Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Assessment Submission Form

Student Name	Vishal Kumar
Student ID Number	19304373
Course Title	MSc. Computer Science- Intelligent Systems
Module Title	Applied Statistical Modelling
Lecturer(s)	Dr. Arthur White
Assessment Title	Assignment 2
Date Submitted	14-05-2020
Word Count	

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at: http://www.tcd.ie/calendar
I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write
I declare that the assignment being submitted represents my own work and has not been taken from the work of others save where appropriately referenced in the body of the assignment.

Signed: 

Date: 14-05-2020

Question 1

Recall Question 2 for Assignment 1, where a company wished to model online traffic for a website using a Poisson distribution. Recall that a Poisson distribution with rate parameter θ has pdf

$$f(y|\theta) = \theta^y \frac{\exp\{-\theta\}}{y!}.$$

The company records covariates $x = x_1, \ldots, x_P$, that they would like to use to predict the number of unique visits that their website home page receives. The predictor variables could include, e.g., website layout, time of day, social media activity, etc., but their specific form is not important.

a) Express the pdf $f(y|\theta)$ in exponential family form, that is, in the form:

$$f(y|\theta) = h(y)g(\theta)\exp\{\phi(\theta)s(y).\}$$

Explicitly identify h, g, ϕ and s. [5 marks]

→ General exponential family of distribution form is given by:

$f(y|\theta) = h(y)\exp\{\phi(\theta)S(y) - a(\theta)\}$, where $h(y)$ is underlying measure
$a(\theta)$ is log-normalizer
$\phi(\theta)$ is natural parameter
$S(y)$ is sufficient statistic

Given, $f(y|\theta) = \theta^y \frac{\exp\{-\theta\}}{y!} = \frac{1}{y!}\exp\{-\theta\}\theta^y$ {where $\theta > 0$ & $y = 1, 2, \ldots$}

$= \frac{1}{y!}\exp\{-\theta\}\theta^y$

$= \frac{1}{y!}\exp\{-\theta\}\exp\{\log(\theta^y)\}$

$\Rightarrow f(y|\theta) = \frac{1}{y!}\exp\{y\cdot\log(\theta) - \theta\}$

∴ Comparing above equation with standard form, we get:

$h(y) = \frac{1}{y!}$ $\phi(\theta) = \log(\theta)$ $S(y) = y$ $a(\theta) = \theta \Rightarrow g(\theta) = \exp(-\theta)$

When we have enough samples, y_i $i = 1, 2, \ldots n$, this PDF becomes as:

$f(y|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\exp\{-\theta\}\theta^{y_i}$

$= \frac{1}{\prod_{i=1}^{n} y_i!}\exp\{-n\theta\}\theta^{\sum y_i} = \frac{1}{\prod_{i=1}^{n} y_i!}\exp\{\sum y_i \log(\theta) - n\theta\}$

$$f(y|\theta) = \frac{1}{\prod_{i=1}^{n} y_i!} \exp\{\sum y_i \log(\theta) - n\theta\}$$

Here, $h(y) = \frac{1}{\prod_{i=1}^{n} y_i!}$ is Normalizing constant

$\phi(\theta) = \log(\theta)$ is Natural parameter

$S(y) = \sum y_i$ is Sufficient statistic

$a(\theta) = n\theta \Rightarrow g(\theta) = \exp\{-n\theta\}$ is link function / log normalizer

b) Explain how the natural parameter ϕ can be used as a link function to incorporate a single predictor variables x into a Poisson regression model. Express the rate parameter θ as function of x , i.e., $\phi^{-1}(x)$. [5 marks]

General canonical form of eqn from Q.1a ,

$$f(y|\eta) = \frac{1}{y!} \exp\{\eta \cdot y - a(\eta)\}$$ where, η is natural parameter

$\eta = \phi(\theta) = \log(\theta)$

$a(\eta) = e^{\eta} = \lambda$

$$E(S(y)) = E(y) = \frac{d}{d\eta} a(\eta) = \frac{d}{d\eta}(e^{\eta}) = e^{\log(\theta)} = \theta$$

$$\Rightarrow \boxed{E(y) = \theta} \quad \text{——— (1)}$$

Now defining link function for eqn (1),

$$\phi(\theta_i) = \eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n = x_i' \beta$$

$$\Rightarrow \boxed{\theta_i = \phi^{-1}(x_i' \beta)} \quad \text{——— (2)}$$

Now, we know that, 'λ' is variance & mean of Poisson distribution

∴ for observation data y_i, $i = 1, 2 \cdots n \Rightarrow E(y_i) = \lambda$

∴ $y_i = E(y_i) + \varepsilon_i$, $i = 1, 2 \cdots n$, where ε_i are disturbance terms.

