Regression Trees for Analysis of Count Data with Extra Poisson Variation

Yunhee Choi[1], **Hongshik Ahn**[2*] **and James J. Chen**[3]

[1]Department of Preventive Medicine, School of Medicine
Seoul National University, 28 Yongeon-Dong, Chongro-Gu, Korea

[2]Department of Applied Mathematics and Statistics
Stony Brook University, Stony Brook, NY 11794-3600

[3]Division of Biometry and Risk Assessment
National Center for Toxicological Research, FDA, Jefferson, AR 72079

Abstract

This article proposes methods for fitting piecewise loglinear models to count data with an extra-Poisson variation. Both SUPPORT (Chaudhuri et al., 1994, *Statistica Sinica,* **4,** 143-167) and GUIDE (Loh, 2002, *Statistica Sinica*, **12,** 361-386) are used for splitting methods. We developed a new bootstrap resampling method performed at each node of the tree to determine the proper size of a tree. The quasi-likelihood approach is used for fitting an extra-Poisson model at each stratum to take into account the extra variability. An adjusted Anscombe residual for the extra-Poisson model is used in this procedure. Performance of the proposed method is evaluated by a Monte Carlo simulation study. The proposed method is used to investigate geographic variability in mortality rates on lung cancer as well as effects of various demographic variability.

Keywords: Carcinogenicity; Generalized linear model; Quasi-likelihood; Recursive partitioning

1 Introduction

Poisson regression models are widely used in the analysis and prediction of counts on potential independent variables. Possible applications include modeling the numbers of colonies of bacteria on various dilutions and experimental conditions, and the counts of failures of machines under different operational conditions. However, count data often exhibit greater variability than would be

*Corresponding author. e-mail: hahn@ams.stonybrook.edu

predicted by simply fitting a Poisson model. For example, studies on number of prostate cancer deaths (Holford, 1983) and number of breast cancer incidence cases (Stevens et al., 1982) considered models with extra-Poisson variates. Faddy and Bosch (2001) studied under-dispersed Poisson variations on the number of fetal implants in mice. For such data, it is desirable to use a model that allows the possibility of an extra-Poisson variation. This is an important issue because ignoring the over-dispersion may lead anti-conservative test results which can mislead the interpretation of the data. To fit an extra-Poisson model, Engel (1984) and Lawless (1987) applied the negative-binomial approach, Breslow (1984) used weighted least squares, and Chen and Ahn (1996) applied the quasi-likelihood approach. In this paper, we modified the tree-structured methods for over-dispersed binomial data by Ahn and Chen. The quasi-likelihood approach is used for fitting extra-Poisson models, and a moment estimation and testing of the over-dispersion (or under-dispersion) parameter is developed.

One regression model for the whole sample is often difficult to interpret, especially when there are numerous covariates, some of which being correlated. One way to overcome these disadvantages is to stratify the data according to selected covariate values and fit separate regression models to each stratum. This can provide insight into the nature of the response and explanatory variables within a stratum. Further, because the data in a stratum would be more homogeneous, they may be fit with models having fewer covariates. In the proposed method, strata are formed by a recursive stratification through a extra-Poisson regression tree. Poisson regression tree without considering extra variation has been studied by Chaudhuri et al. (1995). To our knowledge, however, a regression tree model for count data with extra-Poisson variation has not been attempted. Although existing software such as CART (Classification and Regression Trees) assesses the fit of its models with cross-validation, there is a limitation because the model does not include the over-dispersion.

Marienfeld et al. (1980) conducted an epidemiological investigation of the effect of public drinking water on cancer mortality in Missouri. The data for lung-cancer mortality among the 115 counties during the period of 1972-1981 for both sexes can be obtained from our web site (http://www.ams.stonybrook.edu/~hahn/research/tree.html). Mortality is classified by sex and four age groups: 45-54, 55-64, 65-74, and over 75. Figure 1 shows the scatter plots of log(lung cancer mortality) versus age and population for each sex with smooth ("lowess") curves superimposed. In this figure, the number of deaths per 1000 people is used as the lung cancer mortality. In the plots for population, three counties with total population over 100,000 are excluded for a clear view of the low population

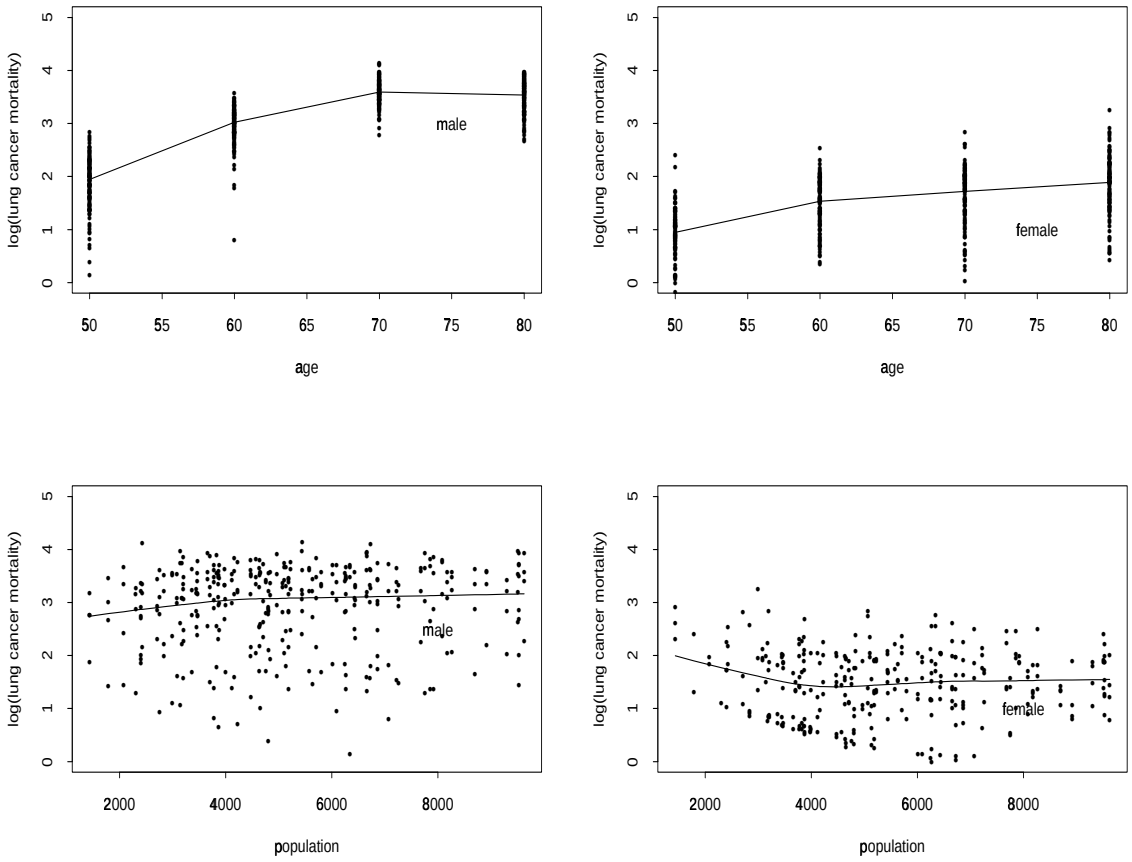

Figure 1: Scatterplots of log(lung cancer mortality) versus age and population for the cancer mortality data in the state of Missouri. In the plots for population, three counties with population over 100,000 were excluded for a clear view of the low population area. The number of deaths per 1000 people are used as the lung cancer mortality.

area. This figure clearly shows that the lung cancer mortality patterns are substantially different between males and females. The lung cancer mortality appears to be piecewise linear on population in a log scale for both sexes. However, it slightly increases for males as population increases, while it decreases for females until it reaches 4000 and then increases for population > 4000. The plots on age also suggest piecewise linear fits in a log scale with different increments in each sex. For each of sex, age and population, piecewise linear fits appear to be adequate. In fact, most cancer types affect different age-sex groups quite differently (Hill, 1977, p198; Tsutakawa, 1988). As will be discussed in Section 5, these count data exhibit a greater variability than a Poisson distribution. Therefore, a tree-structured extra-Poisson regression model might be appropriate for analyzing these data. Through regression trees, lung-cancer mortality will be investigated in relation to age, sex

and population size in this paper.

We employ two splitting methods based on the residual distribution for constructing trees. One is SUPPORT (Smoothed and Unsmoothed Piecewise Polynomial Regression Trees) by Chaudhuri et al. (1994) and the other is GUIDE (Generalized, Unbiased Interaction Detection and Estimation) by Loh (2002). Both methods detect different patterns in residual distributions, but SUPPORT does not deal with categorical predictors. A noticeable point in GUIDE is that it limits the predictor's role to either split, regressor or both. Recursive partitioning is performed using a combination of statistical tests and residual analysis in these methods. This approach has been shown to be effective for tree-structured classification (Loh and Vanichsetakul, 1988), piecewise-polynomials regression (Chaudhuri et al., 1994) and regression with censored data (Ahn and Loh, 1994). A further detail on these methods are provided in Section 3.

