Under-Reporting of Fire Starts: Supporting Evidence

Neil Diamond

November 30, 2011

Introduction

I recently prepared a report entitled "Under-reporting of Fire Starts" for United Energy. In that report I apply a probability model for underreporting to the United Energy dataset. Based on that model I estimate the actual number of fire starts from 2006-2010 to be 940, higher than the recorded number of 561. I have been asked by the Australian Energy Regulator to prepare some supporting material so that they can consider my report. The material requested is given below. I would be very happy to answer any questions about the report or this document. I can be contacted at neil.diamond@monash.edu.

Summary of supporting research and analysis

Literature Review and Theoretical Considerations

The material below is a summary of the published paper by Neubauer, G., Djuras, G., and Fiedl, H. (2011), on which my analysis is based.

Assume $y_t,\ t = 1, \ldots, T$ are a sample of fire starts that are reported at time t and that each time there are the same unknown number λ of fire starts that actually happened. For each fire start a random mechanism decides whether it is reported or not i.e. there is a constant probability π of

reporting the fire start. Here we have a a random Bernoulli variable[1]

$$R_i = \begin{cases} 1 & \text{if fire start reported} \\ 0 & \text{if fire start not reported} \end{cases}$$

so that

$$Y_t = \text{Actual Number of Fire Starts} = \sum_{i=1}^{\lambda} R_i \sim \text{binomial}(\lambda, \pi)$$

Note that

$$\begin{aligned} E(Y_t) &= \mu = \lambda\pi \\ Var(Y_t) &= \mu(1-\pi) = \mu\phi,\ 0 \leq \phi \leq 1. \end{aligned}$$

More realistically $E(Y_t) = \mu_t = \lambda_t\pi$ is allowed to vary with

$$\lambda_t(\beta) = \exp(x^t\beta)$$

where x corresponds to explanatory variables. (For example, I have used mean maximum temperature and total monthly rainfall in the previous month). The Likelihood is given by

$$L(\alpha, \beta | y_t, x_t) = \binom{\lambda_t(\beta)}{y_t} \pi(\alpha)^{y_t} (1 - \pi(\alpha))^{\lambda_t(\beta) - y_t}$$

This model is not appropriate because $Var(Y_t) \leq \mu_t$.

Possible models that would be appropriate include those that allow either λ to vary (over and above the variation due to the variation in the explanatory variables) or to allow π to vary. In the next section some of these models will be outlined.

Allowing π to have a distribution

If we allow π to vary then $Y_t|P \sim$ binomial(λ, p) and $P \sim$ beta(γ, δ). We then obtain the beta-binomial distribution as the marginal distribution of Y_t. Note that

$$Var(Y_t) = \mu(1-\pi)(\lambda + \gamma + \delta)/(1 + \gamma + \delta) = \mu\phi,\ \phi > 0.$$

[1]Details of the various probability distributions used are given in the Appendix.

A re-parameterisation is used with $\theta = \gamma + \delta$ and $\pi = \gamma/\delta$ and with $\lambda_t(\beta) = \exp(x^T\beta)$ and $\pi(\alpha) = \exp(\alpha)/[1+\exp(\alpha)]$, the profile likelihood contribution is

$$L(\alpha, \beta | y_t, x_t, \theta) = \binom{\lambda_t(\beta)}{y_t} \frac{\mathcal{B}(y_t + \pi(\alpha)\theta, \lambda_t(\beta) - y_t + (1-\pi(\alpha))\theta)}{\mathcal{B}(\pi(\alpha)\theta, (1-\pi(\alpha))\theta)}$$

where $\mathcal{B}$ is the beta function and $\gamma(\alpha) = \pi(\alpha)\theta$ and $\delta(\alpha) = (1-\pi(\alpha))\theta$. Maximum Likelihood of α and β given θ is performed and the Method of Moments[2] is used to estimate θ given α and β.

I show that this model is not as good as the reported model later in this document.

