ESTIMATION OF CHANGE POINT IN BINOMIAL RANDOM VARIABLES

S. M. Mundia[1], A. W. Gichuhi and J. M. Kihoro
[1]Department of Actuarial and Statistics, DeKUT
[2]Department of Statistics and Actuarial, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: mainamundia@yahoo.com

Abstract
Statistically, change point is the location or the time point such that observations follow one distribution up to the point and then another afterwards. Change point problems are encountered in our daily life and in disciplines such as economics, finance, medicine, geology, literature among others. In this paper, the change point in binomial observations whose the mean is dependent on explanatory variables is estimated. The maximum likelihood method was used to estimate the change point while the conditional means were estimated using the artificial neural network
The consistency and asymptotic normality of neural network parameter estimates was also proved. We used simulated data to estimate the change point and also estimated the LD50 for the Bliss beetles data.

Keywords: maximum likelihood estimate, binomial distribution, change point, artificial neural- network

1.0 Introduction

A sequence of independent binomial variables is subject to a change in distribution after an unknown point. Formally, we can describe this situation as follows. $m_1,...,m_b$ are independent binomial random variables, such that, for a value k, $1 < k < b$, m_i are distributed as

$$B(n_i, p_i) \qquad 1 \le i \le k$$
$$B(n_i, p_i') \quad k+1 \le i \le b \tag{1}$$

where $p_i(x)$ and $p_i'(x)$ are the success probabilities that depend on the explanatory variables $x = (x_1,...,x_D) \in \Re^D$. Here the assumption made is that there is a single change point at the point *k* . A change point problem will thus be two-fold,

i. Hypothesis testing, to ascertain whether significant change occurred in the distribution.
ii. Estimation of the change-point, *k*, if it exists.

Most analytical approaches, developed for dealing with binomial change-point data, assume the Parameters $p_i(x)$ and $p_i'(x)$, like *k* , to be unknown. Particular attention has been devoted to the case of the m_i being (Bernoulli) zero-one variables i.e. with $n_i = 1$ for all *i* .

Changes in the regression case have been considered by Quandt (1958, 1960) and Hinkley (1970) from the maximum likelihood viewpoint.

In this paper we first define the model used and how artificial neural networks are used to estimate the binomial probabilities. In Section 3 we prove the consistency and the asymptotic normality of the network parameter estimates. In Section 4 we conduct simulation studies and the results are presented in Section 5. In section 6 we have an application to Bliss beetles data.

2.0 The Model

The observations m_i are independently distributed binomial random variables whose probability distribution may be denoted as

$$f(m_i, p_i(x)) = {}_{n_i}C_{m_i} \left[p_i(x)\right]^{m_i} \left[1 - p_i(x)\right]^{n_i - m_i} \tag{2}$$

As the functional form of $p_i(x)$ is not known the output $\varphi(x,\theta)$ of a single hidden-layer feedforward neural network with $H \ge 1$ hidden nodes and a single output node is used to approximate $p_i(x)$

This output may be expressed as

$$\varphi(x,\theta) = \psi(\zeta(x,\theta))$$

$$\zeta(x,\theta) = \alpha_0 + \sum_{h=o}^{H} \alpha_h \left\{ wh_0 + \sum_{d=1}^{D} w_{hd} x_d \right\} \quad (3)$$

where $\theta \in \Omega = (w_{hj}, \alpha_h \quad h = 0,1,\cdots,H \quad j = 0,1,\cdots,D)$ is the vector of network weights and Ψ is the activation function of the network. The unipolar function is used as the activation function due to the fact that its output is in the range [0, 1] which makes it appropriate in the estimation of probabilities. Ω is assumed compact to ensure that it is closed and bounded.

Replacing $p_i(x)$ in equation (2) with $\varphi(x,\theta)$ we have

$$f(m_i, p_i(x)) = {}_{n_i}C_{m_i} \left[\varphi(x,\theta)\right]^{m_i} \left[1 - \varphi(x,\theta)\right]^{n_i - m_i} \quad (4)$$

To determine whether significant change occur the hypothesis problem is stated as

H_0:$p_i(x) = p_0(x) 1 \le i \le k$

Against

H_a:$p_i(x) = p_0(x)$ for some $i \le k$, and for some

$i > k$,$p_i(x) = p_i'(x)$

where $1 < k < b$ is the unknown change-point location and $p_0(x) \neq p_i'(x)$.

If H_0 is rejected then the value of *k* will have to be estimated. We use the likelihood method to do this. Several authors have considered this method. Ruhkin and Gary (1995) established the minimum error probability of the change-point maximum likelihood estimates for fixed binomial probabilities. Hinkley and Hinkley (1970) used the same method to estimate the change point when both the probabilities of success before and after the change-point are known. They also considered the situation where these probabilities are unknown and they replaced them with their maximum likelihood estimates (m.l.e's). If H_a is true and the probabilities $p_0(x)$ and $p_i'(x)$ are known then the likelihood function is given by

$$L_k(m, x, p_0, p') = \prod_{i=1}^{k} \binom{n_i}{m_i} \left[p_i(x)\right]^{m_i} \left[1 - p_i(x)\right]^{n_i - m_i} \prod_{i=k+1}^{b} \binom{n_i}{m_i} \left[p_i'(x)\right]^{m_i} \left[1 - p_i'(x)\right]^{n_i - m_i}$$

(5)

while for unknown probabilities their m.l.e.'s are used.