We employ a multi-step look-ahead stopping rule with cross-validation and bootstrap resampling to determine the size of a tree. In order to obtain a tree with an optimal size, we propose a new bootstrap approach which is conducted at each node of the tree. This method controls the tree size better than the method of Ahn and Loh (1994) which conducts the bootstrap estimation only at the root node. As an alternative stopping rule, the backward-elimination method (V-fold cross-validation cost-complexity pruning) by CART after growing a huge tree is also examined and compared with the look-ahead procedure. The performance of the proposed extra-Poisson regression tree method is evaluated via a Monte Carlo simulation study. Further, we examine how the split affects the Type I error (the error of discovering two or more strata when there is only one) rate and power, especially when a design matrix contains categorical variables, by comparing SUPPORT and GUIDE.

The deviance based on the Anscombe residuals for the extra-Poisson regression is used as the measure of goodness of split. Throughout this paper, the regression tree with an extra-Poisson model will be called the extra-Poisson tree, and the tree with a standard Poisson regression model will be called the Poisson regression tree.

2 Extra-Poisson Variate and Loglinear Model

We start this section with a review of loglinear model for data with extra-Poisson variate. Suppose $Y_1, Y_2, \cdots, Y_n$ are count data from n clusters or batches with $\mathrm{E}(Y_i) = \mu_i$, $i = 1, \ldots, n$. If there exits

a correlation within a cluster or a batch, then $\mathrm{Var}(Y_i) = \mu_i(1 + \phi_i\mu_i)$, where ϕ_i is the dispersion parameter. If $\phi_i > 0$, Y_i has an over-dispersed Poisson variate and if $\phi_i < 0$, it has an under-dispersed Poisson variate. When $\phi_i = 0$, Y_i reduces to a Poisson random variable with $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \mu_i$. We assume $\phi_1 = \cdots = \phi_n = \phi$ in this paper. Ignoring this over-dispersion may lead anti-conservative test results which can mislead the interpretation of the data.

A classic approach to an extra-Poisson variate is to assume that the mean of the Poisson has a gamma distribution which leads to a negative binomial distribution for the observed data. See Margolin et al. (1981), Lawless (1987) and Moore and Tsiatis (1991). Suppose the conditional distribution of Y given μ is Poisson with mean μ, then $\mathrm{E}(Y|\mu) = \mathrm{Var}(Y|\mu) = \mu$. Assume that the prior distribution of μ is Gamma(α, γ), then the marginal distribution of Y is Negative binomial$(\alpha, (1+\gamma)^{-1})$. Therefore, $\mathrm{E}(Y) = \alpha\gamma$ and $\mathrm{Var}(Y) = \alpha\gamma(1 + \gamma)$. Since the extra-Poisson variate has the same mean as the standard Poisson distribution, we obtain $\mu = \mathrm{E}(Y) = \alpha\gamma$. Thus, $\mathrm{Var}(Y) = \mu(1 + \phi\mu)$ if we define $\phi = 1/\alpha$. This model is used to generate the extra-Poisson data in our simulation study.

An alternative to the parametric models is the quasi-likelihood approach. This approach is used in the proposed study for fitting an extra-Poisson model. Define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Wedderburn (1974) defined the quasi-likelihood Q for an observation $\boldsymbol{y}$ with mean $\boldsymbol{\mu}$ and variance $V(\boldsymbol{\mu})$ as

$$\frac{\partial Q(\boldsymbol{Y}|\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \frac{\boldsymbol{Y} - \boldsymbol{\mu}}{V(\boldsymbol{\mu})}$$

for a given variance structure $\mathrm{Var}(\boldsymbol{Y}) = \mathrm{v}(\boldsymbol{\mu}) = \boldsymbol{\mu}(1 + \phi\boldsymbol{\mu})$. Let $\boldsymbol{\beta}$ be a vector of the regression coefficients. The estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which maximizes the quasi-likelihood is

$$0 = \frac{\partial Q}{\partial \boldsymbol{\beta}} = \frac{\partial Q}{\partial \boldsymbol{\mu}}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial [Q(Y_i|\mu_i)]}{\partial \mu_i}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V(\mu_i)}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}),$$

where $D_{ij} = \partial\mu_j/\partial\beta_i$ and $\boldsymbol{V} = \mathrm{diag}[v(\mu_1), v(\mu_2), \cdots, v(\mu_n)]$. Let $\boldsymbol{b}$ be the true value of $\boldsymbol{\beta}$ and $\boldsymbol{b}_0$ an initial guess of $\boldsymbol{\beta}$, then by the Newton-Raphson Method, we obtain

$$\boldsymbol{b} = \boldsymbol{b}_0 + (\boldsymbol{b} - \boldsymbol{b}_0) = \boldsymbol{b}_0 + I(\boldsymbol{b}_0)^{-1}\frac{\partial Q}{\partial \boldsymbol{\beta}}\Big|_{b_0} = \boldsymbol{b}_0 + [\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}]^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}).$$

In this paper, we propose a moment estimator of ϕ which does not require a numerical method. In each Newton-Raphson iteration, the estimate of ϕ can be obtained as

$$\hat{\phi} = \frac{\{\sum_{i=1}^{n}(Y_i - \mu_i)^2 - \mu_i\}}{\sum_{i=1}^{n}\mu_i^2}. \tag{1}$$

Using the fact that $\text{Var}(Y_i) = \mu_i(1+\mu_i\phi)$, the above moment estimator can be obtained by equating $n^{-1}\sum_{i=1}^{n}\mu_i(1+\phi\hat{\mu}_i)$ with an estimate $\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2/(n-k)$ of $\text{Var}(Y)$. This estimator is unbiased since

$$\text{E}(\hat{\phi}) = \frac{\sum_{i=1}^{n}\{\text{E}(Y_i-\mu_i)^2-\mu_i\}}{\sum_{i=1}^{n}\mu_i^2} = \frac{\sum_{i=1}^{n}\{\mu_i(1+\phi\mu_i)-\mu_i\}}{\sum_{i=1}^{n}\mu_i^2} = \phi.$$

The variance of $\hat{\phi}$ can be obtained as

$$\text{Var}(\hat{\phi}) = \frac{\sum_{i=1}^{n}\text{Var}\{(Y_i-\mu_i)^2\}}{(\sum_{i=1}^{n}\mu_i^2)^2} = \frac{\sum_{i=1}^{n}(\mu_i+2\mu_i^2)}{(\sum_{i=1}^{n}\mu_i^2)^2}. \tag{2}$$

To test H_0: $\phi = 0$ versus H_1: $\phi \neq 0$, the test statistic can be obtained using Equations (1) and (2) as

$$Z = \frac{\hat{\phi}}{\sqrt{\text{Var}(\hat{\phi})}}.$$

We may use a Poisson regression model if H_0 is not rejected because it is a simpler model and it will prevent an overfit by adding an unnecessary parameter. However, the extra-Poisson model can also be used for such data because the Poisson model is a special case of the extra-Poisson model, and including ϕ will prevent a possible Type I error.

The conventional residual is $Y_i - \mu_i$, but since the response Y_i is distributed as a Poisson with $\text{Var}(Y_i) = \mu_i(1+\phi\mu_i)$, some adjustment is required to define the residuals for diagnostic purpose. Generalized linear models require a generalization of residuals, applicable to all the distributions which can be used for the same purposes as the standard normal residuals (see McCullagh and Nelder, 1989). McCullagh and Nelder provide several kinds of residuals in Poisson regression. These residuals can be modified for extra-Poisson models.

Pearson residuals can be modified for the extra-Poisson variation as

$$R_p(Y_i,\mu_i) = \frac{Y_i - E(Y_i)}{\text{Var}(Y_i)} = \frac{Y_i-\mu_i}{\sqrt{\mu_i(1+\phi\mu_i)}}.$$

However, the distribution of these residuals can be highly skewed. Anscombe (1953) proposed a transformation $t(y)$ to obtain a normal-like residuals (see McCullagh and Nelder, 1989). This residual for a Poisson distribution is defined as

$$R_A(Y_i,\mu_i) = \frac{t(Y_i)-\text{E}\{t(Y_i)\}}{\text{Var}[t(Y_i)]} = \frac{Y_i^{2/3}-\mu^{2/3}}{\frac{2}{3}\mu^{1/6}}. \tag{3}$$

When the μ_i's are close to zero, the response is likely to be zero. Hence, the Anscombe residuals

will be negative in these cases. To avoid this situation, Pierce and Schafer (1986) proposed the adjusted Anscombe residuals by adding a correction term to E$[t(Y_i)]$. For an extra-Poisson variate, we derived the variance of $t(Y_i)$ using the delta-method, and obtained the residual as

$$R(Y_i, \hat{\mu}_i) = \frac{Y_i^{2/3} - (\hat{\mu}_i^{2/3} - \hat{\mu}^{-1/3}/9)}{\frac{2}{3}\hat{\mu}^{1/6}(1+\phi\hat{\mu}_i)^{1/2}}. \tag{4}$$

In this paper, this adjusted Anscombe residual is used.