Allowing λ to have a distribution

Alternatively, $Y_t|L \sim$ binomial(l, π) and $L \sim$ Poisson(λ) and then $Y_t \sim$ Poisson$(\lambda\pi)$ with $Var(Y_t) = \mu$ i.e $\phi = 1$. Allowing randomness in λ, $L_t|K \sim$ Poisson$(k\lambda_t)$ and assuming $K \sim$ Gamma(ω, ω) we obtain a negative binomial distribution with parameters ω (but in this case ω is the expected number of unreported cases) and π. Here

$$Var(Y_t) = E(Y_t) + E(Y_t)^2/\omega = \mu\phi \ \phi \geq 1.$$

With $\omega_t(\beta) = \exp(x^T\beta)$ and $\pi(\alpha) = \exp(\alpha)/(1+\exp(\alpha))$, the Likelihood contribution is

$$\binom{\omega_t(\beta) + y_t - 1}{y_t} \pi(\alpha)^{y_t}(1-\pi(\alpha))^{\omega_t(\beta)}.$$

I show that this model is not as good as the reported model later in this document.

Let Y be a discrete random variable, taking only non-negative values. Y follows a generalised Poisson distribution if its probability distribution function is given by:

$$p(y|\theta, \tau) = \begin{cases} (1/y!))\theta(\theta + y\tau)^{y-1}e^{-\theta-y\tau} & y = 0, 1, 2, \ldots \\ 0 & \text{if } y > m, \text{ when } \tau < 0 \end{cases}$$

[2]Estimation based on equating the sample mean and variance to the population mean and variance.

where $\theta > 0$, $\max(-1, -\theta/m) < \tau$, and $m(\geq 4)$ is the largest integer such that $\theta + m\tau > 0$. Note

$$\begin{aligned} E(Y) &= \theta(1-\tau)^{-1} \\ Var(Y) &= \theta(1-\tau)^{-3} \end{aligned}$$

Note that the θ parameter used for the generalised Poisson has a different meaning to the θ parameter used for the β distribution discussed in the previous section.

Neubauer et al (2011) use a result that the generalised Poisson distribution is "equivalent" to a binomial distribution if $\tau < 0$, equivalent to a Poisson distribution if $\tau = 0$, and "equivalent" to a negative binomial distribution[3] if $\tau > 0$. Equating the moments of the Negative Binomial distribution to the generalised Poisson distribution $\pi = 1-(1-\tau)^2$ and $\lambda_t = \theta_t(\beta)\pi^{-1}(1-\pi)^{-1/2}$ where $\theta_t(\beta) = \exp(x^T\beta)$. Finally, to ensure the mean is positive τ is transformed to α with $\tau(\alpha) = 1 - \exp(-\alpha)$. The likelihood function is given by

$$L(\alpha, \beta | y_t, x_t) = \frac{\theta_t(\beta)[\theta_t(\beta) + y\tau(\beta) + y\tau(\alpha)]^{y-1} \exp[-(-\theta_t(\beta) + y\tau(\alpha))]}{y!}.$$

The estimated value of τ is positive and hence the model is equivalent (or almost equivalent) to a negative binomial distribution. The reported model corresponds to this case of a generalised Poisson distribution.

Allowing both π and λ to have a distribution

Neubauer et al (2011) also use a model where both π and λ have a distribution resulting in a beta-Poisson distribution. I have been unable to get the maximum likelihood estimation of this model to converge.

Other Literature

During the preparation of the report I consulted a number of related papers:

In one of the earliest papers on modelling under-reporting, Winkelmann (1996) used Markov Chain Monte Carlo rather than likelihood methods to estimate the parameters. He used it to study workers' absenteeism data

[3] While the equivalence is exact if τ is zero, it is only approximate, but very close, when $\tau \neq 0$.

from the German Socio-Economic Panel. This method is quite an attractive option and I considered using it, but I preferred to use a more recent article.

Fader and Hardie (2000) derived analytical expressions for the posterior distributions for a simple model where the count process does not depend on explanatory variables. The results are interesting but not directly applicable in the present context.

