Semester 1	Generalized Linear Models - Revision exercise	2016

1. ***Estimation.*** Consider the Pareto model for survival times $X_1, \ldots, X_n$:

$$X_i|\ \lambda, \alpha \overset{\text{ind}}{\sim} \text{Pareto}(\lambda, \alpha);$$

where $\lambda > 0$ is the *fixed* scale parameter and $\alpha > 0$ is the shape parameter. The probability density function (pdf) is

$$f(x) = \frac{\alpha\lambda^\alpha}{(x+\lambda)^{\alpha+1}},\ x \geq 0$$

which can be equivalently written in terms of $y = x + \lambda$ as

$$f(y) = \frac{\alpha\lambda^\alpha}{y^{\alpha+1}},\ y \geq \lambda \tag{1}$$

and the cumulative distribution function (cdf) is

$$F(y) = 1 - \left(\frac{\lambda}{y}\right)^\alpha, \quad y \geq \lambda. \tag{2}$$

(a) Show that $\log\left(\frac{Y}{\lambda}\right) \sim \text{Exp}(\alpha)$ where $\text{Exp}(\alpha)$ is the exponential distribution with parameter α.

(b) Show that $E(Y^h) = \frac{\alpha\lambda^h}{\alpha - h}$ if $\alpha > h$. Hence show that $Var(Y) = \frac{\mu^2}{\alpha(\alpha-2)}$ if $\alpha > 2$ where $\mu = E(Y)$.

(c) Based on a given λ and an iid sample $y_1, \ldots, y_n$ with sample mean $\bar{y}$,

(i) Show that the moment estimate of α is $\hat{\alpha}^{(M)} = \frac{\bar{y}}{\bar{y} - \lambda}$.

(ii) Show that the maximum quasi-likelihood estimate (MQLE) for μ is the sample mean. Explain why the MQLE for α can not be obtained.

(iii) Show that the maximum likelihood estimate (MLE) for α is given by

$$\hat{\alpha}^{(\text{MLE})} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \log y_i - \log\lambda}.$$

(d) A distribution belongs to the exponential family if its pdf can be written as:

$$f(y) = \exp\left\{\frac{T(y)\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{3}$$

where $T(y)$ is the sufficient statistic, θ is the natural parameter, and $b(\theta)$ is the normalization factor.

Show that Pareto distribution with fixed λ is a member of exponential family by identifying θ, $T(y)$, $a(\phi)$, $b(\theta)$ and $c(y,\phi)$ in (3).

Hence find $E(T(Y))$ and $Var(T(Y))$. Verify your answer for $E(T(Y))$.

2. ***Estimation.***

(a) Find a function g such that $\mathrm{Var}(g(X))$ is approximately constant when $X = Z/n$, where $Z \sim \mathrm{Bin}(n,p)$.

(b) Show that $\sum_{i=1}^{n} h_{ii} = p$, where $h_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$ and $\mathrm{rank}(\boldsymbol{X}) = p$.

(c) Let the correlation matrix for a set of exchangeable random variables be $\boldsymbol{\Sigma} = (1-\alpha)\boldsymbol{I} + \alpha\mathbf{1}\mathbf{1}'$. You are given that the inverse of the matrix $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1-\alpha}[\boldsymbol{I} - k\mathbf{1}\mathbf{1}'],$$

where $k = \alpha/[1+(n-1)\alpha]$. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, .., n$$

where the ϵ_i are assumed to be zero mean, exchangeable random variables. Show that the WLS estimate for the slope is just the simple ordinary least squares estimate.
[Hint: use the Bartlett's Identity: $(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}')^{-1} = \boldsymbol{A}^{-1} - (\boldsymbol{A}^{-1}\boldsymbol{u})(\boldsymbol{v}'\boldsymbol{A}^{-1})/(1+\boldsymbol{v}'\boldsymbol{A}^{-1}\boldsymbol{u})$ to find the inverse.]