The deviance as a goodness-of-fit measure for a Poisson regression model is defined as

$$d(\boldsymbol{y}|\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\{\ell(\boldsymbol{y},\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}},\boldsymbol{y})\} = 2\sum_{i=1}^{n}\{y_i\log(y_i) - y_i - \log(y_i!)\}. \tag{5}$$

The deviance residuals based on (5) are often close to the Anscombe residuals in numerical value (McCullagh and Nelder, 1989; Pierce and Schafer, 1986). For an extra-Poisson regression model, the deviance can be defined as

$$\begin{aligned} d(\boldsymbol{y}|\boldsymbol{\mu}) &= \sum_{i=1}^{n}[2Q(\boldsymbol{y},\boldsymbol{y}) - 2Q(\boldsymbol{\mu},\boldsymbol{y})] = 2\int_{\mu}^{y}\frac{y-t}{t(1+\phi t)}dt \\ &= 2\sum_{i=1}^{n}\left\{y_i\log\frac{y_i}{\mu_i} - \left(\frac{1}{\phi}+y_i\right)\log\frac{1+\phi y_i}{1+\phi\mu_i}\right\}. \end{aligned}$$

This deviance function is a standard measure of a discrepancy in quasi-likelihood models. It is used as the splitting criterion in cross-validation in the proposed extra-Poisson tree. The main idea of this is from Ahn and Chen (1997).

3 Tree-Structured Methods

Splitting methods based on either SUPPORT or GUIDE are used to construct the regression trees in this study. The look-ahead method utilizing the cross-validation (Chaudhuri et al., 1994; see also Ahn and Loh, 1994; Ahn and Chen, 1997) is used for determining the size of the trees.

3.1 Splitting Rules

SUPPORT detects the differences of mean and variance between residual distributions of two groups at each subnode and decides the most significant predictor for a split. At each node of a binary regression tree, a covariate vector is assigned to class 1 if the corresponding modified adjusted Anscombe residual of the extra-Poisson regression is larger than the median of the residual given

in (4) and to class 2 otherwise. For each covariate, the two-sample t-test for differences in means and Levene's tests (1960) for differences in variances of each covariate are performed to detect the heterogeneity of the two groups, and the p-value from the larger of the two statistics is computed. For the covariate with the smallest p-value, the data in the node are split into two parts, with one subset containing all cases with that covariate value less than the average of the two sample means, and the other containing the remaining cases. This process is repeated at each subsequent node until either the smallest p-value is less than the significance level determined by cross-validation or there are too few cases left at the node.

GUIDE (Loh, 2002) is proposed to eliminate the variable selection bias. GUIDE has advantages compared to SUPPORT in two ways. GUIDE can be applied for a categorical predictor which is not appropriately dealt with in SUPPORT. Further, it can include interactions of the variables as splitting variables. At each node, the χ^2 statistic is computed for each predictor and interaction to decide a splitting variable. GUIDE limits a predictor's role as n-variable (a numerical predictor used for both regressor and split), f-variable (a numerical predictor used for only regressor, not for split), s-variable (a numerical predictor used for only split) and c-variable (a categorical predictor used for only split). An extra-Poisson regression model is fit to the data and the residuals are calculated at the node. The χ^2 statistic and p-value are calculated based on the signs of the residuals according to the type of the predictor and/or interaction of the predictors. The predictor with the smallest p-value is chosen as the split variable. See Loh (2002) for a detail.

A categorical predictor is used for only split. An ordered predictor can play a role of s-variable. If an n-variable competes with s- or c-variables, the variable selection will be biased. Since residual represents the part which is not explained by a fitted model, the residual distribution of an n-variable is more homogeneous than other variables and has a less chance to be selected for a split. To correct this bias, GUIDE uses the constant factor. After p-values are obtained from the χ^2 test, they are converted into two-tailed z-values and multiplied by γ (> 1) to the z-values for only the n-variables. Loh (2002) estimates the optimal γ value by bootstrap resampling. The bootstrap calibration for γ is carried out at the root node in this study.

3.2 Stopping Rule

In order to determine whether a node should be split or declared terminal, a V-fold cross-validation is used. The cases in the node are randomly divided into V subsets. A nested sequence of trees is

constructed with the data from $(V-1)$ subsets and the remaining subsets are used as test sample to decide if the data in the former should be partitioned. This procedure is applied V times, each time leaving out a different subset as test sample. Due to the variability caused by cross-validation, a cross-validation tree is preferred to the trivial tree (a tree consisting of the root node only) if it has the measure of goodness of fit that is at least $(1-f)$ times smaller than that for the trivial tree. If the frequency (out of V) that a preferred cross-validation tree exceeds another pre-specified number η, the node is split. The pruning parameters $f \in [0,1]$ and $\eta \in [0,1]$ are estimated by bootstrap.

In this paper, we modify the bootstrap estimation method of the pruning parameters f and η introduced by Ahn and Loh (1994). The hypothesis that a node is split when a single extra-Poisson regression model suffices for the sample in the node is tested. The probability of a Type I error is defined as

$$\alpha = P(\text{split the node} \mid H_0 : \text{ the node should not be split})$$

at each node of the tree. By evaluating this probability using different values of f and η via bootstrap resampling, the parameter values are chosen to be the value for which the probability is closest to the nominal significance level α using the grid of 0.1. In this paper, we fix $f = \eta$ and select the optimal parameter value because this searching approach turned out to perform better than the other two search methods proposed by Ahn and Loh. Ahn and Chen (1997) used the nested grid with an increment of 0.01 since the tree size is sensitive to the values of f and η. In this paper, the nested grid is not applied to η because the change of η in the nested grid does not affect the tree size with a 10-fold cross-validation. Further details on this bootstrap method is provided in Ahn and Loh (1994) and Choi (2002).

The estimates $\hat{f}$ and $\hat{\eta}$ of the splitting parameters at the root node can be viewed as a cost of split. Hence, if the split at the node makes the error reduction less than the threshold given by f and η, the node will not be split in the look-ahead procedure. Since the bootstrap methods introduced by Ahn and Loh estimates these parameters at the root node, it can control the probability of a Type I error (split of the root node when it should not be split). However, these values from the root node may not be sufficient to control the splits at the subsequent nodes properly. In this paper, we apply the bootstrap estimation method at each node so that the values of the bootstrap parameters are updated based on the current sample. This approach is expected to control the size of the tree more efficiently. Although a combination of cross-validation and bootstrap resampling at each node

is extremely computer intensive task, the rapid improvement of computing speed and memory space enables this approach.

3.3 Estimation of the Over-Dispersion

We assume a constant over-dispersion for the whole sample in this paper. The common over-dispersion parameter can be estimated for the entire sample at the root node and this estimated value is used at the subsequent nodes without estimating it again for the current subset of the data. This approach is more reliable than estimating the over-dispersion parameter at each node because the estimation is based on more information from the whole sample. However, this may not be a good approach if the over-dispersion is substantially different in each stratum. Another approach is to consider estimation of the over-dispersion parameter at each node. A drawback of this second approach is that the over-dispersion parameter can be poorly estimated if it is based on a very small sample, and it would cause highly biased test results for the regression coefficients. The estimation of the over-dispersion parameter may be highly influenced by the fluctuation of the data points if the sample is too small. We can see from (1) that the variance is increased by the over-dispersion and decreased by the under-dispersion from the standard loglinear model.

4 Simulation Study

A Monte Carlo simulation study is conducted to evaluate the performance of the proposed loglinear regression trees. For data with over-dispersed Poisson variates, the extra-Poisson tree is compared with the Poisson regression tree in order to investigate the improved performance of the tree by considering the over-dispersion in the model. The Poisson regression tree and extra-Poisson regression tree are also evaluated with data from a Poisson distribution without over-dispersion. For the stopping rule, the look-ahead method is compared with the backward-elimination method.

The simulation study consists of two parts. In the first part, data are generated from one loglinear model. In the second part, data are generated from more than one loglinear model. No split is expected from the trees in the first part, but at least one split is expected from the second part. We evaluate the Type I error rate from the first part and the power of the regression trees in the second part. Each part is described in Sections 4.1 and 4.2. The simulated models in these two sections have only numerical predictors. Hence SUPPORT is used for variable selection in these

sections.

