$$

with $\Gamma(u) = \int_0^{+\infty} x^{u-1}\exp(-x)\ dx$

[5 marks]

(c) Suggest a link function that can be used with the Weibull distribution. Explain your answer.

[4 marks]

(d) Explain the usage of the Weibull distribution for survival analysis.

[4 marks]

(e) Compute the hazard function for the Weibull distribution. Explain what *accelerated failure time* is.

[6 marks]

(f) In survival analysis, what does it mean that the response y_i of group i is censored?

[4 marks]

(25 marks)

2. The dataset Titanic provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarized according to economic status (classes 1st, 2nd, 3rd or Crew), sex (Male or Female), age (Child or Adult) and survival (survived: No or Yes). 2201 observations have been collected and are summarised in table 1 (subtables a,b,c,d). We want to assess the chance of survival on the Titanic with respect to age, sex and economic status.

Table 1: Titanic dataset.

(a) Age = Child, Survived = No

Class	Sex: Male	Female
1st	0	0
2nd	0	0
3rd	35	17
Crew	0	0

(b) Age = Adult, Survived = No

Class	Sex: Male	Female
1st	118	4
2nd	154	13
3rd	387	89
Crew	670	3

(c) Age = Child, Survived = Yes

Class	Sex: Male	Female
1st	5	1
2nd	11	13
3rd	13	14
Crew	0	0

(d) Age = Adult, Survival = Yes

Class	Sex: Male	Female
1st	57	140
2nd	14	80
3rd	75	76
Crew	192	20

(a) Rewrite the information in table 1 in a new table where the first column corresponds to the group index, the response variable y_i is the number of people who have survived in group i, n_i is the total number of people in group i, Age_i is the age of group i, Sex_i is the sex of group i and $Class_i$ is the economic status of group i (cf. tab. 2).

[6 marks]

9/03/16 ALSM2

EXAM PAPER 2014

Q1 A. Exponential Distribution is a special case of the weibull distribution with $\lambda = 1$

$= (1)(\theta)(y^{1-1}) \exp[-\theta y^1] = \theta \exp[-\theta y]$

B. $\mathbb{E}[y] = \int_0^\infty y \, p(y|\lambda,\theta) \, dy$

$= \int_0^\infty y \lambda \theta y^{\lambda-1} \exp[-\theta y^\lambda] \, dy$

Substitution: $\mu = \theta y^\lambda \quad d\mu = \theta \lambda y^{\lambda-1} dy$

$= \int_0^\infty y \lambda \mu \exp[-\mu] \, d\mu$

$= \int_0^\infty (\mu/\theta)^{1/\lambda} \exp[-\mu] \, d\mu$

$= (\frac{1}{\theta}) \Gamma(1 + 1/\lambda) = \quad \Gamma = \int_0^\infty s^{N-1} \exp(-s) \, ds$

C. $\theta > 0 \quad \mathbb{E}[y] \xi \theta^{-\lambda} \in \mathbb{R}^+$

Log: $\mathbb{R}^+ \to \mathbb{R}$ for $x^T \beta$

$\rightarrow$ suitable link function to map $\mathbb{R}^+$ onto $\mathbb{R}$

The inverse function is exp (link function is invertible)

D. - $y \in \mathbb{R}^+$

- Corresponds to a duration or interpreted as duration
- Suitable for modelling time
- Can compute probability of failure between time 0 and time t by integration

$F(t) = \int_0^T p(y|\theta,\lambda) \, dy \Rightarrow$ Probability of failure between time 0 and T

$S(T) = 1 - F(t)$ probability of survival beyond time t.