Hubert, Lauretto, and Stern (2009) give a good description of the Generalised Poisson Distribution (2009).

Source Data, Program Codes, and Output Files

Source Date

The source data is given in "UEcomparisonCFAvMFBTFisher.csv". In addition the MoreMoorabbinMeanMax.csv and MoreMoorabbinRainfall.csv files were used. The rainfall for December 2005 was 81.4 mm.

Program Codes and Output Files

The Program Code is given in firestartsnegbinom.Rnw, which contains the text and R code used in the report. To use it you need to Sweave the report. The R Code itself is in firestartsnegbinom.R. The Output is given below.

```
> ### R code from vignette source 'firestartsnegbinom.Rnw'
>
> ###################################################
> ### code chunk number 1: firestartsnegbinom.Rnw:116-132
> ###################################################
> setwd("\\\\ad.monash.edu/buseco/b01users02/diamond/Documents/
  Stat Consulting/Statistical Consulting/Rothfield/firestarts")
> set.seed(020256)
> require(foreign)
> require(MASS)
> ###United Energy Data
> fire <- read.csv("UEcomparisonCFAvMFBTFisher.csv",header=T,sep=",")
> fire$Date <- chron(as.character(fire$Date),format=list(dates="d/m/y"),
+                    out.format=list(dates="d/m/y"))
> numbfirestarts <- table(cut(fire$Date,"months"))
```

```
> numbfirestarts

Jan 06 Feb 06 Mar 06 Apr 06 May 06 Jun 06 Jul 06 Aug 06 Sep 06 Oct 06 Nov 06 Dec 0
    13      9      5      6      5      1      4      2      4     19     10     3
Feb 07 Mar 07 Apr 07 May 07 Jun 07 Jul 07 Aug 07 Sep 07 Oct 07 Nov 07 Dec 07 Jan 0
    40     13      6      4      2      5      1      4      9      5      4     1
Mar 08 Apr 08 May 08 Jun 08 Jul 08 Aug 08 Sep 08 Oct 08 Nov 08 Dec 08 Jan 09 Feb 0
    10      5      4      6      6      5      3      2     14      7     26     4
Apr 09 May 09 Jun 09 Jul 09 Aug 09 Sep 09 Oct 09 Nov 09 Dec 09 Jan 10 Feb 10 Mar 1
    10      6      3      2      7      6      2      3      6     10     18
May 10 Jun 10 Jul 10 Aug 10 Sep 10 Oct 10 Nov 10 Dec 10
     6      4      4      5      5      4      6     10
> ###################################################
> ### code chunk number 2: firestartsnegbinom.Rnw:143-144
> ###################################################
> plot(numbfirestarts,xlab="Month",ylab="Recorded Number of Fire Starts")
>
>
> ###################################################
> ### code chunk number 3: model2
> ###################################################
> temperature <- as.numeric(unlist(strsplit(
   readLines("MoreMoorabbinMeanMax.csv")[-c(1:11,17)],
+   split=","))[-c(1,14,15,28,29,42,43,56,57,70)])
> temperature
 [1] 27.4 24.8 25.1 18.3 15.7 13.5 14.1 15.7 18.6 21.0 22.0 25.0 27.0 28.8 25.4 22
[18] 13.3 13.5 16.3 17.9 20.5 23.2 26.2 27.1 24.1 26.5 20.3 16.7 15.3 13.7 13.7 18
[35] 22.3 22.2 27.8 27.3 23.7 20.0 16.6 15.0 14.5 16.0 17.7 18.8 25.7 24.8 26.4 27
[52] 21.9 17.1 13.8 13.7 13.6 15.1 19.3 22.1 23.5
> rainfall <- as.numeric(unlist(strsplit(readLines(
  "MoreMoorabbinRainfall.csv")[-c(1:7,13)],
+   split=","))[-c(1,14,15,28,29,42,43,56,57,70)])
> rainfall
 [1]  46.0  80.2  18.4  77.4  63.2  14.0  35.4  35.0  25.6   7.4  24.6  44.6  28.6
[15]  40.8  16.8  50.8  56.0  78.0  26.0  26.4  27.4  73.6 127.2  15.0  26.0  25.4
[29]  44.4  40.6  62.8  61.8  24.8  18.2  50.6  69.2   1.6   1.6  40.6  70.6  17.6
[43]  66.0  48.4  73.6  59.0  81.8  34.4  46.0  32.4  58.4  65.4  40.8  72.6  25.2
[57]  57.2 163.6 113.2 111.2
```

```
> lagrainfall <- rep(0,60)
> lagrainfall[1] <- 81.4
> for (i in 2:60){lagrainfall[i] <- rainfall[i-1]}
>
> mylogl <- function(alphabeta,y=numbfirestarts){
+   thetabeta <- exp(alphabeta[2]+alphabeta[3]*(temperature-mean(temperature))+
+     alphabeta[4]*log(lagrainfall/mean(lagrainfall)))
+     taualpha <- 1-exp(-alphabeta[1])
+  (log(thetabeta)+(y-1)*log(thetabeta+y*taualpha)-
+    (thetabeta+y*taualpha)-log(factorial(y)))
+  }