In addition to the study of Type I error rate and power, we investigate the tree sizes obtained by our methods. The existing bootstrap estimation method carried out at the root node shows a reasonable control of the Type I error rate. However, the control of the tree size is expected to be improved by the proposed bootstrap estimation method by updating the parameter values at each node. In this simulation study, our bootstrap method is compared with the existing method.

Section 4.3 presents a comparison of two different splitting methods using residual distributions. One is based on SUPPORT and the other is based on GUIDE. SUPPORT employs the Levene's statistic and two sample t-test, and GUIDE uses the chi-squared test (see Loh, 2002 for a detail) in the analysis of the residuals. We examine how the split affects the Type I error rate and power, especially when a design matrix has categorical variables. The look-ahead and backward-elimination methods are considered in each splitting method.

Two hundred data sets are generated for each model given in the subsequent sections with the standard Poisson and extra-Poisson responses. Each data set contains 120 data points. For all the examples in this paper, 10-fold cross-validation is used and the value of α in the bootstrap is chosen to be 0.05. For a Poisson response, Y is generated from the Poisson distribution with mean μ. For the extra-Poisson data, The response Y is distributed with the over-dispersion ϕ which is set to be 1 in this simulation study. The response Y with $\mathrm{Var}(Y) = \mu(1+\phi\mu)$ is generated using the negative binomial model. For a given value of μ, suppose the distribution of Y given μ is Poisson(μ) and the prior distribution of μ is Gamma($1/\phi, \mu\phi$), then from Section 2, Y has an extra-Poisson variation.

The extra-Poisson random variate is generated sequentially as follows. First, a random number y_i from a gamma distribution with $(1/\phi,\ \mu\phi)$ is generated. Using the value of this random number as the mean, a Poisson random number is generated. The resulting random number follows an extra-Poisson variation with mean μ and over-dispersion of ϕ.

Since the gamma random number generator allows only an integer value of α in generating a Gamma(α, γ) distribution, two different algorithms are considered for generating a gamma random variable depending on an integer and a non-integer value of α. If α is an integer, the gamma distribution becomes the sum of exponential distributions with mean γ. Thus a gamma distribution can be generated by a transformation, which is

$$-\sum_{i=1}^{\alpha} \log U_i/\gamma \sim \mathrm{Gamma}(\alpha, \gamma),$$

where $U_i \sim \text{Uniform}(0,1)$, $i = 1, \ldots, \alpha$. If α is not an integer, the Accept-Reject Method (Robert and Casella, 1999) is applied.

4.1 Data from One Loglinear Model

The following four models are considered in this section.

Model 1: $\mu_i = \exp(-0.6931)$

Model 2: $\mu_i = \exp(-0.6931 + 0.497 d_i)$, where $d_i \in \{0, 1, 2, 3, 4, 5\}$

Model 3: $\mu_i = \exp(-0.6931 + 0.497 X_i)$, where $X_i \sim \text{Uniform}(0, 5)$

Model 4: $\mu_i = \exp(-0.6931 + 0.497 d_i + 0.597 X_i)$, where $X_i \sim \text{Uniform}(0, 3)$.

The above simulation models are based on the context of a dose response experiment. Models 1 and 2 are based on the model $\mu_i = \exp(\beta_0 + \beta_1 d_i)$, $i = 1, \ldots, n$, where $\boldsymbol{\beta} = (\beta_0, \beta_1) = (-0.6931, 0)$ for Model 1 and $\boldsymbol{\beta} = (-0.6931, 0.4970)$ for Model 2. For Model 1, the values of β_0 and β_1 give a constant mean $\mu_i = 0.5$, and for Model 2, the parameter values give $\mu_i = 0.5$ at $d_i = 0$ and $\mu_i = 6$ at $d_i = 5$. Model 3 replaces the discrete variable d_i in Model 2 to the continuous variable X. For Model 4, the parameters give $\mu_i = 0.5$ at $(d_i = 0, X = 0)$, $\mu_i = 3$ at $(d_1 = 0, X = 3)$ and $\mu_i = 6$ at $(d_i = 5, X = 0)$.

Because the data are generated from a loglinear model for all the cases, a single Poisson regression fit is sufficient and no split is expected. Simulation results are given in Table 1. For the look-ahead approach, the Type I error rate is supposed to be the same for the proposed bootstrap estimation method and the old bootstrap method except for the random error because this is based on the pruning parameters obtained at the root node. The extra-Poisson tree with the look-ahead method controls the Type I error rate satisfactorily, while the backward-elimination method fails to control it. The trees with Models 1 or 2 are either trivial or with one split since the models include only one predictor with 6 possible distinct values. The backward-elimination method often selects the tree with the minimum error even when the split is insignificant compared to the cost of split in these examples. However, the look-ahead procedure splits the root node if a significant error reduction is made since the f and η are estimated in the bootstrap and appropriately used as a cost of excessive split. For Models 3 and 4 which include a continuous predictor, the backward-elimination gives a better result than for Models 1 and 2. Since the predictor is continuous, various sizes of subtrees are

Table 1: Simulated Type I error rate (% trivial trees) with different tree-structured methods for data from null models. Nominal significance level is $\alpha = 0.05$; 10-fold cross-validation; 200 trials.

		Model with $\phi > 0$		Model with $\phi = 0$	
Data	Case	B-E[1]	L-A[2]	B-E	L-A
$\phi > 0$	Model 1	81.0	4.5	80.5	7.0
	Model 2	81.0	5.5	81.5	9.0
	Model 3	7.5	5.0	11.0	8.5
	Model 4	7.5	4.0	11.0	14.5
$\phi = 0$	Model 1	78.0	4.5	76.5	4.5
	Model 2	80.5	2.5	80.5	3.0
	Model 3	8.0	5.0	8.5	6.5
	Model 4	4.5	1.5	8.5	2.5

[1] backward-elimination method
[2] look-ahead method

produced in cross-validation. It manages to avoid selecting a nontrivial tree. Even for this model, the look-ahead procedure keeps the Type I error rate better than the backward-elimination. The extra-Poisson regression tree with the look-ahead procedure performs well even for the Poisson data without over-dispersion because the extra-Poisson model reduces to Poisson when the value of the over-dispersion parameter becomes zero.

The Type I error rate of the Poisson regression tree is quite high for the data with over-dispersion although it controls the Type I error rate quite well for the standard Poisson data. These results indicate that the over-dispersion should be included in the model when the data possess over-dispersion.

4.2 Data from More Than One Loglinear Model

In this section, the following four models are considered.

Model 5:

$$\mu_i = \begin{cases} \exp(\beta_0 + \beta_1 d_i) & \text{if } d_i = 0, 1, 2 \\ \exp(\gamma_0 + \gamma_1 d_i) & \text{if } d_i = 3, 4, 5 \end{cases}$$

where $(\beta_0, \beta_1) = (-0.6931, 0)$ and $(\gamma_0, \gamma_1) = (-3.178, 0.994)$. The values of β_0 and β_1 give a constant mean $\mu_i = 0.5$ for $d_i \in (0, 2.5)$ and the values of γ_0 and γ_1 give the mean $\mu_i = 0.5$ at $d_i = 2.5$ and $\mu_i = 6$ at $d_i = 5$.

Model 6:

$$\mu_i = \begin{cases} \exp(\beta_0 + \beta_1 X_i) & \text{if } X_i \leq 2.5 \\ \exp(\gamma_0 + \gamma_1 X_i) & \text{if } X_i \geq 2.5 \end{cases}$$

where β_i's and γ_i's are the same as in Model 4.

Model 7: $\mu_i = \exp(|X_i|)$, where $X_i \sim$ Uniform$(-2, 2)$

Model 8:

$$\mu_i = \begin{cases} \exp(2|X_i|) & \text{if } -1 < X_i < 1 \\ \exp(2X_i + 4) & \text{if } -2 < X_i < -1 \\ \exp(-2X_i + 4) & \text{if } 1 < X_i < 2 \end{cases}$$

For the above models, it is expected that a single loglinear regression fit to the root node is not adequate and each tree should have a split. Models 5 through 7 are expected to give one split and Model 8 is expected to give three splits. Table 2 displays the simulation results. The expected number of splits is denoted with boldface. The power can be obtained by adding the frequencies of the trees with at least one split. The power is supposed to be the same for both the bootstrap methods for the look-ahead approach except the random error. The power of the look-ahead method and the backward-elimination method are comparable. For extra-Poisson trees, they are 100%, 60%, 69% and 70.5% for the backward-elimination method, and 69%, 49.5%, 94% and 66.5% for the look-ahead method with the new bootstrap estimation method. For the look-ahead method, the trees with the new bootstrap estimation method produces the right-sized tree more often than the trees with the existing bootstrap method. For the over-dispersed Poisson data, the extra-Poisson trees produce exactly one split for 15% and 58.5% from the old bootstrap method and 36.5% and 83% from the new bootstrap method for Models 6 and 7, respectively. For Model 8, the extra-Poisson trees produce three splits for 10.5% and 31% for the old and new bootstrap methods, respectively. We observe similar results from the Poisson regression trees.