3. ***Estimation.***

(a) For the Poisson variables with link function $\mu_i^2 = \eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, prove that

$$\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0.$$

Hence prove that the deviance is

$$2\sum_{i=1}^{n} y_i \ln(y_i/\hat{\mu}_i).$$

(b) Suppose that Y_1 and Y_2 are independent Poisson random variables such that $Y_1 \sim \mathcal{P}(\mu)$ and $Y_2 \sim \mathcal{P}(\rho\mu)$. Show that

$$\begin{aligned} Y_1 + Y_2 &\sim \mathcal{P}(\mu + \rho\mu) \\ Y_1 |\ Y_1 + Y_2 = m &\sim \mathcal{B}(m, \frac{1}{1+\rho}) \end{aligned}$$

Show how you might use this result to test the composite null hypothesis $H_0 : \rho = 1$ against the one-sided alternative $H_1 : \rho > 1$.

4. ***Estimation.*** Let Y be a random variable with mean μ and variance given by

$$\mathrm{Var}(Y) = \mu(\mu + \alpha)/\alpha$$

for some $\alpha \geq 0$.

(a) Show that the quasi-likelihood for this case is

$$Q(\mu, y) = y \ln \frac{\mu}{y} - (y + \alpha) \ln \frac{\mu + \alpha}{y + \alpha}.$$

(b) Hence show that the maximum quasi-likelihood estimator for μ based on a sample of n independent observations is the sample mean.

(c) Obtain the defining equation for the maximum quasi-lielihood estimation for α.

5. ***Logit model.*** Consider the following data collected by Erickson (1987) as part of a study on the measurement of anaesthetic depth. The potency of an anaesthetic agent is measured in terms of the minimum alveolar concentration (MAC) of the agent at which 50% of patients exhibit no response to stimulation (i.e. do not move - moving means jerking or twisting, not twitching or grimacing - in response to a surgical incision). Thirty patients were adminstered an anaestheic agent which was maintained at a predetermined alveolar concentration (actually, anaesthetists refer to concentration when they mean partial pressure - hence alveolar concentration is measured as a percentage of one atmosphere) for 15 minutes before a singel incision was made in each patient. For each patient, the alvecolar concentration (AC) of the anaesthetic agent and the patient's response to incision (1 for 'no move' and 0 for 'move') was recorded.

Patient i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Resp. y_i	1	0	1	0	0	1	1	0	1	0	0	1	1	1	1
AC x_i	1.0	1.2	1.4	1.4	1.2	2.5	1.6	0.8	1.6	1.4	0.8	1.6	2.5	1.4	1.6

Patient i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Resp. y_i	1	1	0	1	1	0	0	0	0	0	1	0	1	0	1
AC x_i	1.4	1.4	0.8	0.8	1.2	0.8	0.8	1.0	0.8	1.0	1.2	1.0	1.2	1.0	1.2

(a) Fit a logistic regression model to this data and assess its fit, i.e. fit $\ln[\pi_i/(1-\pi)] = \beta_0 + \beta_1 x_i$ where x_i represents the alveolar concentration and π_i is the probability of movement in the ith patient.

(b) Use the model to estimate the MAC value.

(c) Express your MAC estimator as a function of β_0 and β_1, $h(\boldsymbol{\beta})$. Recalling

$$\mathrm{Var}[h(\widehat{\boldsymbol{\beta}})] \simeq h'(\widehat{\boldsymbol{\beta}})^T \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) h'(\widehat{\boldsymbol{\beta}})$$

where $h'(\boldsymbol{\beta}) = \Big(\dfrac{\partial h}{\partial \beta_0}, \dfrac{\partial h}{\partial \beta_1}\Big)^T$. Obtain an approximate standard error for your estimator in (b).

6. ***Survival analysis.*** Dellaportas and Smith (1993) analysed data from Grieve (1987) on photocarcinogenicity in four groups, each containing 20 mice, who have recorded a survival time and whether they died or were censored at that time. A portion of the data, giving survival times in weeks for two groups, Irradiated Control (IC) and Vehicle Control (VC) are shown below. A * indicates censoring.