The m.l.e. $\hat{k}$ of *k* is the value of *k* that maximizes the likelihood function. Thus

$$\hat{k} = \max_{2 \le k \le b-1} L_k(m, x, p_0, p') \quad (6)$$

For unknown probabilities $p_0(x)$ and $p_i'(x)$ their respective m.l.e's. are $\frac{M}{N}$ and $\frac{M_k'}{N_k'}$ where

$M = \sum_{i=1}^{b} m_i$, $M_k = \sum_{i=1}^{k} m_i$, $N = \sum_{i=1}^{b} n_i$, $N_k = \sum_{i=1}^{k} n_i$, $M_k' = \sum_{i=1}^{b} m_i - \sum_{i=1}^{k} m_i$ and $N_k' = \sum_{i=1}^{b} n_i - \sum_{i=1}^{k} n_i$

Thus the loglikelihood function is

$$l_k = \log L_k(\hat{p}_0, \hat{p}')$$

$$= \sum_{i=1}^{b} \log \binom{n_i}{m_i} + \sum_{i=1}^{k} \left[m_i \log \frac{M_k}{N_k} + (n_i - m_i) \log (1 - \frac{M_k}{N_k}) \right] + \sum_{i=1}^{k} \left[m_i \log \frac{M_k'}{N_k'} + (n_i - m_i) \log (1 - \frac{M_k'}{N_k'}) \right]$$

$$= \sum_{i=1}^{b} \log \binom{n_i}{m_i} + \left[M_k \log \frac{M_k}{N_k} + (N_k - M_k) \log (\frac{N_k - M_k}{N_k}) \right] \quad (7)$$

Thus the maximum likelihood estimate of the change point *k* is

$$\hat{k} = \max_{2 \le k \le b-1} l_k \quad (8)$$

3.0 Consistency and Asymptotic Normality of the Neural Network Parameter Estimates

The random variables (m_i, X_i) are independent with parameters $(n_i, p(X_i))$ $i = 1, \cdots, b$.

An output of a neural network, $\varphi(x, \theta)$ is used to estimate $p_i(x)$ by minimizing the negative of the loglikelihood function divided by *b*. That is the function

$$l(\theta) = -\frac{1}{b} \left\{ \sum_{i=1}^{b} \log \binom{n_i}{m_i} + \left[ni \log \varphi(x, \theta) + (n_i - m_i) \log \varphi(x, \theta) \right] \right\} \quad (9)$$

is minimized.

The expected value of this target function $l_0(\theta)$ is

$$l_0(\theta) = -E\left[\frac{1}{b}\left\{ \sum_{i=1}^{b} \log\binom{n_i}{m_i} + \left[n_i \log\varphi(x,\theta) + (n_i - m_i)\log\varphi(x,\theta)\right] \right\} \right]$$

$$= -E\left[\left\{ \binom{n_1}{m_1} + \left[n_1 \log\varphi(X_1,\theta) + (n_1 - m_1)\log\varphi(X_1,\theta)\right] \right\} \right] \qquad (10)$$

Assuming that $l_0(\theta)$ has a unique minimum if θ is in the compact set Ω , then this minimum is characterised by

$$\nabla l_0(\theta) = -n_1 E\left\{ \frac{p(X_1)}{\varphi(X_1,\theta)} - \frac{1-p(X_1)}{1-\varphi(X_1,\theta)} \right\} \nabla\varphi(X_1,\theta) = 0 \qquad (11)$$

Since the neural network output functions are continuous in *x* and in θ and continuously differentiable in θ it is possible to interchange expectation and differentiation.

If the model is correctly specified then $p(x) = \varphi(x,\theta')$ for some $\theta' \in \Omega$

and equation (11) is solved but in a general situation θ' is defined as

$$\theta' = \arg\min_{\theta\in\Omega} l_0(\theta) \qquad (12)$$

For an estimator $\hat{\theta}$ of θ' obtained by minimising equation (10), its consistency implies that $\hat{\theta} \to \theta'$ as $b \to \infty$.

In the context of classical regression our model may written as

$$m_i = n_i p(X_i) + \varepsilon_i \quad i = 1,\cdots,b \qquad (13)$$

where the residuals are

$$\varepsilon_i = m_i - n_i p(X_i) \quad i = 1,\cdots,b \qquad (14)$$

Since the observations (m_i, X_i) are independent and $P(m_i \mid X_i) = \frac{1}{n_i} E(m_i \mid X_i)$

we have that $E(\varepsilon_i = 0)$ and

$$Var(\varepsilon_i) = E(m_i - n_i p_i(x))^2 = E\{(E(m_i - n_i p_i(x))^2 \mid X_i)\}$$

$$= E\{E(m_i^2 - 2m_i n_i p_i(x) + (n_i p_i(x))^2) \mid X_i)\}$$

$$= \sigma_\varepsilon^{\,2} < \infty$$

Also we note that $Var(\varepsilon_i)$ is independent of θ and

$$Var(\varepsilon_i \mid X_i) = n_i p(X_i)(1 - p(X_i))$$

Theorem 3.1 Let $U_1, U_2, \cdots$ be independent random vectors in $\Re^D$, $\Omega \subseteq \Re^M$ be compact,

$\Upsilon : \Re^D \times \Omega \to \Re$ be measurable such that

1. $E \mid \Upsilon(U_1;\theta \mid < \infty \quad \forall \theta \in \Omega$
2. $\Upsilon(U_1;\theta)$ is Lipschitz continuous in θ that is for some $L(u) > 0$
3. $E(L(_1U)) < \infty$

Then $\sup_{\theta\in\Omega} \mid \frac{1}{b}\sum_{i=1}^{b} \Upsilon(Ui;\theta) - E(\Upsilon(U_1;\theta)) \mid \to 0$ in probability.