4.3 Comparison of Two Splitting Methods

Choice of the split variable and the splitting point at each node is important in tree-structured regression. Since inefficient splits do not generate enough error reduction for a split, the resulting

Table 2: Comparison of the tree sizes with different tree-structured methods for data from more than one extra-Poisson model. Entries are frequencies (%) of splits. The expected number of splits is denoted with boldface. Nominal significance level is $\alpha = 0.05$; 10-fold cross-validation; 200 trials.

			Model with $\phi > 0$			Model with $\phi = 0$		
Data	Case	#splits	B-E[1]	L-A old[2]	L-A new[3]	B-E	L-A old	L-A new
$\phi > 0$	Model 5	0	0.0	31.0	31.0	1.0	30.0	30.0
		1	100.0	69.0	69.0	99.0	70.0	70.0
	Model 6	0	40.0	49.5	50.5	40.5	42.0	44.0
		1	42.5	15.0	36.5	39.5	16.0	39.0
		2	8.5	5.0	5.5	8.0	3.0	8.5
		≥ 3	9.0	30.5	7.5	12.0	39.0	8.5
	Model 7	0	31.0	7.0	6.0	9.5	2.5	3.0
		1	62.5	58.5	83.0	77.5	55.0	79.0
		2	2.5	5.0	6.0	3.0	7.0	9.5
		≥ 3	4.0	29.5	5.0	10.0	35.5	8.5
	Model 8	0	29.5	33.0	33.5	15.0	25.0	27.0
		1	0.5	0.0	5.0	0.0	0.5	4.5
		2	19.5	3.5	24.5	27.5	2.0	22.5
		3	40.0	10.5	31.0	34.5	6.5	28.0
		≥ 4	10.5	53.0	6.0	23.0	66.0	18.0
$\phi = 0$	Model 5	0	0.0	5.5	6.5	0.0	7.0	6.5
		1	100.0	94.5	93.5	100.0	93.0	93.5
	Model 6	0	48.5	52.5	54.5	47.5	54.5	55.5
		1	42.5	32.0	39.0	45.5	30.5	37.5
		2	2.0	2.0	3.5	2.0	1.5	3.0
		≥ 3	7.0	13.5	3.0	5.0	13.5	4.0
	Model 7	0	3.5	0.0	0.0	1.5	0.0	0.0
		1	87.5	65.0	88.0	89.0	68.5	91.0
		2	2.0	5.5	9.0	2.0	6.5	6.0
		≥ 3	7.0	29.5	3.0	7.5	25.0	3.0
	Model 8	0	2.5	0.0	0.0	1.0	0.0	0.0
		1	0.0	0.0	0.0	0.0	0.0	0.0
		2	2.5	0.0	0.0	1.5	0.0	0.0
		3	82.0	20.0	76.5	82.0	19.5	76.5
		≥ 4	13.0	80.0	23.5	15.5	80.5	23.5

[1] backward-elimination method
[2] look-ahead method with bootstrap at the root node
[3] look-ahead method with bootstrap at each node

tree may be larger or shorter than the optimal tree depending on situations.

Both the methods based on SUPPORT and GUIDE use residual distributions for choosing a splitting variable. SUPPORT detects the difference of the mean and variance between residual distributions of two groups at each subnode and decides the most significant predictor for a split. GUIDE tests if there is any association between the residual distribution and each predictor. These two tests detect different patterns in residual distributions similarly, but SUPPORT is not appropriate for categorical predictors. Moreover, GUIDE includes a test for interaction terms as a split.

To compare these two splitting methods, the following three models are simulated.

Model 9: $\mu_i = \exp(-0.5 + 0.8X_{1i} - 0.5X_{2i})$ where $X_{1i}, X_{2i} \sim \text{Uniform}(0, 3)$

Model 10: $\mu_i = \exp\{(-1)^{C_i}(0.4X_{1i} - 0.4X_{2i})\}$
where $C_i = \{0,1\}$, $X_{1i} \sim \text{Uniform}(0, 5)$ and $X_{2i} \sim \text{Uniform}(0, 2)$

Model 11: $\mu_i = \exp(2.5|X_{1i}| - 2X_{2i})$ where $X_{1i} \sim \text{Uniform}(-1, 1), X_{2i} \sim \text{Uniform}(0, 1)$.

Two hundred data sets are simulated for each Model. Model 9 has only numerical predictor. For Model 9, the Type I error rate is obtained by SUPPORT and GUIDE with both look-ahead and backward-elimination procedures. Model 10 contains one categorical variable and two numerical variables. In SUPPORT, the categorical variable C plays a role for both a regressor and a split variable. In GUIDE, C is used only for split. It does not appear in the model at any node. For Model 10, power is obtained according to the variable selection at the root node. Model 11 contains two numerical predictors and the power is calculated for this model.

The results are given in Table 3. For Model 9, the Type I error rate is not significantly affected by the splitting method. It is reasonable because Model 9 has numerical predictors to be non-significant about split.

Choice of the splitting method is very influential in Model 10. GUIDE gives higher power than SUPPORT for both the look-ahead and backward-elimination procedures. GUIDE chooses C as the splitting variable for 187 out of 200 simulation data sets at the root node, while SUPPORT chooses only 135 data sets. This shows that SUPPORT often fails to choose the correct splitting variable when a categorical predictor is supposed to be split. Model 11 contains two numerical predictors. GUIDE selects the correct split variable X_1 193 out of 200 data sets at the root node, while SUPPORT selects 189. The powers of GUIDE and SUPPORT are similar in this model.

Table 3: Simulated Type I error rate and power (% trivial trees) with different split methods. Nominal significance level is $\alpha = 0.05$; 10-fold cross-validation; 200 trials.

			L-A[1]		B-E[2]	
Model	Data	Case	GUIDE	SUPPORT	GUIDE	SUPPORT
$\phi > 0$	$\phi > 0$	Model 9	3.5	4.5	5.0	5.0
		Model 10	94.5	63.5	69.5	45.5
		Model 11	95.5	93.5	64.0	74.0
	$\phi = 0$	Model 9	3.5	2.5	8.5	6.5
		Model 10	100.0	86.0	89.5	72.5
		Model 11	100.0	100.0	90.0	91.0
$\phi = 0$	$\phi > 0$	Model 9	8.5	7.0	6.5	9.5
		Model 10	96.5	66.5	87.0	34.0
		Model 11	97.5	97.5	82.0	83.0
	$\phi = 0$	Model 9	4.5	4.5	7.5	3.5
		Model 10	100.0	89.5	95.0	72.5
		Model 11	100.0	100.0	94.0	94.5

[1]backward-elimination method
[2]look-ahead method

Regardless of the splitting method, the look-ahead method provides better power than the backward-elimination procedure for Models 10 and 11.

5 Missouri Lung Cancer Data

The proposed extra-Poisson regression tree procedure is applied to the data discussed in Section 1 on lung cancer mortality in Missouri. Parts of the data provided on our web site were given in Tsutakawa (1985, 1988). The mortality frequencies due to lung cancer among males aged 45-54 in eighty-four of the largest cities in Missouri were given in Tsutakawa (1985), and the male data of the 115 counties were given in Tsutakawa (1988). Tsutakawa (1988) analyzed the male data using a mixed model for finding geographic variability in mortality rates. In addition to demographic parameters and random geographic parameters, his gamma-Poisson model includes additional random-effects parameters to adjust for extra-Poisson variability.

As mentioned in Section 1, Figure 1 indicates that one log-linear model may not be sufficient to explain the effect of public drinking water on cancer mortality. The pattern of the cancer mortality appears to be different in urban and rural areas, different sexes, and old and young residents. Thus, our tree-structured model might be suitable for analyzing these data. In this paper, the number of

deaths from lung cancer per 1000 persons in the 115 counties of Missouri is considered the response variable. The rates are used instead of counts because the relationship between the lung cancer mortality rate (not the number of lung cancer cases in the county) and population is of interest. These rates can be viewed as approximately Poisson because they are obtained based on a relatively large population ($n = 1000$). There are 920 cases because the numbers are obtained for four age groups and both sexes in each county.

Three predictors, sex, age, and population, are included in the regression model in the SUPPORT tree, while only age and population are regressors in the GUIDE tree. Sex is used for split in GUIDE. Sex is coded '0' for male and '1' for female. The mid-point is used for each age group (for the group of age > 75, 80 is used) in this analysis. GUIDE tree might be more appropriate for these data because a categorical variable is included. However, we will study both GUIDE and SUPPORT in order to find the difference caused by handling the categorical variable differently.