IC	12	1	21	25	11	26	27	30	13	12	21	20	23	25	23	29	35	40*	31	36
VC	32	27	23	12	18	40*	40*	38	29	30	40*	32	40*	40*	40*	40*	25	30	37	27

A survival model with Weibull distribution for the lifetime T is fitted as below.

```
> t1=c(12,1,21,25,11,26,27,30,13,12,21,20,23,25,23,29,35,40,31,36)
> t2=c(32,27,23,12,18,40,40,38,29,30,40,32,40,40,40,40,25,30,37,27)
> w1=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1)
> w2=c(1,1,1,1,1,0,0,1,1,1,0,1,0,0,0,0,1,1,1,1)
> t=c(t1,t2)
> w=c(w1,w2)
> c=factor(c(rep(0,20),rep(1,20)))
> l=log(t)
> omega=sum(w)
> alpha=1
> c=as.numeric(c)
> iter=15
> result=matrix(0,iter,3)
> for (i in 1:iter) {
+ par=glm(w~offset(alpha*l)+c,family=poisson)$coeff
+ beta0=par[1]
+ beta1=par[2]
+ result[i,]=c(alpha,par[1],par[2])
+ mu=t^alpha*exp(beta0+c*beta1)
+ sum1=sum((mu-w)*log(t))
```

```
+ alpha=0.5*(alpha+omega/sum1)
+ }
> colnames(result)=c("alpha","beta0","beta1")
> result
         alpha     beta0      beta1
 [1,] 1.000000 -2.481399 -0.7075598
 [2,] 2.326261 -6.562449 -1.0109130
 [3,] 2.510650 -7.146599 -1.0475881
 [4,] 2.567484 -7.327335 -1.0586655
 [5,] 2.585766 -7.385543 -1.0622067
 [6,] 2.591718 -7.404499 -1.0633572
 [7,] 2.593663 -7.410694 -1.0637329
 [8,] 2.594300 -7.412721 -1.0638558
 [9,] 2.594508 -7.413385 -1.0638960
[10,] 2.594576 -7.413602 -1.0639092
[11,] 2.594598 -7.413673 -1.0639135
[12,] 2.594606 -7.413696 -1.0639149
[13,] 2.594608 -7.413704 -1.0639153
[14,] 2.594609 -7.413706 -1.0639155
[15,] 2.594609 -7.413707 -1.0639155
```

(a) State the shape parameter α and the scale parameter λ for the Weibull distribution.

(b) Write down the hazard function $h(t)$ and survival function $S(t)$ for the IC group and compare them with the VC group. Estimate the hazard and survival probability for the IC group when $t = 20$.

7. ***Survival analysis.*** Let $t_i,\ i = 1, \ldots, n$ be a set of censoring or survival times and ω_i the corresponding non-censoring indicators. Consider a composite hazard function $h(t) = a + b(t-\tau)^2 I(t)$, that is,

$$h(t) = \begin{cases} a, & 0 \le t < \tau, \\ a + b(t-\tau)^2, & t \ge \tau, \end{cases}$$

where $a > 0$ and $b > 0$ are unknown parameters, $I(t) = I(t \ge \tau)$ and τ is fixed.

(a) Show that the cumulative hazard function is given by

$$H(t) = at + \frac{b}{3}(t-\tau)^3 I(t).$$

Hence find the survival function $S(t)$ and the pdf $f(t)$.

(b) Define the sum of squares as

$$SS(a,b) = \sum_{i=1}^{n} [H_0(t_i) - H(t_i)]^2$$

where the observed cumulative hazard $H_0(t_i) = -\ln S_0(t_i)$ is based on

$$S_0(t_i) = \prod_{j:t_j \le t_i} \left(1 - \frac{d_j}{r_j}\right)$$

when there are d_j failures among the r_j subjects at risk at time t_j.

If we write $SS(a,b) = \sum_i (y_i - az_{1i} - bz_{2i})^2$ for a suitable linear model $y_i = az_{1i} + bz_{2i}$, identify y_i, z_{1i} and z_{2i} in terms of $H_0(t_i)$, observations and parameters.

(c) Find the least square estimates of a and b and describe how their standard errors can be obtained.