This is the *Uniform Law of Large Numbers* (ULLN) whose proof is found in Andrews (1992). We use this theorem to prove the consistency of $\hat{\theta}$

Franke and Neumann (2000) in their work discussed nonlinear least square estimates for neural network parameters. To follow their work we make the following assumptions. We note that the residuals in equation (14) are independent and bounded in absolute value by m_i if

i. The activation function ψ is bounded and twice continuously differentiable and $E(m_i \mid X_i)$ is also bounded. This assumption is usually satisfied if the activation function is either unipolar or bipolar .

ii. $l_0(\theta)$ has a unique global minimum at θ' in the interior of Ω and $\nabla^2 l_0(\theta)$,which is the Hessian matrix is positive definite. This is a standard assumption in regression analysis.

iii. Ω is chosen such that for some $\delta > 0,\ \delta \le \varphi(x;\theta) \le 1-\delta$ for all $x \in \Re^D\ \theta \in \Omega$,.
This is a standard assumption.

iv. (m_i, X_i) are independent with some density $\upsilon(x)$ and $E \parallel X_1 \parallel^2 < \infty$.
This is a standard assumption since the observed values of X_1 will have to be finite.

v. $p(x)$ is continuous and for some $\upsilon > 0,\ 0 < \upsilon \le p(x) \le 1-\upsilon < 1$.

This assumption ensures that the experiments do not become degenerate. i.e. we do not have all events in the experiment with probability of one or zero.

To derive the asymptotic normality of $\hat{\theta} - \theta'$ we separately consider its two asymptotically in-dependent components $\hat{\theta} - \tilde{\theta}$ and $\tilde{\theta} - \theta'$ where

$\tilde{\theta} = \arg\min_{\theta\in\Omega} \tilde{l}(\theta)$ is generated by replacing m_i by

$E(m_i \mid X_i)$ in equation (9) to obtain

$$\tilde{l}(\theta) = -\frac{1}{b}\left\{ \sum_{i=1}^{b} \log \binom{n_i}{n_i p_i} + \left[ni \log \varphi(x,\theta) + (n_i - n_i p_i) \log \varphi(x,\theta) \right] \right\} \quad (16)$$

We use the following theorem which is similar to Theorem 1 of Franke and Neumann (2000) to show the asymptotic normality of $\hat{\theta} - \theta$

Theorem 3.2 Suppose assumptions (i)-(v) are satisfied. Let $(m_i, X_i) \sim B(n_i, p_i(x))$. Then as

$b \to \infty$, with $\hat{\theta}$ and θ' as defined above

$$\sqrt{b}\begin{pmatrix} \hat{\theta} - \tilde{\theta} \\ \tilde{\theta} - \theta' \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right)$$

that is $\sqrt{b}\,(\hat{\theta} - \tilde{\theta})$ and $\sqrt{b}\,(\tilde{\theta} - \theta')$ are asymptotically independent normal random vectors with co-variance matrices Σ_1 and Σ_2 respectively, where

$\Sigma_1 = A^{-1}(\theta')B_1(\theta')A^{-1}(\theta')$

$\Sigma_2 = A^{-1}(\theta')B_2(\theta')A^{-1}(\theta')$

With

$$B_1(\theta') = E\left[\frac{(n_1 p(X_1))(1 - p(X_1))}{\varphi^2(X_1;\theta')(1 - \varphi(X_1;\theta'))^2} \right] \nabla\varphi\ (X_1;\theta')\varphi^t(X_1;\theta')$$

$$B_2(\theta') = E\left[\frac{(n_1 p(X_1)) - \varphi(X_1;\theta'))^2}{\varphi^2(X_1;\theta')(1 - \varphi(X_1;\theta'))^2} \right] \nabla\varphi\ (X_1;\theta')\varphi^t(X_1;\theta')$$

An immediate consequence of this theorem is that $\sqrt{b}\,(\hat{\theta} - \tilde{\theta})$ and $\sqrt{b}\,(\tilde{\theta} - \theta')$ is asymptotically normal with mean 0 and covariance matrix $\Sigma_1 + \Sigma_2$. In a correctly specified model $B_2(\theta') = 0$ since there is essentially no effect due to the randomness of X_i's implying that the difference $(\hat{\theta} - \theta')$ is asymptotically of order smaller than $b^{-0.5}$ while in a miss-specified case this difference is of order $b^{-0.5}$. We also note that

$B_1(\theta')$ contains the variance of m_1 indicating its randomness while $B_2(\theta')$ contains the modeling bias and hence it would be zero if the model were correctly specified.

We prove the consistency and asymptotic normality of the network parameter estimates in four parts.