> require(maxLik)
> test <- maxLik(mylogl,start=c(0,1.93,0.12,-0.28))
> summary(test)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 6 iterations
Return code 1: gradient close to zero
Log-Likelihood: -167.1823
4  free parameters
Estimates:
      Estimate Std. error t value   Pr(> t)
[1,]  0.453977   0.102387  4.4339 9.254e-06 ***
[2,]  1.481167   0.108112 13.7002 < 2.2e-16 ***
[3,]  0.110620   0.016081  6.8789 6.033e-12 ***
[4,] -0.337236   0.054588 -6.1779 6.496e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
> ptest <- 1-(1-(1-exp(-test$estimate[1])))^2
> ptest
[1] 0.5966518
> varcov1 <- solve(-test$hessian)
> varcov1
              [,1]          [,2]          [,3]         [,4]
[1,]  0.0104831906 -0.0075565369 -0.0004182572 0.0003037887
[2,] -0.0075565369  0.0116882938 -0.0001411038 0.0013523018
[3,] -0.0004182572 -0.0001411038  0.0002586012 0.0001705260
```

```
[4,]  0.0003037887  0.0013523018  0.0001705260 0.0029798024
> testsim <- mvrnorm(1000,test$estimate,solve(-test$hessian))
> testsim <- cbind(testsim,1-(1-(1-exp(-testsim[,1])))^2)
> write.csv(testsim,"testsim.csv")
> ptestsim <- 1-(1-(1-exp(-testsim[,1])))^2
> summary(ptestsim)
> numbsim <- matrix(0,1000,60)
> for(i in 1:1000){for(j in 1:60){
+   numbsim[i,j] <- exp(testsim[i,2]+
+     testsim[i,3]*(temperature[j]-mean(temperature))+
+     testsim[i,4]*log(lagrainfall[j]/mean(lagrainfall)))/
+     (testsim[i,5]*sqrt(1-testsim[i,5]))
+   }}
> write.csv(numbsim,"numbsim.csv")
> estnumb <- round(sum(exp(test$estimate[2]+test$estimate[3]*
  (temperature-mean(temperature))+
+ test$estimate[4]*(log(lagrainfall/mean(lagrainfall))))/(ptest*sqrt(1-ptest))))
> estnumb
[1] 940
>
>
>
> ####################################################
> ### code chunk number 4: firestartsnegbinom.Rnw:249-250
> ####################################################
> summary(test)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 6 iterations
Return code 1: gradient close to zero
Log-Likelihood: -167.1823
4  free parameters
Estimates:
      Estimate Std. error t value   Pr(> t)
[1,]  0.453977   0.102387  4.4339 9.254e-06 ***
[2,]  1.481167   0.108112 13.7002 < 2.2e-16 ***
[3,]  0.110620   0.016081  6.8789 6.033e-12 ***
[4,] -0.337236   0.054588 -6.1779 6.496e-10 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
-----------------------------------------
>
>
> ###################################################
> ### code chunk number 5: firestartsnegbinom.Rnw:261-271
> ###################################################
> plot(numbfirestarts,xlab="Month",ylab="Number of Fire Starts",ylim=c(0,80))
> lines(1:60+0.25,(exp(test$estimate[2]+test$estimate[3]*
 (temperature-mean(temperature))+
+ test$estimate[4]*(log(lagrainfall/mean(lagrainfall))))/
  (sqrt(1-ptest))),col=2,lwd=2)
> points(1:60+0.25,(exp(test$estimate[2]+test$estimate[3]*
 (temperature-mean(temperature))+
+ test$estimate[4]*(log(lagrainfall/mean(lagrainfall))))/
  (sqrt(1-ptest))),col=2,pch=16)
> lines(1:60+0.25,(exp(test$estimate[2]+test$estimate[3]*
(temperature-mean(temperature))+
+ test$estimate[4]*(log(lagrainfall/mean(lagrainfall))))/
  (ptest*sqrt(1-ptest))),col=4,lwd=2)
> points(1:60+0.25,(exp(test$estimate[2]+test$estimate[3]*
(temperature-mean(temperature))+
+ test$estimate[4]*(log(lagrainfall/mean(lagrainfall))))/
  (ptest*sqrt(1-ptest))),col=4,pch=16)
>
>
>
> ###################################################
> ### code chunk number 6: firestartsnegbinom.Rnw:291-292
> ###################################################
> hist(1-(1-(1-exp(-testsim[,1])))^2,
  xlab="Probability Fire Start reported",main="")
>
>
> ###################################################
> ### code chunk number 7: firestartsnegbinom.Rnw:301-303
> ###################################################
```

```
> hist(apply(numbsim,1,"sum"),breaks=seq(500,2000,50),
  xlab="Number of Fire Starts (Reported and Unreported)",
  main="")
> abline(v=561,col=2)
>
```