Throughout the section, the analysis is based on the robust variance estimate since it is shown to be consistent even if the variance structure is misspecified (Liang and Zeger, 1986). Before presenting the analysis, let node(i, j) denote the jth node from the left (including the empty nodes) at the ith level of the tree. The root node is node(1, 1).

5.1 Extra-Poisson Regression Trees Using SUPPORT

With the extra-Poisson regression model, the look-ahead procedure generates a different tree from the backward-elimination method. However, both the procedures show different patterns of the lung cancer mortality for the ages over and under 65. The estimate of ϕ at the root node is 0.1. The estimate of ϕ using the μ_i's obtained at the terminal nodes is 0.043 from both the methods and the z-value for testing H: $\phi = 0$ is 18.00 (p-value ≈ 0). This suggests that the data contain over-dispersion. Therefore, the extra-Poisson tree seems to be more suitable than the Poisson regression tree for the data.

5.1.1 Look-ahead procedure

For the look-ahead procedure, our new bootstrap estimation method for estimating f and η is performed at each node. Figure 2 shows the tree generated by this method. In this tree, the root node is split at age 65. For the groups of age under 65, the node is further split on sex, and for the groups of age over 65, the node is split on age (over and under 75). The node for age > 75 is further

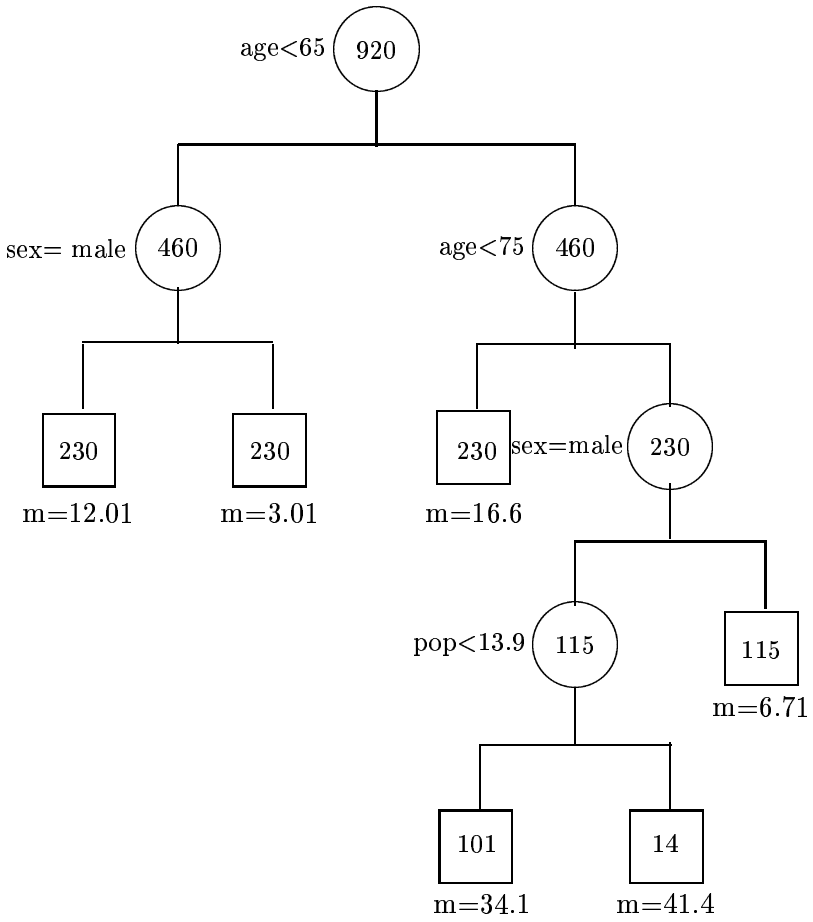

Figure 2: The extra-Poisson tree using SUPPORT with the look-ahead procedure for the analysis of lung cancer mortality in Missouri. The bootstrap estimation is performed at each node. Population is in thousands. The number in each node is the sample size and the number below each terminal node is the median number of deaths per 1000.

split on sex. For males with age over 75, the node is split on population (over and under 13,900). Table 4 presents the estimated values of the bootstrap parameters at each node of the tree for the data. The bootstrap estimation is not conducted at node(5,14) because it is declared as terminal due to a small sample size. This table shows that the parameter estimates are quite different at each node. This implies that the updated parameters should be used in order to obtain a tree with an optimal size.

Table 5 shows the regression estimates of the coefficients at the terminal nodes for the data using this method. In general, the size of county (in terms of population) adversely affects the cancer mortality. For the people of age under 65, age is a significant factor on lung cancer mortality. For the group of age over 75, population is marginally significant for females, while it is highly significant for males. In younger patients, population is a significant factor. From Table 5 and the median

Table 4: Bootstrap parameter estimates at each node for the extra-Poisson tree in Figure 2.

Node	f	η	Sample size
(1,1)	0.055	0.1	920
(2,1)	0.065	0.2	460
(2,2)	0.090	0.1	460
(3,1)	0.125	0.2	230
(3,2)	0.115	0.2	230
(3,3)	0.125	0.2	230
(3,4)	0.135	0.2	230
(4,7)	0.155	0.3	115
(4,8)	0.155	0.3	115
(5,13)	0.17	0.3	101

Table 5: The estimates of the regression coefficients at the terminal nodes of Figure 2. The test is based on the robust variance estimates.

Node		Constant	Sex	Age	Population
node(3,1)	$\hat{\beta}$	-3.25		0.104	0.0012
age <65	se	0.28		4.93E-5	5.47E-4
male	$\hat{\beta}$/se	-11.5		21.1	2.19
	p-value	$\simeq 0$		$\simeq 0$	0.029
node(3,2)	$\hat{\beta}$	-2.87		0.073	0.0023
age <65	se	0.52		0.009	3.99E-4
female	$\hat{\beta}$/se	-5.49		7.98	5.88
	p-value	$\simeq 0$		$\simeq 0$	$\simeq 0$
node(3,3)	$\hat{\beta}$	3.59	-1.93		0.0018
$65 \leq$ age <75	se	0.022	0.003		4.2E-4
	$\hat{\beta}$/se	159.8	-33.7		4.23
	p-value	$\simeq 0$	$\simeq 0$		$\simeq 0$
node(4,8)	$\hat{\beta}$	1.93			9.38E-4
age ≥ 75	se	0.062			4.65E-4
female	$\hat{\beta}$/se	31.1			2.02
	p-value	$\simeq 0$			0.044
node(5,13)	$\hat{\beta}$	3.28			0.037
age ≥ 75	se	0.076			0.0098
male	$\hat{\beta}$/se	43.4			3.83
pop <13,900	p-value	$\simeq 0$			0.0001
node(5,14)	$\hat{\beta}$	3.68			7.8E-4
age ≥ 75	se	0.036			2.19E-4
male	$\hat{\beta}$/se	103.5			3.54
pop $\geq$13,900	p-value	$\simeq 0$			0.0004

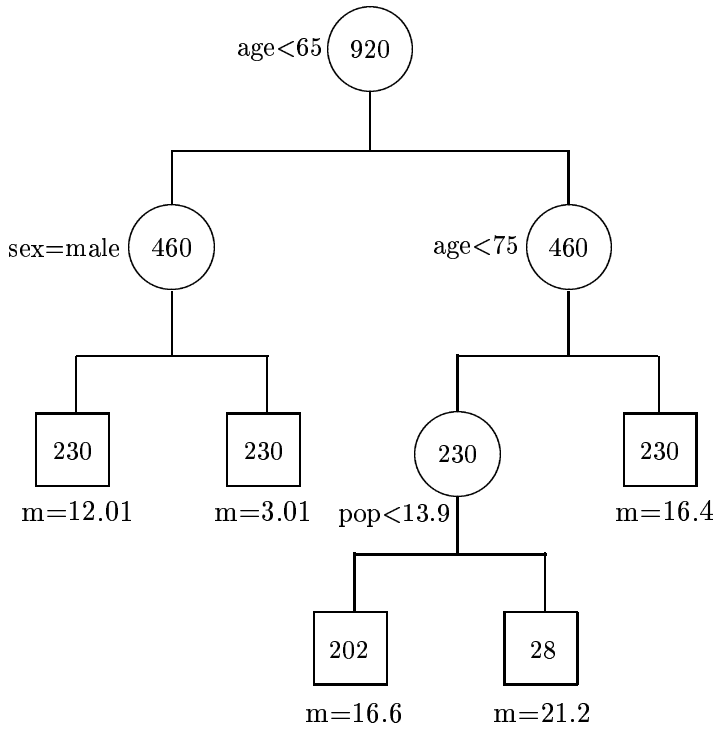

Figure 3: The extra-Poisson tree using SUPPORT with the backward-elimination method for the analysis of lung cancer mortality in Missouri. Population is in thousands. The number in each node is the sample size and the number below each terminal node is the median number of deaths per 1000.

number of deaths at each node given in Figure 2, we see that males have substantially higher lung cancer mortality than females.