Part I

Using theorem 3.1, and taking $U_i = X_i$ then

$$\Upsilon(X_i;\theta) = -\left\{ \log\binom{n_i}{n_i p(X_i)} + \left[n_i p(X_i)\log\varphi(x,\theta) + (n_i - n_i p(X_i))(1-\log\varphi(x,\theta))\right] \right\} \quad (17)$$

Thus

$$\sup_{\theta\in\Omega} | \tilde{l}(\theta) - l_0(\theta) | = \sup_{\theta\in\Omega} | \frac{1}{b}\sum_{i=1}^{b} \Upsilon(Ui;\theta) - E(\Upsilon(U_1;\theta)) | = o_p(1) \quad (18)$$

Similarly,

$$| \tilde{l}(\theta) - l_0(\theta) | = | \frac{1}{b}\sum_{i=1}^{b} \log\frac{\binom{n_i}{n_i p(X_i)}}{\binom{n_i}{m_i}} - (n_i - | n_i p(X_i)) \log\frac{\varphi(x,\theta)}{1-\varphi(x,\theta)}$$

$l(\theta)$ and $\tilde{l}(\theta)$ as defined in equations (9) and (16).

Taking $U_i = (m_i, X_i)$ we obtain

$$\Upsilon(m_i, x, \theta) = \log\frac{\binom{n_i}{n_i p(X_i)}}{\binom{n_i}{m_i}} - (n_i - n_i p(X_i)) \log\frac{\varphi(x,\theta)}{1-\varphi(x,\theta)}$$

Then

$$\sup_{\theta\in\Omega} | \tilde{l}(\theta) - l_0(\theta) | = \sup_{\theta\in\Omega} | \frac{1}{b}\sum_{i=1}^{b} \Upsilon(m_i, x, \theta) | = o_p(1) \quad (19)$$

as $E(\Upsilon(m_1, X_1, \theta)) = 0$

We have to confirm whether the three conditions of Theorem 3.1 are satisfied in both cases. The activation function ψ is twice continuously differentiable and is bounded and so is $\varphi(x,\theta)$.

As the derivative of $\varphi(x,\theta)$ is bounded then we have a constant ω so that for all $x \in \Re^D, \theta \in \Omega$

$$| \frac{\partial}{\partial\theta_j}\varphi(x,\theta_j) | \le \omega \quad \text{if } \alpha_0, \cdots, \alpha_h, \quad w_{10}, \cdots, w_{h0} \le 1$$

$$|\frac{\partial}{\partial\theta_j}\varphi(x,\theta_j)|\leq\omega|X_i| \quad \text{if} \quad w_{1i},\cdots,w_{hi}\leq 1 \tag{20}$$

Hence it follows that for a suitable constant ω'

$$\|\varphi(x,\theta)\|\leq\omega'\|X_i\|$$

In a corresponding manner, for some constant ω''

$$\|\nabla\log\varphi(x,\theta)\|=\omega''\frac{\|\varphi(x,\theta)\|}{\varphi(x,\theta)}\|X_i\| \text{ and}$$

$$\|\nabla\log(1-\varphi(x,\theta))\|=\omega''\frac{\|\varphi(x,\theta)\|}{1-\varphi(x,\theta)}\|X_i\|$$

Hence for $\mathrm{Y}(X_i;\theta)$ in equation (17)

$$|\mathrm{Y}(u;\theta)-\mathrm{Y}(u;\theta')|\leq\sup_{\theta\in\Omega}\|\mathrm{Y}(u;\theta)\|\|\theta-\theta'\|$$

$$\leq\{\ n_i p(x)+(n_i-n_i p(x))\omega''\|x_i\|\|\theta-\theta'\|=n_i\omega''\|x_i\|\|\theta-\theta'\|$$

(21)

The assumption that are independent with finite variance makes conditions (ii) and (iii) of Theorem 3.1 to be satisfied with $L(u)=\omega''\|u\|$.

Also from the third assumption made after the statement of Theorem 3.1 and that $0\leq p(x)\leq 1$ we have that $\mathrm{Y}(u,\theta)$ is uniformly bounded in $x\in\Re^D,\theta\in\Omega$.

Since $m_i's$ are bounded binomial random variables then a similar argument to the above is used for

$\mathrm{Y}(x,\theta)$ in equation (19) and therefore we have from equations (18) and (19)

$\|\hat{\theta}-\tilde{\theta}\|=o_p(1)$ and $\|\tilde{\theta}-\theta'\|=o_p(1)$

Hence it follows by assumption (ii) and with increasing probability that $\tilde{\theta}$ and $\hat{\theta}$ are interior points in Ω. In particular

$$\nabla l(\hat{\theta})=\nabla l(\tilde{\theta})=\nabla l_0(\theta')$$

with probability close to 1 as $b\to\infty$

Part II

With probability close to 1 we have that

$$0=\nabla l(\tilde{\theta})-\nabla\tilde{l}\,(\theta')+\nabla\tilde{l}\,(\theta')$$

$$=(\tilde{\theta}-\theta')\nabla^2 l_0(\theta)\frac{1}{b}\sum_{i=1}^{b}\left\{\frac{n_i p(X_i)}{\varphi(X_i,\theta')}-\frac{(n_i-m_i)(1-p(X_i))}{1-\varphi(X_i,\theta')}\right\}\nabla\varphi(X_i,\theta')+F_1 \tag{22}$$

where ,

$$F_1 = \nabla l(\tilde{\theta}) - \nabla \tilde{l}\,(\theta') - (\tilde{\theta} - \theta') - \nabla^2 l_0(\theta') + (\tilde{\theta} - \theta')(\nabla^2 l_0(\theta')) - \nabla^2 l_0(\theta')$$
$$= o_p(\| \tilde{\theta} - \theta' \| \tag{23}$$

But

$$0 = \nabla l_0(\theta')$$
$$= n_1 E\left\{ \frac{p(X_1)}{\varphi(X_1,\theta')} - \frac{1-p(X_1)}{1-\varphi(X_1,\theta')} \right\} \nabla \varphi(X_1,\theta') \tag{24}$$

and by the central limit theorem the middle term of equation (22) is of the order $b^{-0.5}$. Since it is possible to interchange expectations and differentiation and $\varphi(x,\theta)$ is bounded and bounded away from zero uniformly in $x \in \Re^D, \theta \in \Omega$ then the logarithms in the functions $l_0(\theta), l(\theta)$ and $\tilde{l}\,(\theta)$ will all be defined.