E Hazard Function

$F(y) = 1 - \exp(-\theta T^\lambda)$

$S(T) = 1 - F(t) = \exp(-\theta T^\lambda)$

$H(T) = \frac{d \, \text{Log}(S(t))}{d \, (S(T))} = \lambda \theta T^{\lambda-1}$

$H(T) \Rightarrow$ Accelerated failure time $\Rightarrow$ can't do this ~~without~~ with exponential distribution, it is dependent on the choice of λ

$H(T)$ → Probability of failure between time T and $T + \delta T$ having already survived up to T

F Censoring

- Under estimate of how long a person has survived - i.e. has survived longer than the duration of the study
- Time until failure y_i is not known or observed
- Or when person disappears or relocates from the study

- Left censored: origin unknown - don't know start date of disease
- Right censored: didn't die - ~~[illegible]~~ use probability of success $S(T)$

Use the survival function in the likelihood function:

$$S y_i \quad \pi_{i=1}^{n} P(y_i | \theta_i, \lambda) \quad \text{or} \quad S(y_2) S(y_3)$$

- Censoring - condition in which the value of measurement or observation is only partially known ie know age of death is atleast 75 but no more
- May occur when a value is outside the range of measuring instrument - example bathroom weighing scales
- Interval censoring - know value lies between two bounds

? WEEK 1 SEMESTER 2

24/03/16 AWM 2

EXAM PAPER 2014

Q2 A

i	y_i	n_i	Age_i	Sex_i	$Class_i$
1	5	5+0	Child	Male	1st
2	1	1+0	Child	Female	1st
3	11	11+0	Child	Male	1st
⋮					
16	20	20+3	Adult	Female	Crew

B Use binomial or Poisson distribution to model response $y_1, y_2 \ldots y_{16}$

The response variable is binary (0 or 1) with 16 groups

Binomial Distribution can be used to model the survival of each group

C. Binomial distribution given as: $\binom{n}{y}\theta^y(1-\theta)^{n-y}$ for any individual group

$P(y_1, \ldots, y_{16} | \theta_1, \ldots, \theta_{16}) = \prod_{i=1}^{16} p(y_i | \theta_i)$

$= \prod_{i=1}^{16} \binom{n_i}{y_i} (\theta_i)^{y_i} (1-\theta_i)^{n_i - y_i}$

D Indicator Variable:

- A binary variable 0 or 1
- Could also be something like sex male/female
- Use to represent an attribute with two levels like a boolean
- Indicates the presence/non-presence of a variable

E For sex, easiest to define it as a 0,1 event, 0 for male, 1 for female

For age, there are 2 categories, child, adult. Use 0,1 indicator variable

Could use an indicator variable for levels of class

F. Could use logit link function $\theta = \log\left(\frac{1}{1+\theta}\right)$ or probit function

M1: $b_0 + b_1 Sex_i + b_2 Age_i + b_3 Class_i$

M2: $b_0 + b_1 Sex_i + b_2 Age_i^2 + b_2 Class_i^2$

M3 $b_0 + b_1 Sex_i Class_i + b_2 Sex_i + b_3 Sex_i^2$

29/03/16 ALM 2

EXAM PAPER 2014

Q2 A.

i	y_i	n_i	Age_i	Sex_i	$Class_i$
1	5	5+0	Child	Male	1st
2	1	1+0	Child	Female	1st
3	11	11+0	Child	Male	1st
⋮					
16	20	20+3	Adult	Female	Crew

B Use binomial or Poisson distribution to model response $y_1, y_2 \dots y_{16}$

The response variable is binary (0 or 1) with 16 groups

Binomial Distribution can be used to model the survival of each group

C. Binomial distribution given as $\binom{n}{y}\theta^y(1-\theta)^{n-y}$ for any individual group

$P(y_1, \dots, y_{16} | \theta_1, \dots, \theta_{16}) = \prod_{i=1}^{16} p(y_i | \theta_i)$

$= \prod_{i=1}^{16} \binom{n_i}{y_i}(\theta_i)^{y_i}(1-\theta_i)^{n_i-y_i}$

D Indicator Variable:

- A binary variable 0 or 1
- Could also be something like sex male/female
- Use to represent an attribute with two levels like a button
- Indicates the presence/non-presence of a variable