When the bootstrap estimation of f and η is performed at only the root node, an excessively large extra-Poisson tree is obtained with 8 levels and 20 terminal nodes. The splits at the root node and the nodes at the next level occur at the same places as the tree in Figure 2. We do not provide this tree in this paper. This result supports the need for estimating the bootstrap parameters at each node. Table 4 shows that the bootstrap parameter values are increased at the lower nodes in order to prevent excessive splits when we estimate the parameters at each node. This result agrees with the finding from our simulation that our new bootstrap method gives tree with the optimal size more often than the old bootstrap method.

5.1.2 Backward-elimination method

As seen in Figure 3, the backward-elimination method generates a different regression tree from the look-ahead procedure with our new bootstrap estimation approach. This method splits the nodes for $65 \leq$age < 75 and age ≥ 75 differently from the look-ahead method. While the node for age

over 75 is not split any more, the node for 65 $\leq$age $<$ 75 is split at population 13,900. According to the median survival rates in Figure 3, the lung cancer occurs more frequently for more densely populated areas. As observed from the tree with the look-ahead method, males have a higher risk of death from lung cancer.

5.2 Extra-Poisson Regression Trees Using GUIDE

The first split by GUIDE occurs at sex, while it occurred at age in SUPPORT. The main reason for this difference is that sex is used as a splitting variable, but not included in the regression in GUIDE. Based on the considerably different patterns between males and females in lung cancer mortality shown in Figure 1, it is not surprising that the first split occurs on sex. The look-ahead procedure generates a similar size of tree as the backward-elimination method. For both methods, the first and the second splits occur at sex and age, respectively, and the subsequent splits occur at population. However, the split patterns are different. In the look-ahead procedure, males with age over 65 have more complicate splits according to the population size while females with age less than 65 have many splits in the backward-elimination method. The estimates of ϕ using the μ_i's obtained at the terminal nodes are 0.033 (p-value ≈ 0) and 0.042 (p-value ≈ 0) from the look-ahead and backward-elimination methods, respectively.

5.2.1 Look-ahead procedure

Figure 4 shows the tree generated by the look-ahead procedure using GUIDE. Table 6 presents the regression estimates of the coefficients at the terminal nodes for the data using this method. The estimates for node(3,3) is excluded because it is included in Table 5. The relationship between lung cancer mortality and predictors at each node is similar to that in the SUPPORT tree. According to Figure 4 and Table 6, males with age over 65 in the area with population more than 13,900 have the highest lung cancer mortality and the females under 65 have the lowest lung cancer mortality. The bootstrap estimates for f and η at each node are quite different as in the SUPPORT trees. Choi (2002) provides the estimates.

5.2.2 Backward-elimination method

Figure 5 displys the tree obtained from this method. Each sex and age under or over 65 groups make different patterns in lung cancer mortality. Females with age under 65 have further splits on

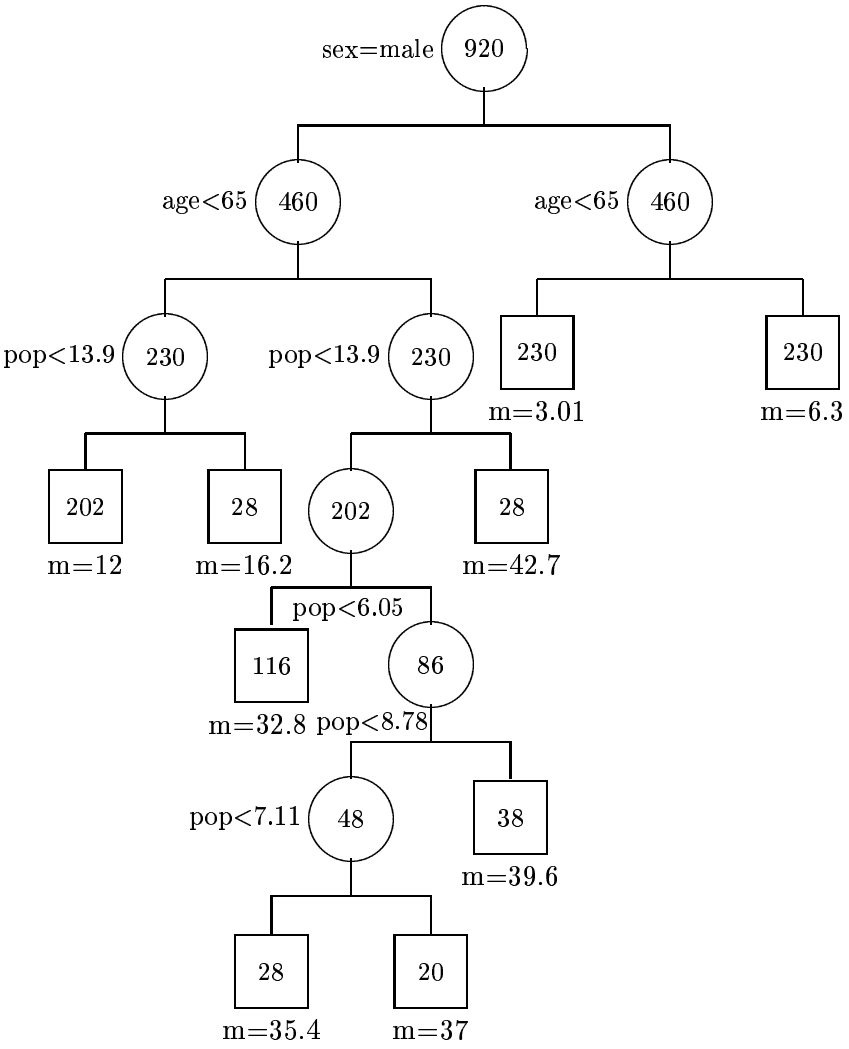

Figure 4: The extra-Poisson tree using GUIDE with the look-ahead method for the analysis of lung cancer mortality in Missouri. Population is in thousands. The number in each node is the sample size and the number below each terminal node is the median number of deaths per 1000.

Table 6: The estimates of the regression coefficients at the terminal nodes of Figure 4. The test is based on the robust variance estimates.

Node		Constant	Age	Population
node(3,4)	$\hat{\beta}$	-0.10	0.025	0.0019
age$\geq$ 65	se	0.58	0.0078	4.20E-4
female	$\hat{\beta}$/se	-0.17	3.22	4.53
	p-value	0.86	0.0013	$\simeq 0$
node(4,1)	$\hat{\beta}$	-3.40	0.11	0.022
age$<$ 65	se	0.33	0.0055	0.0077
male	$\hat{\beta}$/se	-10.3	18.9	2.87
pop$<$ 13.9	p-value	$\simeq 0$	$\simeq 0$	0.0042
node(4,2)	$\hat{\beta}$	-3.07	0.10	8.55E-4
age$<$ 65	se	0.48	0.0087	5.92E-4
male	$\hat{\beta}$/se	-6.33	11.77	1.44
pop$\geq$ 13.9	p-value	$\simeq 0$	$\simeq 0$	0.15
node(4,4)	$\hat{\beta}$	4.11	-0.0052	4.71E-4
age$\geq$ 65	se	0.35	0.0044	2.10E-4
male	$\hat{\beta}$/se	11.6	-1.16	2.24
pop$\geq$ 13.9	p-value	$\simeq 0$	0.25	0.025
node(5,5)	$\hat{\beta}$	3.95	-0.010	0.078
age$\geq$ 65	se	0.42	0.0055	0.026
male	$\hat{\beta}$/se	9.31	-1.87	3.04
pop$<$ 6.05	p-value	$\simeq 0$	0.061	0.0024
node(6,12)	$\hat{\beta}$	3.49	0.0022	0.0013
age$\geq$ 65	se	0.46	0.0062	0.021
male	$\hat{\beta}$/se	7.59	0.36	0.063
$8.78 \leq$ pop $<$ 13.9	p-value	$\simeq 0$	0.72	0.95
node(7,21)	$\hat{\beta}$	4.11	-0.019	0.13
age$\geq$ 65	se	1.15	0.0093	0.15
male	$\hat{\beta}$/se	3.56	-2.00	0.90
$6.05 \leq$ pop $<$ 7.11	p-value	0.004	0.045	0.37
node(7,22)	$\hat{\beta}$	2.95	0.0014	0.071
age$\geq$ 65	se	0.98	0.0087	0.11
male	$\hat{\beta}$/se	2.99	0.16	0.64
7.11 $\leq$pop$<$ 8.78	p-value	0.0027	0.88	0.52