Hence equation (22) becomes

$$\nabla^2 l_0(\theta')(\tilde{\theta} - \theta') + o_p(\| \tilde{\theta} - \theta' \|) = O(b^{-0.5}) \tag{25}$$

and since $\nabla^2 l_0(\theta')$ the Hessian is positive definite by assumption (ii) we have that

$$o_p(\| \tilde{\theta} - \theta' \|) = O(b^{-0.5}) \tag{26}$$

Replacing $\nabla^2 l_0(\theta')$ with $A(\theta')$ then equation (22) becomes

$$\sqrt{b}(\tilde{\theta} - \theta') = A(\theta')^{-1} \frac{1}{\sqrt{b}} \sum_{i=1}^{b} \left\{ \frac{n_i p(X_i)}{\varphi(X_i,\theta')} - \frac{(n_i - m_i)(1-p(X_i))}{1-\varphi(X_i,\theta')} \right\} \nabla \varphi(X_i,\theta') + o_p(1) \tag{27}$$

and hence for a suitable function s_1 satisfying $E(s_1(X_i)) = 0$ we get

$$\sqrt{b}(\tilde{\theta} - \theta') = b^{-0.5} \sum_{i=1}^{b} s_1(X_i) + o_p(1) \tag{28}$$

Part III

From equations (13), (19) and that $E(\varepsilon_i \mid X_i) = 0$, then with probability going to 1 we have

$$\mathrm{Y}(m_i, x, \theta) = \log \frac{\begin{pmatrix} n_i \\ n_i p(X_i) \end{pmatrix}}{\begin{pmatrix} n_i \\ m_i \end{pmatrix}} - (\varepsilon_i) \log \frac{\varphi(x,\theta)}{1-\varphi(x,\theta)} \tag{29}$$

and

$$0 = \nabla l(\hat{\theta}) - \nabla (l(\hat{\theta}) + \nabla \tilde{l}\,(\hat{\theta}))$$

$$= \nabla \tilde{l}(\theta) + \frac{1}{\sqrt{b}} \sum_{i=1}^{b} \Upsilon(m, x, \hat{\theta})$$

$$= \nabla \tilde{l}(\theta) - \frac{1}{\sqrt{b}} \sum_{i=1}^{b} \varepsilon_i \frac{\nabla \varphi(X_i, \hat{\theta})}{\varphi(X_i, \hat{\theta})(1 - \varphi(X_i, \hat{\theta}))} \tag{30}$$

As in part II of the proof we have that

$$\sqrt{b}(\hat{\theta} - \tilde{\theta}) = A(\theta')^{-1} \frac{1}{\sqrt{b}} \sum_{i=1}^{b} \left\{ \frac{n_i p(X_i)}{\varphi(X_i, \theta')} - \frac{(n_i - m_i)(1 - p(X_i))}{1 - \varphi(X_i, \theta')} \right\} \nabla \varphi(X_i, \theta') + o_p(1) \tag{31}$$

and hence for a suitable function s_2 satisfying $E(s_2(X_i)) = 0$ we get

$$\sqrt{b}(\hat{\theta} - \theta') = b^{-0.5} \sum_{i=1}^{b} s_2(X_i) + o_p(1) \tag{32}$$

Hence we have for some constants ω_1, ω_2 and for all $x \in \Re^D$

$$\| s_1(x) \| \leq \omega_1 \| x \| \text{ and } \| s_2(x) \| \leq \omega_2 \| x \|$$

since $\nabla \| \varphi(x; \theta) \|$ is bounded. As $E(X_i)$ is finite, ε_i bounded and $(X_i \varepsilon_i)$ are independent we have

$$\sqrt{b} \begin{pmatrix} \hat{\theta} - \tilde{\theta} \\ \tilde{\theta} - \theta' \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right) \tag{33}$$

as for all f, g

$$b\ Cov(\tilde{\theta}_f - \theta'_f, \hat{\theta}_f - \tilde{\theta}_f) = b^{-1} \sum_{i,j=1}^{b} E(s_{1f}(X_i) s_{2g}(X_j) \varepsilon_j) + o_p(1)$$

$$= b^{-1} \sum_{i,\neq j}^{b} E(s_{1f}(X_i) s_{2g}(X_j)\ \varepsilon_j)$$

$$+ b^{-1} \sum_{i,=1}^{b} E(s_{1f}(X_i) s_{2g}(X_i) \varepsilon_i) + o_p(1)$$

$$= o_p(1) \tag{34}$$

as $E(X_i \mid \varepsilon_i) = 0$

Part IV

We now require the form of Σ_1 and Σ_2 .

$$\Sigma_1 = E(s_1(X_1)s_1^{\,t}(X_1) = A^{-1}(\theta')B_1(\theta')A^{-1}(\theta') \quad (35)$$