E For sex, easiest to define it as a 0,1 event, 0 for male, 1 for female

For age, there are 2 categories, child, adults. Use 0,1 indicator variable

Could use an indicator variable for levels of class

F. Could use logit link function $\theta = \log\left(\frac{1}{1-\theta}\right)$ or probit function

M1: $b_0 + b_1 sex_i + b_2 age_i + b_3 class_i$

M2: $b_0 + b_1 sex_i + b_2 age_i^2 + b_2 class_i^2$

M3: $b_0 + b_1 sex_i{:}class_i + b_1 sex_i + b_1 sex^2$

G. AIC - Akaike Information Criterion = $-2\log(L) + 2m$
- Where L is the likelihood function of the data as mentioned before
- m is the # of parameters estimated in the model
- Rewards model accuracy/fit while penalising complexity by # of parameters
- Choose lowest AIC value

H. Look at the deviance of the model
This is roughly equal to the likelihood ratio of $\frac{\text{Saturated Model}}{\text{Given Model}}$
This should be near 0 if it is a good model

Compare D_{calc} against a χ^2 dist with $(n-m, n=0)$ DF.
If this D value is below $D_{0.95}$, it is a good model

7 WEEK 1 SEMESTER 2

9/03/16 ALSM2 1

EXAM PAPER 2014

Q3 A. Distribution of Data

- Is the data a binary outcome? i.e. M/F yes/No live/die?
- Is the data studying a time until survival such as in an experiment for a drug dose
- Are we modelling a time until failure or are we modelling the number of occurances of an event.

- First, the dependent variable of interest may have a non-continuous distribution, and thus, the predicted values should also follow the retrespective distribution; any other predicted values are not logically possible
- For example, a researcher may be interested in predicting one of three possible discrete outcomes. In this case, the dependent variable can take only 3 distinct values, and the distribution of the dependent variable is said to be multinomial.
- Or suppose you are trying to predict peoples family planning choices, specifically how many children, as a function of income and various other indicators
- The dependent variable - #children is discrete and most likely the distribution of that variable is highly skewed. In this case, would be reasonable to assume dependent variable follows a poisson distribution

Poisson - # of occurances $y \in \mathbb{N}$

Binomial - # of successes in n trials $y \in \{0,1,\dots,n\}$

Exponential - time until failure → survival analysis $y \in \mathbb{R}^+$

Weibull - time until failure $y \in \mathbb{R}^+$

Multinomial - Binomial but with multiple choice outcomes

B What function of the mean will be modelled as linear in the predictors?

- Choice of link function ⇒ must be invertible
- A second reason why the linear (multiple regression) model might not be inadequate to describe a particular relationship is that the effect of the predictors on the dependent variable may not be linear in nature
- For example, the relationship between a persons age and various indicators of health is most likely not linear in nature: During early adulthood

PTO.

the average health status of people who are 70 compared to that of someone who is 60 is markedly different

- However, health difference between 60-70 is probably greater
- thus, the relationship is non-linear in nature.
- the link between age and health status is best described as non-linear or as a power relationship in this example.

logit: $\log\left(\frac{1}{1-\mu}\right)$ probit: invnorm(z) or log

identity $\mathbb{R} \Rightarrow \mathbb{R}$ Square and square root

- A smooth invertible linearising link function $g()$ which transforms the expectation of the response under $\mu_i = E[y_i]$ to a linear predictor
- Provides the relationship between linear predictor and the mean of the distribution function

C What will the Predictors Be?

- The predictors used requires some fine tuning
- As a basis, all predictors should be included as linear terms. Calculate AIC.
- Create interactions between predictors or use the square of a predictor to attempt to create a more accurate model.
- Calculate AIC for these models
- Choose Model with lowest AIC
- Investigate Deviance value to determine if model provides a good fit or not

limitation of likelihood function - assumes responses are independent

- outliers in data can give a zero, all data except one group voting for parameter B
- May not be robust because of outliers.
- Hard to remove outliers in large dataset
- Same problem with bayesian approach