See also `testsim.csv` for the 1000 simulations used to construct the confidence intervals; and `ptestsim.csv` and `numbsim.csv` for the estimated reporting frequency and number of estimated fire starts for each simulation.

Choice of statistical models

I fitted three main models while preparing my report: the beta-binomial and the (usually defined) negative binomial distribution model and the generalised Poisson model. I compared them on the basis of the Bayesian Information Criterion

$$BIC = -2\text{Maximum}(\text{Log-Likelihood}) + k\log(n)$$

where k = Number of Parameters and n = Number of Data Points. The model with the smallest BIC is the preferred model.

I found the beta-binomial model quite difficult to fit as there is an implied constraint that both parameters in the Beta function need to be positive. The profile likelihood depends on θ. The best value of θ, obtained by trial and error, was $\theta = 2.95$.

```
> mylog3 <- function(alphabeta,y=numbfirestarts,theta=2.95){
+   lambdabeta <- exp(alphabeta[2]+alphabeta[3]*(temperature-mean(temperature))+
+     alphabeta[4]*log(lagrainfall/mean(lagrainfall)))
+     pialpha <- exp(alphabeta[1])/(1+exp(alphabeta[1]))
+     lchoose(lambdabeta,y)+
+       lbeta(y+rep(pialpha*theta,60),lambdabeta-y+rep((1-pialpha)*theta,60))-
+       lbeta(pialpha*theta,(1-pialpha)*theta)
+   }
> test3 <- maxLik(mylog3,start=c(0.3,7,0.2,-0.3),method="BFGS")
> summary(test3)
--------------------------------------------
Maximum Likelihood estimation
BFGS maximisation, 94 iterations
```

```
Return code 0: successful convergence
Log-Likelihood: -167.9604
4  free parameters
Estimates:
      Estimate Std. error t value   Pr(> t)
[1,] -0.061397   0.175587 -0.3497  0.726590
[2,]  2.661585   0.061459 43.3070 < 2.2e-16 ***
[3,]  0.144232   0.013345 10.8075 < 2.2e-16 ***
[4,] -0.417724   0.137845 -3.0304  0.002442 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
```

The BIC for this model is 352.30, which is higher than that for the generalised Poisson model, and hence the generalised Poisson model is preferred.