Since $s_1(X_i)\ \ i=1,\cdots,b$ are independent and where $E(s_1(X_i))=0$, where

$$B_1(\theta') = E\left[\frac{(n_1 p(X_1))(1-p(X_1))}{\varphi^2(X_1;\theta')(1-\varphi(X_1;\theta'))^2} \right]\nabla\varphi\ (X_1;\theta')\varphi^t(X_1;\theta')$$

Similarly as $E(\varepsilon_i \mid X_1) = \sigma_{\varepsilon i}$ as in equation (15) we have

$$\Sigma_2 = E(s_2(X_1)s_2^{\,t}(X_1)\varepsilon_i^2)$$
$$= E(s_2(X_1)s_2^{\,t}(X_1)n_i p(X_1)(1-n_i p(X_1)))$$
$$= A^{-1}(\theta')B_2(\theta')A^{-1}(\theta') \quad (36)$$

where

$$B_2(\theta') = E\left[\frac{(n_1 p(X_1)) - \varphi(X_1;\theta'))^2}{\varphi^2(X_1;\theta')(1-\varphi(X_1;\theta'))^2} \right]\nabla\varphi\ (X_1;\theta')\varphi^t(X_1;\theta')$$

Thus the theorem is proved.

4.0 Simulation Studies

For simulation purposes under H_a, the following model was used

$$P(m_i \mid X_i = x) = \begin{cases} (1+\exp(-(-1.5+x_{1i}+x_{2i})))^{-1} & 1<i\le k \\ (1+\exp(-(-1.5+2x_{1i}+1.8x_{2i})))^{-1} & k+1<i\le b-1 \end{cases}$$

(37)

The change point *k* was at fixed for a sample size *b*= 200. x_{1i} and x_{2i} were generated as uniform [0, 1]. n_i, the size of the i^{th} group were generated as the integer part of uniform [2, b]. Then the binomial random variable m_i is generated in line with equation (37). A simulation was done when change point was fixed half way through the data with the aim of testing whether change existed.

At each estimated change point the values of the loglikelihood and the test statistic calculated. A plot of the test statistic is presented in Figure 1. A plot of the values of loglikelihood against the estimated change point is given in Figure 3. A further simulation is carried out to test for a change when it is actually not present and at each estimated change point the values of loglikelihood and the test statistic

calculated. A plot of the test statistic is presented in Figure 2. 1000 simulationswere carried out with the change point fixed half way through the data. That is the change point likelihood estimates of the change point is presented in Figure was fixed at 100.

In each simulation the change point is estimated. A histogram of the maximum likelihood estimates of the change point is presented in Figure 4.

Another 1000 simulation were carried out when actually there were no change in the data. A histogram of the maximum likelihood estimates of the change point is presented in Figure 5.

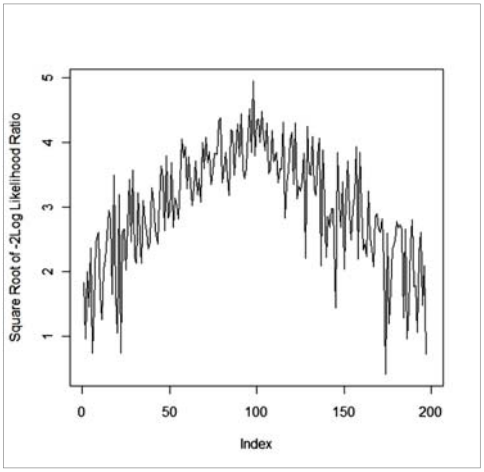

Figure 1: Plot of the values of the test statistic when the alternative is true

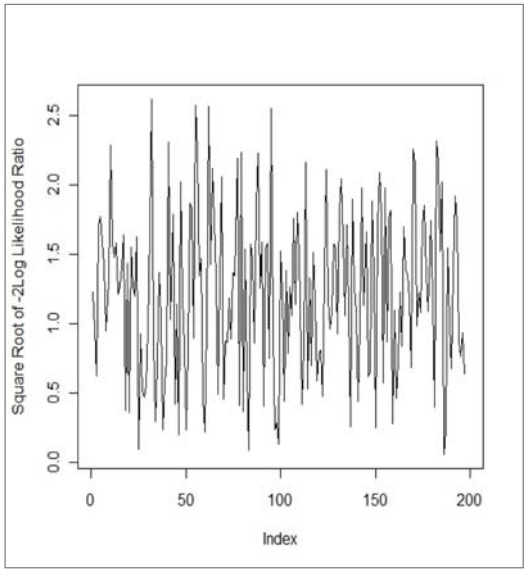

Figure 2: Plot of the values of the test statistic when the alternative is false

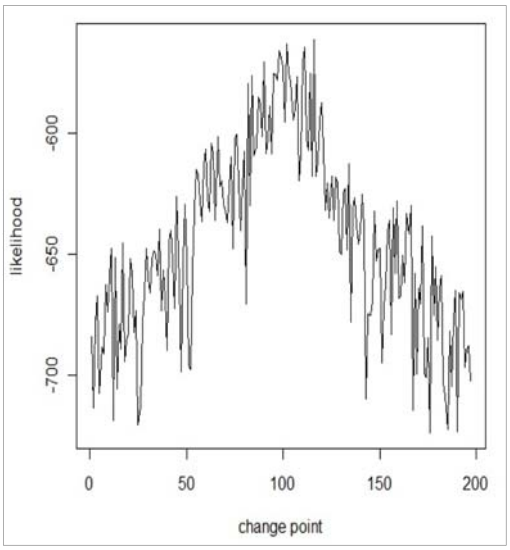

Figure 3: Plot of the values of the loglikelihood maximum likelihood estimates of change point when the null hypothesis is false.