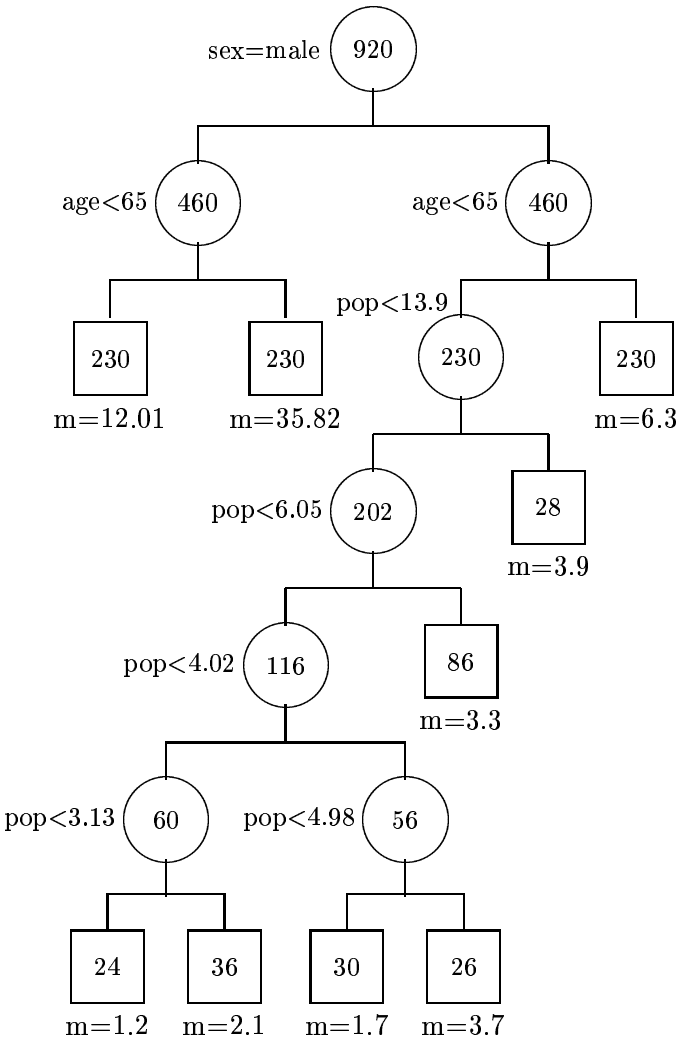

Figure 5: The extra-Poisson tree from the backward-elimination method using GUIDE for the analysis of lung cancer mortality in Missouri. Population is in thousands. The number in each node is the sample size and the number below each terminal node is the median number of deaths per 1000.

population. The estimates of regression coefficients at the terminal nodes are not reported here, but provided in Choi (2002).

6 Discussion

The primary goal of this research is to develop piecewise regression models for testing carcinogenic potential of environmental and geographic effects. To achieve this goal, we developed the loglinear regression tree methods for analysis of count data with extra-variation. The extra-Poisson regression model is fit at each stratum by a recursive stratification. This approach provides an easier interpretation of covariates on response and avoids strong assumptions on models.

Ahn and Loh (1994) conducts bootstrap estimation of the pruning parameters at the root node and use it as the cost of split at the subsequent nodes. While this method controls probability of a Type I error (false split of the root node) appropriately, it often fails to produce the tree with the optimal size. In this paper, these parameters are updated at each node for constructing a tree with a desirable size.

A new moment estimation method for the over-dispersion parameter is developed. It is easier to compute than the existing estimator. The existing estimate requires a numerical solution, and does not always give the solution. A test for the over-dispersion is also developed for the proposed extra-Poisson regression trees. If the over-dispersion is not significant, we may use the standard Poisson regression trees because the model is simpler and it will prevent an overfit by adding an unnecessary parameter. However, the extra-Poisson regression tree can also be used for such data because the Poisson model is a special case of the extra-Poisson model, and including ϕ may prevent the Type I error.

Generalized linear regression models are often used to analyze count data. However, one regression model for the whole sample is often difficult to interpret, especially when there are numerous covariates, and some of them are correlated. Since the generalized regression trees by Chaudhuri et al. (1995) do not take into account the extra variation, the test of regression parameters often fails to control Type I error rate (Choi, 2002). In order to overcome this problem, we developed extra-Poisson regression trees.

According to our simulation, the look-ahead splitting method adopted in this study performs better than the backward-elimination method in terms of power and control of Type I error rate.

The proposed bootstrap method produces trees with the correct size more often than the trees with the existing bootstrap method. As expected, SUPPORT turned out to be not appropriate for categorical predictors. For the model with a categorical predictor which is supposed to be split, the power of SUPPORT is lower than that of GUIDE.

The simulation and data analysis were carried out on an Ultra Spark 60 SUN Workstation, an Alpha Workstation DS20E 833 MHZ machine, and a departmental super computer system with 256 CPU's of Pentium III's. For each model, it took 2 to 3 hours for simulating 200 data sets for Models 1 through 4, and less than 24 hours for the other models using the super computer. For the Missouri lung-cancer data, computing times for obtaining the trees on the super computer were 4 minutes for SUPPORT and 6 minutes for GUIDE with the look-ahead method, and 7 seconds for SUPPORT and less than a second for GUIDE with backward-elimination method on an Alpha Workstation.

Geographic variability in mortality rates on lung cancer as well as effect of various complicated demographic variability is investigated in this study. The count data discussed in Section 5 exhibit a greater variability than would be predicted by simply fitting a Poisson model. Therefore, an extra-Poisson regression tree appears to be appropriate for analyzing these data. Both trees based on SUPPORT and GUIDE seem to perform well in explaining the lung cancer mortality at each subgroup of the Missouri population. However, there is a substantial difference between the trees from the two methods. The main reason is the inclusion of categorical variables which can be more appropriately handled by GUIDE. GUIDE is recommended since it is developed for reducing the selection bias, especially in presence of a categorical variable such as sex. Our results show that most cancer types affect different age-sex groups quite differently. Therefore, one model for the whole data may not be sufficient.

Acknowledgements

We wish to thank Dr. Wei-Yin Loh for his helpful comments on this paper. We also thank Dr. Robert Tsutakawa for kindly providing a full set of the Missouri lung cancer data.

References

Ahn, H. and Chen, J. J. (1997). Tree-structured logistic model for over-dispersed binomial data with application to modeling developmental effects. *Biometrics,* **53,** 435-455.

Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics,* **50,** 471-485.

Anscombe, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. *Journal of the Royal Statistical Society,* **B, 15,** 229-230.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics,* **33,** 38-44.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica,* **4,** 143-167.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica,* **5,** 641-666.

Chen, J. J. and Ahn, H. (1996). Fitting mixed Poisson regression models using quasi-likelihood methods. *Biometrical Journal,* **38,** 81-96.

Choi, Y. (2002). Tree-structured regression for a loglinear model with an extra-Poisson variation. Unpublished Ph.D. Thesis. Department of Applied Mathematics and Statistics, State University of New York at Stony Brook.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statistica Neerlandica,* **38,** 159-167.

Faddy, M. J. and Bosch, R. J. (2001). Likelihood-based modeling and analysis of data underdispersed relative to the Poisson distribution. *Biometrics,* **57,** 620-624.

Hill, A. B. (1977). *A Short Textbook of Medical Statistics,* (10th ed.) Philadelphia: Lippincott.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics,* **39,** 311-324.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics,* **15,** 209-225.

Levene, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics*, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. G. Mann (eds), 278-292. Stanford, California: Stanford University Press.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* **73,** 13-22.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction decision. *Statistica Sinica,* **12,** 361-386.

Loh, W.-Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association* **83,** 715-728.

Margolin, B. H., Kaplan, N. and Zeiger, E. (1981). Statistical analysis of the Ames salmonella /microsome test. *Proceedings of National Academy of Science,* **76,** 3779-3783.

Marienfeld, C. J., Collins, M., Wright, H., Reddy, R., Shoop, G., Roberts, K. K., and Rust, P. (1980). Cancer mortality and public drinking water in St. Louis city and county. *American Water Works Association Journal, 72,* 649-654.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Moore, D. F. and Tsiatis, A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics,* **47,** 383-401.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association,* **81,** 977-986.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical methods,* p45-52. Springer, New York.

Stevens, R. G., Moolgavkar, S. H. and Lee, J. A. H. (1982). Temporal trends in breast cancer. *American Journal of Epidemiology,* **115,** 759-777.

Tsutakawa, R. K. (1985). Estimation of cancer mortality rates: A Bayesian Analysis of Small Frequencies. *biometrics,* **41,** 69-79.

Tsutakawa, R. K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association,* **83,** 37-42.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika,* **61,** 439-447.