An alternative model is based on the negative-binomial distribution as usually defined, as distinct from the generalised Poisson distribution. The fitting of this model was straightforward compared to the beta binomial model. The code and results are given below:

```
> mylog2 <- function(alphabeta,y=numbfirestarts){
+   omegabeta <- exp(alphabeta[2]+alphabeta[3]*(temperature-mean(temperature))+
+     alphabeta[4]*log(lagrainfall/mean(lagrainfall)))
+     pialpha <- exp(alphabeta[1])/(1+exp(alphabeta[1]))
+     lchoose(omegabeta+y-1,y)+y*log(pialpha)+omegabeta*log(1-pialpha)
+   }
> test2 <- maxLik(mylog2,start=c(0.3,1.6,0.12,-0.33))
> summary(test2)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -167.5509
4  free parameters
Estimates:
      Estimate Std. error t value   Pr(> t)
[1,]  0.318390   0.325212  0.9790    0.3276
```

```
[2,]  1.610473   0.311671  5.1672 2.376e-07 ***
[3,]  0.112942   0.015793  7.1516 8.575e-13 ***
[4,] -0.337413   0.053345 -6.3251 2.531e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
-----------------------------------------
```

The BIC of the generalised Poisson model given in the report[4] was 350.742 while the BIC of the negative binomial model was 351.479. Hence the generalised Poisson model is the preferred model.

Regression parameters and regression model specification

The reported number of fire starts has a marked seasonal pattern so it was necessary to either incorporate seasonal dummy variables or explanatory variables that have such a pattern. The mean maximum temperature was an obvious choice. In addition, monthly rainfall was also used. Some experimentation showed that the maximal correlation occurred when the monthly rainfall was lagged one month i.e. when it was used as a leading indicator.

The model fits the data reasonably well, but does not fully capture the number of fire starts over the 2006/2007 summer period. All the parameters are statistically significant and the signs are sensible-that is the number of estimated fire starts increases with increases in temperature and decreases with increases in total monthly rainfall in the previous month.

Estimation method

The model was fitted by maximum likelihood which is the standard method for estimating parameters of complex statistical models. The likelihood of a set of parameter values given the observed data is equal to the probability of those observed outcomes given the parameter values. More likely parameter values will have a higher likelihood than less likely parameter values. The method of maximum likelihood chooses the set of parameter values which

[4]The previous report was titled "Under-reporting of fire starts" and was submitted to the AER on 20th November 2011. Table 1 of that document shows the results from a fitted negative binomial model. Those results were actually obtained from a generalised Poisson model. However, as explained earlier in this explanatory note, in the section on "allowing λ to have a distribution", the negative binomial distribution is approximately equal to a generalised Poisson distribution under certain conditions.

has the highest likelihood. The method is iterative and requires starting values, although I've used various starting values and got the same maximum value for the reported model. The negative of the inverse of the matrix of second derivatives gives an estimated variance covariance matrix from which standard errors of the estimated parameters can be determined. Both the estimated probability of reporting a fire start and the estimated actual number of fire starts is a non-linear function of the estimated parameters. Standard errors could be computed using the Delta method but I prefer to use the essentially equivalent simulation method which also captures the non-symmetric nature of the estimation distribution-the estimated reporting probability is skewed to the left, while the estimated actual number of fire starts is skewed to the right. It should be noted that the confidence intervals are predicated on the selected model being the correct one.