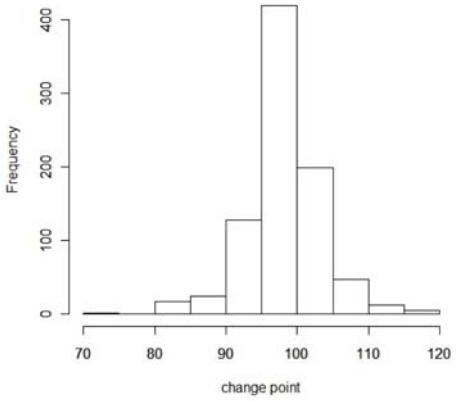

Figure 4: Histogram of maximum likelihood estimates of change point when the null hypothesis is false

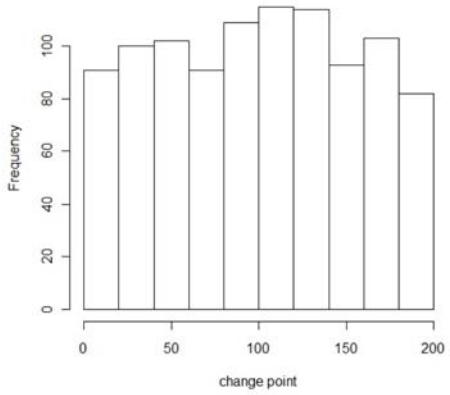

Figure 5: Histogram of maximum likelihood estimates of change point when there is no change

5.0 Results

Figure 1 indicates that the hypothesis of change is accepted at all levels as the value of the test statistic is greater than the critical values C_2 given in Table 1. But the hypothesis of change is accepted at 1% when using the critical values C_1. This is due to the fact that the rejection regions given by C_1 are conservative as noted by Gombay and Horvath (1996)

From Figure 2 the hypothesis of no change is not rejected at all the three levels of significance. In both cases the critical values used are in Table 1 with sample size of 200. The critical values are generated in line with Gombay and Horvath (1996).

Table 1: Critical values

Sample size	α	C_1	C_2
200	0.01	5.268792	4.908558
	0.05	4.467199	4.449472
	0.1	3.982043	4.22199
481	0.01	5.310178	4.73092
	0.05	4.456003	4.274104
	0.1	4.078778	4.049254

Figure 3 indicates that we have large values of the loglikelihood near the change point. A histogram of the maximum likelihood estimates of the change point shows that most of them are between 95-100 which is near our actual value of the change point.

We look at the asymptotic properties of the change point estimates. Figure 7 shows the histogram of the biases of the change point estimates. The biases have an approximate mean of 0. To evaluate the goodness of fit we draw normal curve whose mean and variance are those of the bias. This is presented in Figure 6. Further a quantile-quantile plot, in Figure 8 confirms the normality of the change point estimates. We further performed Kolmogorov-Smirnov test of normality on the biases of the change point estimates. The test gives a p-value of 0.01121. Thus the null hypothesis of normality is accepted at 1%. This is an indication that the biases of the estimates are asymptotically normally distributed with a mean of zero.

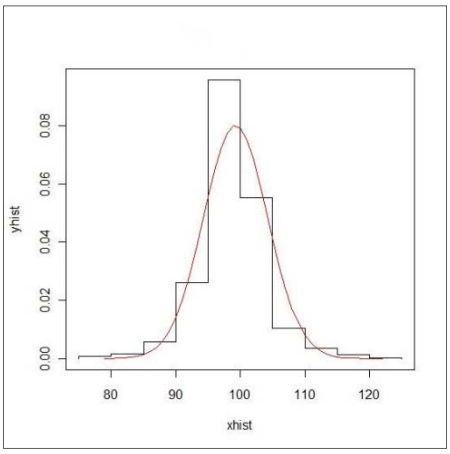

Figure 6: Normal curve and histogram together

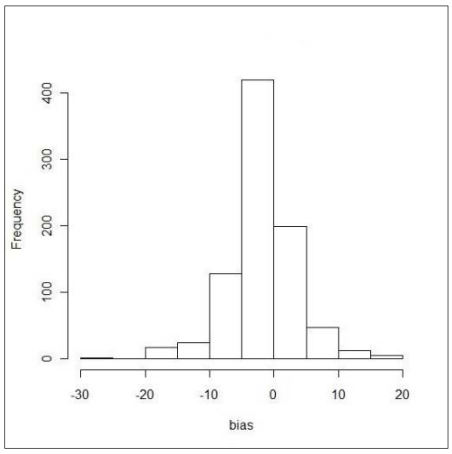

Figure 7: Histogram of the biases of the change point estimates

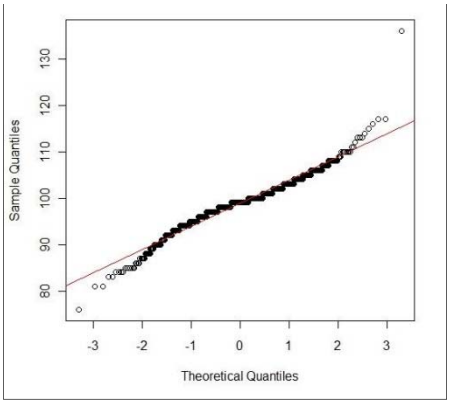

Figure 8: qqplot of the change-point estimates

6.0 Application to Real Data

To demonstrate the use of artificial neural networks in the estimation of the conditional means and maximum likelihood estimate of change point, we use the famous beetles data of Bliss (1935) batches of adult beetles were exposed to

gaseous carbon disulphide for five hours. This data has been extensively used by statisticians in studies of generalized link functions e.g., Prentice (1976) Stukel (1988) and are used by Spiegelhalter *et al* (1996) to demonstrate how BUGS handles generalized linear models for binomial data. The data is given below.