Using eqn (2), Identity link function becomes:

$$\theta_i = \phi^{-1}(x_i' \beta)$$

$\Rightarrow \phi(\theta_i) = x_i'\beta$

Taking log on both sides of above link function

$\Rightarrow \boxed{\theta_i = \exp(x_i'\beta)}$

c) Use the link function $\phi(x)$ to specify the log-likelihood of the Poisson regression model. Compute the first derivative of the model with respect to β_1, the regression coefficient for x. [5 marks]

We have, $P(y_i|x_i) = \dfrac{\exp\{-\theta_i\}\cdot\theta_i^{y_i}}{y_i!}$ ——— (1)

and, from previous Q1.b :
we have, log link function, $\theta_i = \exp\{x_i\beta\}$

Since, y_i & x_i are independent random variable,

Calculating joint probability from equation (1)

$$P(Y|X) = \prod_{i=1}^{n}\left(P(y_i|x_i)\right) = \prod_{i=1}^{n}\frac{\exp\{-\theta_i\}\cdot\theta_i^{y_i}}{y_i!} = L(\beta) \quad \text{——— (2)}$$

Now taking log likelihood of equation (2)

$$\Rightarrow \log(L(\beta)) = \sum_{i=1}^{n}\left(-\theta_i + y_i\log(\theta_i) - \log(y_i!)\right)$$

using $\theta_i = \exp\{x_i\beta\}$

$$\Rightarrow \log(L(\beta)) = \sum_{i=1}^{n}\left(-\exp\{x_i\beta\} + y_i\cdot x_i\beta - \log(y_i!)\right)$$

$$= \sum_{i=1}^{n}\left(y_i x_i\beta - \exp\{x_i\beta\} - \log(y_i!)\right)$$

Now differentiating w.r.t β,

$$\Rightarrow \frac{d}{d\beta}\left(\log(L(\beta))\right) = \sum_{i=1}^{n}\left(y_i x_i - \exp\{x_i\beta\}\cdot x_i\right)$$

$$= \sum_{i=1}^{n}\left(y_i - \exp\{x_i\beta\}\right)\cdot x_i \quad \text{——— (3)}$$

And, to calculate max-log-likelihood, equating 1st derivative to zero,

$$\Rightarrow \frac{d}{d\beta}\left(\log(L(\beta))\right) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} (y_i - \exp\{x_i\beta\}) \cdot x_i = 0}$$

d) The company fit a Poisson regression model to data using a single covariate x. This variable recorded which of two advertising strategies (Strategy 0 or Strategy 1) were used on the preceding day. Traffic to the website between 12 and 12:15pm was then recorded for the following day.

The following output from the regression was produced:

```
Call:
glm(formula = y ~ strategy, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.64317  -0.81635   0.09564   0.59717   1.49792

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3001     0.1924   1.559    0.119
strategy1     1.6094     0.2108   7.634 2.27e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 120.941  on 39  degrees of freedom
Residual deviance:  42.344  on 38  degrees of freedom
AIC: 154.09

Number of Fisher Scoring iterations: 5
```

Interpret this output. Which strategy is better, and by how much? [5 marks]

From given Regression Output, we can interpret the following:

→ Number of Fisher Scoring iteration = 5, signifies that 5-iteration were performed to fit the model and re-weight the model parameters.

→ Null deviance = 120.941
Residual deviance = 42.344
} i.e Null deviance > Residual deviance which signifies that model improvement upon including 1 out of 39 predictors, which can be inferred from DF (degree of freedom).

→ Residual deviance: 42.344 on 38 degree of freedom
i.e residual deviance is greater than Degree of freedom.
It represents: — Correctness of model estimation.
— Smaller Standard error than expected.

→ In Strategy 1, P-value is too small that even using quasi-poisson method will make significant predictor. Signif-Codes is *** which indicates p-value is significant.

→ Strategy-0 is better, because, poisson regression by default uses log-link function that helps in exponential increase in traffic of website between 12 and 12:15 pm.
And estimated increase will be by: exp{1.6094}

e) The company have another data set that they would like to analyse using Poisson regression. For GDPR reasons, they are unwilling to release individual responses $y_1, \ldots, y_n$.

(The same restrictions do not apply to predictor variables x in this case.) Data summaries for y are allowed, however, since these do not reveal sensitive information to the same extent. Using your answer from part c), explain what data summaries are required to perform a Poisson regression. [5 marks]

From Q1.c we have $\sum_{i=1}^{n} (y_i - \exp\{x_i\beta\}) \cdot x_i = 0$

i.e It is clear that log-likelihood is maximized for coefficient-β when 1st derivative is equal to 0.

And,
To perform a poisson regression, need to find coefficient β of model. we need only $(y^T x)$ matrix data.

Question 2

Two teams of epidemiologists are trying to model a disease outbreak. The goals of each team, are similar, but different:

- Team A wish to predict the number of cases as accurately as possible, with a view to ensuring that enough public resources (e.g., hospital beds, protective equipment, etc.) are allocated to manage the disease as effectively as possible.
- Team B would like to identify the variables that appear to be most influential at driving the disease; their goal is to inform policy changes (e.g., closing businesses, restricting travel, administering medicines, etc.) that could lead to a reduction in the number of disease cases that are being observed.

Both teams use a similar disease model, which can take a large number of covariates as input to the model. Neither team expects that all inputs are necessary for their models, but they are unsure about which models to include.

Using the different paradigms for model selection discussed in class (prediction, understanding and regularisation) and their associated methods (AIC, cross-validation, BIC, Lasso, etc.), discuss which approaches might be most suitable for Team A and B respectively. (Suggested word limit: 100-150 words.)

[15 marks]

Team A aims for accurate prediction for number of potential cases. So, team A might use prediction or understanding models with methods as AIC or cross-validation that performs better than BIC method in making prediction. Because, Cross-validation do not assume anything about model, directly generates test errors and is more effective in comparing different models and, AIC helps in measuring unknown true likelihood with fitted likelihood of data along with constant term.

Team B requires to identify most influential parameters. So team B could use regularisation model with BIC method because, they reduce variance among selected parameters thus reducing model variability. They remove least important parameters by penalising strongly or reduces model size by retaining most important parameters only. BIC would better in this situation because here a false positive would be as misleading, or more than, a false negative.