Other Considerations

In a report I am currently finalising I use Capture-Mark-Recapture methods to estimate the number of United Energy fire starts over the period 2006-2010. I obtain a higher estimate of the number of fire starts with the Capture-Mark-Recapture method than using the modelling approach. It is arguable that the Capture-Mark-Recapture method is a stronger method than the modelling approach, since at least two lists are involved rather than one.

Appendix: Probability Distributions

Distribution	Probability Function	Range	Mean	Variance
Bernoulii	$\pi(1-\pi)$	$0, 1$	π	$\pi(1-\pi)$
binomial	$\binom{n}{y}\pi^y(1-\pi)^{n-y}$	$0, 1, \ldots, n$	$n\pi$	$np(1-p)$
beta	$\frac{\pi^{\gamma-1}(1-\pi)^{\delta-1}}{\mathcal{B}(\gamma,\delta)}$	$0 < \pi < 1$ $\gamma, \delta > 1$	$\frac{\gamma}{\gamma+\delta}$	$\frac{\gamma\delta}{(\gamma+\delta)^2(\gamma+\delta+1)}$
beta-binomial	$\binom{\lambda}{y}\frac{\mathcal{B}(y+\gamma,\lambda-y+\delta)}{\mathcal{B}(\gamma,\delta)}$	$0, 1, \ldots, \lambda$	$\mu = \frac{\lambda\gamma}{\gamma+\delta}$	$\frac{\mu\delta}{\alpha+\beta}\frac{\lambda+\gamma+\delta}{1+\gamma+\delta}$
Poisson	$\frac{e^{-\lambda}\lambda^y}{y!}$	$y = 0, 1, \ldots$	λ	λ
Gamma	$\frac{y^{\alpha-1}\beta^\alpha e^{-\beta y}}{\Gamma(\alpha)}$	$y > 0$ $\alpha, \beta > 0$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Negative Binomial	$\binom{y+r-1}{r-1}\pi^r(1-\pi)^y$	$y = 0, 1, \ldots$	$\frac{r(1-\pi)}{\pi}$	$\frac{r(1-\pi)}{\pi^2}$
Generalised Poisson	$\begin{cases} (1/y!))\theta(\theta+y\tau)^{y-1}e^{-\theta-y\tau} \\ 0 \end{cases}$	y=0,1,... if $y > m$, when $\tau < 0$	$\theta(1-\tau)^{-1}$	$\theta(1-\tau)^{-3}$

Table 1: Details of Probability Functions used in this Document

Bibliography

Fader, P.S., and Hardie, B. (2000), "A note on modelling underreported Poisson Counts," *Journal of Applied Statistics,* **27**, 953-964.

Hubert, P. C., Lauretto, M. S., Stern, J. M., Goggans, P. M., & Chan, C.-Y. (2009). "FBST for Generalized Poisson Distribution". Retrieved from `http://link.aip.org/link/APCPCS/v1193/i1/p210/s1&Agg=doi`

Neubauer, G., Djuras, G., and Fiedl, H. (2011), "Models for underreporting: A bernoulli sampling approach for reported counts," *Austrian Journal of Statistics,* **40**, 85-92.

Win kelman, R. (1996) "Markov Chain Monte Carlo Analysis of Underreported Count Data with an Application to Worker Absenteeism," *Empirical Economics,* **21**, 575-581.

United Energy Distribution Pty Limited
ABN 70 064 651 029

UNITED ENERGY

Pinewood Corporate Centre
43-45 Centreway Place
Mt Waverley VIC 3149

P O Box 449
Mt Waverley VIC 3149

Telephone (03) 8846 9900
Facsimile (03) 8846 9999

7th November 2011

Our Reference: UE.ED.07.02

By email: Neil.Diamond@buseco.monash.edu.au

Dr Neil Diamond
Room 674, Building 11E
Department of Econometrics and Business Statistics
Monash University
CLAYTON VICTORIA 3800
Australia

Dear Dr Diamond,

Expert report in relation to the historical data on fire starts

The Australian Energy Regulator is responsible for the administration and operation of the f-factor scheme, and has recently released a draft determination, which is to apply over the period from 2012 to 2015[1]. The scheme aims to provide incentives for Distribution Network Service Providers (DNSPs) to reduce the risk of fire starts, and to reduce the risk of loss or damage caused by fire starts[2]. The scheme was developed by the Victorian Government.