Table 2: Beetles Data

Dosage (CS2mg/litre)	**No. of beetles**	**No. of beetles killed**
49.057	59	6
52.991	60	13
56.911	62	18
60.842	56	28
64.759	63	52
68.691	59	53
72.611	62	61
76.542	60	60

Here we assume that $P(m_i \mid X_1) = \beta_0 + \beta_1 X_{1i}$ where m_i is the number of deaths due to the i^{th} dose and X_{1i} is the respective dose. We want to determine the dosage at which 50% of the beetles are killed as this indicates a significant change in the structure of the probability of death. In line with Gombay and Horvath [4] we generated the critical values C_1 and C_2, using a sample size as 481 which we presented in Table 1 . The graph in Figure 9 gives us a value test statistic of 15.44, the is the maximum and this leads to the rejection hypothesis of no change. In Figure 10 we present a plot of the values of the loglikelihood against the dosage. The maximum of the curve corresponds to the third change point location which is the fourth dosage. From the data the fourth dosage of 60.842 CS_2mg/litre kills 50% of the beetles.

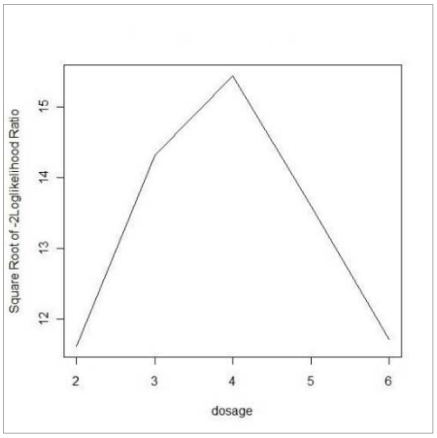

Figure 9: Plot of the values of the test statistic for the Bliss data

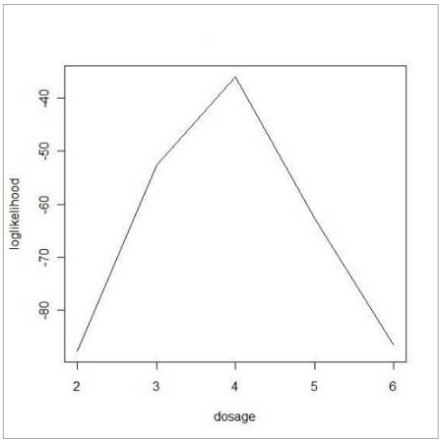

Figure 10: Plot of the values of the loglikelihood against the dosage for the Bliss data

This paper proposes the use of artificial neural network in the estimation of the conditional binomial probabilities and then the use the maximum likelihood to estimate the change point. If the change point is identified, then the conditional probabilities are estimated from the explanatory variables.
Simulation under the alternative hypothesis shows that the change point estimator is consistent, asymptotically unbiased and normally distributed. An empirical example on analyzing the Bliss beetles data is given.
The simulation and data analysis programs in R are available from the first author.

Acknowledgments
The authors wish to thank the Editor-in-chief and referees for their valuable comments and suggestions that helped to improve the presentation of the paper.

References

Andrews, D. Generic uniform convergence. *Econometric Theory*, 8, 2(1992), pp. 241-257.

Bliss, C. I. The calculation of the dosage-mortality curve. *Annals of Applied Biology* **22 (1935)**, pp. 134-167.

Franke, J. and Neumann, M. Bootstrapping neural networks .*Neural Computation* **12 (2000)**, pp. 1929-1949.

Gombay, E. and Horvath, L. On the rate of approximation for maximum likelihood tests in change-points models. *Journal of Multivariate Analysis*, **56 (1996)**, pp. 120-152.

Hinkley , D., and Hinkley, E.Inference about change point in a sequence of binomial variable. *Biometrika*, 57(1970), 477-488.

Prentice, and Ross, L. A generalization of the probit and logit methods for dose response curves. *Biometrics* 32 (1976), 134-167.

Quandt , R. The estimation of parameters of a linear regression system that obeys two separate regimes. *Journal of American Statistical Association*, 53 (1958), 73-88.

Quandt, R. Tests of hypothesis that a linear regression system obeys two separate regimes.

Journal of American Statistical Association, 55 (1960), 324-330.

Ruhkin, A., and Gary, M. Asymptotic behavior of estimators of change-point in binomial probability. *Applied Statistical science* 2, 1 (1995), 1-12.

Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. R. BUGS: Bayesian inference Using Gibbs sampling, version 0.5 (version ii). Cambridge, UK: *Biostatistics Unit*.

(1996).

Stukel, T. A. Generalized logistic models. *Journal of the American Statistical Association,* 83 (1988),426-431.