(d) A sample of $n = 49$ Kelvar Epoxy Spherical pressure vessels were subjected to constant sustained pressure at the 70% stress level until all had failed, so that the complete data with exact failure time (in thousands hours) are given as below:

1.051	1.337	1.389	1.921	1.942	2.322	3.629
4.006	4.012	4.063	4.921	5.445	5.620	5.817
5.905	5.956	6.068	6.121	6.473	7.501	7.886
8.108	8.546	8.666	8.831	9.106	9.711	9.806
10.205	10.396	10.861	11.026	11.214	11.362	11.604
11.608	11.745	11.762	11.895	12.044	13.520	13.670
14.110	14.496	15.395	16.179	17.092	17.568	17.568

Fit the model in (a) using the methods in (b) and (c), setting $\tau = 3.4$. Plot the hazard function, cumulative hazard function and survivial functions for the model.

8. ***Survival analysis.*** The Gompertz distribution has a hazard function $h_0(t) = bc^t,\ t \ge 0,\ b > 0,\ c > 1$.

Let t_i denote failure or censored time for patient i and ω_i the non-censoring indicator. The observations t_i come from two treatments, indicated by $x_i = 0$ and $x_i = 1$, respectively. To allow for the treatment effects, the hazard function $h(t_i)$ is further multiplied by $\exp(\beta x_i)$.

(a) Find the hazard function $h(t)$, cumulative hazard function $H(t)$ and survival function $S(t)$ for the Gompertz distribution.

(b) Conditional on c and β, show that maximum likelihood estimator (MLE) for b is

$$\hat{b} = \frac{\ln c \sum_{i=1}^{n} \omega_i}{\sum_{i=1}^{n} (c^{t_i} - 1) \exp(\beta x_i)}$$

(c) Derive the Newton Raphson (NR) iterative procedures for $\hat{c}$ and $\hat{\beta}$ conditional on b.

(d) (More challenging; Optional) Write a R program to fit the model to the Leukaemia data and estimate the MLE for b, c and β iteratively.

(e) (More challenging; Optional) Plot $h(t)$, $H(t)$ and $S(t)$ for the two groups. You may compare the plots with those from the models of exponential and Weibull distributions.

9. ***Survival analysis.*** For a sample of n subjects, the time t_i is a lifetime if $w_i = 1$ indicates death or a censored time if $w_i = 0$ indicates termination of observation. The corresponding covariate for t_i is z_i.

(a) An *additive* hazard model is defined as:

$$h(t_i) = \beta_0 + \beta_1 z_i.$$

Write down the log-likelihood function ℓ using

$$\ell = \sum_{i=1}^{n} [w_i \ln h(t_i) - H(t_i)].$$

Hence show that the parameter estimates of $\beta_j,\ j = 0, 1$ are solution to the equations:

$$\sum_{i=1}^{n} \frac{w_i z_{ij}}{\beta_0 + \beta_1 z_{i1}} - \sum_{i=1}^{n} t_i z_{ij} = 0, \quad j = 0, 1 \tag{4}$$

where $z_{i0} = 1$ and $z_{i1} = z_i$.

(b) A *linear* hazard model from Rayleigh distribution is defined as:

$$h(t_i) = \beta_0 + \beta_1 t_i.$$

Show that the parameter estimates of $\beta_j,\ j = 0, 1$ are solution to the equations:

$$\sum_{i=1}^{n} \frac{w_i z_{ij}}{\beta_0 + \beta_1 z_{i1}} - \sum_{i=1}^{n} z_{i,j+1} = 0, \quad j = 0, 1 \tag{5}$$

where $z_{i0} = 1$, $z_{i1} = t_i$ and $z_{i2} = \dfrac{t_i^2}{2}$.

(iii) Show that the solution to (5) can be obtained by solving β_1 from:

$$\sum_{i=1}^{n} \frac{w_i}{\sum\limits_{i=1}^{n} w_i - \dfrac{\beta_1}{2}\left(\sum\limits_{i=1}^{n} t_i^2\right) + \beta_1 t_i \left(\sum\limits_{i=1}^{n} t_i\right)} = 1$$

using `uniroot` in R or Newton-Raphson procedure and β_0 can be calculated from:

$$\beta_0 = \frac{\sum\limits_{i=1}^{n} w_i - \dfrac{\beta_1}{2} \sum\limits_{i=1}^{n} t_i^2}{\sum\limits_{i=1}^{n} t_i}.$$