An f-factor target has been set, which has been based, in part, on the historical occurrence of fire starts in each distribution network (including the United Energy distribution network) over the period from 2006 to 2010. United Energy has examined its data and has become aware that there was systematic under-reporting of fire starts over the five years from 2006 to 2010. The distribution management system used by the business was aimed at gathering information on faults, with a lesser degree of effort directed towards the gathering of data on fire starts.

An examination of the records in the distribution management system shows that evidence of fires and fire starts was reported in an *ad hoc* fashion. Inconsistent terminology has been used, spelling is inaccurate, and the descriptions in the text field are sometimes incomplete. The questions posed by SKM in relation to specific records in the UE Distribution Management System (DMS) are indicative of some of the problems with the historic recording of information pertaining to fire starts[3].

---

[1] AER, Draft determinations and Explanatory statement for the draft determinations, F-factor scheme determinations 2012-15 for Victorian electricity distribution network service providers, Australian Energy Regulator, 5th October 2011.
[2] Energy and Resources Legislation Amendment Bill 2010, Explanatory Memorandum, page 10.
[3] See AER – Guide to Questions – F-Factor Data Verification, questions posed by Terry Krieg, Sinclair Knight Merz, 2nd September 2011.

We are aware that linesmen were not fully briefed on the methods for reporting fire starts, although this situation began to change in 2010. Considering the 2006 to 2010 period as a whole, field personnel appear to have recorded the evidence for fire starts somewhat sporadically. Linesmen were not obliged to note down fire-related symptoms.

Previously, United Energy had formed the view that the reporting of pole and cross-arm fires from 2006 to 2010 was reasonably rigorous and well-founded. However, from a detailed examination of the records, and from discussions with field staff, we are confident that there were a number of pole fires that occurred which have not been documented.

In future, we expect more rigorous reporting of fire starts, because additional effort has been expended on re-training linesmen, and a new and enhanced reporting template has been created. The new template provides for answers to be chosen from among a menu of responses. Hence, there will be less reliance on the direct comments provided by linesmen.

In this context, we would like you to undertake and report on the following task:

- Review and assess the methods which have been applied by the AER in its draft determination to allow, and compensate for past under-recording of fire starts.
- Analyse a number of approaches which might assist in correcting for the past under-reporting of data on fire starts.
- Apply the methods making use of the various databases provided by United Energy.
- Determine a result which can be used as an appropriate benchmark to be adopted by United Energy as its "target" under the f-factor scheme.

Guidelines in preparing your report

Attached are Expert Witness Guidelines issued by the Federal Court of Australia. Although this brief is not in the context of litigation, the Victorian electricity distribution businesses are seeking a rigorously prepared independent view for use in the context of regulatory decision making and you are requested to follow the Guidelines to the extent reasonably possible in the context.

In particular, please:

Identify your relevant area of expertise and provide a curriculum vitae setting out the details of that expertise:

1.1.1. only address matters that are within your expertise;

1.1.2. where you have used factual or data inputs please identify those inputs and the sources;

1.1.3. if you make assumptions, please identify them as such and confirm that they are in your opinion reasonable assumptions to make;

1.1.4. if you undertake empirical work, please identify and explain the methods used by you in a manner that is accessible to a person not expert in your field;

UNITED ENERGY

1.1.5. confirm that you have made all the inquiries that you believe are desirable and appropriate and that no matters of significance that you regard as relevant have, to your knowledge, been withheld from your report; and

1.1.6. please do not provide legal advocacy or argument and please do not use an argumentative tone.

Yours sincerely,

Jeremy T. Rothfield.

Jeremy Rothfield
Network Regulation and Compliance Manager