Pu et al. BMC Medical Research Methodology (2017) 17:20
DOI 10.1186/s12874-017-0303-y

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Impact of communities, health, and emotional-related factors on smoking use: comparison of joint modeling of mean and dispersion and Bayes' hierarchical models on add health survey

Jie Pu[1], Di Fang[2*] and Jeffrey R. Wilson[3]

Abstract

Background: The analysis of correlated binary data is commonly addressed through the use of conditional models with random effects included in the systematic component as opposed to generalized estimating equations (GEE) models that addressed the random component. Since the joint distribution of the observations is usually unknown, the conditional distribution is a natural approach. Our objective was to compare the fit of different binary models for correlated data in Tabaco use. We advocate that the joint modeling of the mean and dispersion may be at times just as adequate. We assessed the ability of these models to account for the intraclass correlation. In so doing, we concentrated on fitting logistic regression models to address smoking behaviors.

Methods: Frequentist and Bayes' hierarchical models were used to predict conditional probabilities, and the joint modeling (GLM and GAM) models were used to predict marginal probabilities. These models were fitted to National Longitudinal Study of Adolescent to Adult Health (Add Health) data for Tabaco use.

Results: We found that people were less likely to smoke if they had higher income, high school or higher education and religious. Individuals were more likely to smoke if they had abused drug or alcohol, spent more time on TV and video games, and been arrested. Moreover, individuals who drank alcohol early in life were more likely to be a regular smoker. Children who experienced mistreatment from their parents were more likely to use Tabaco regularly.

Conclusions: The joint modeling of the mean and dispersion models offered a flexible and meaningful method of addressing the intraclass correlation. They do not require one to identify random effects nor distinguish from one level of the hierarchy to the other. Moreover, once one can identify the significant random effects, one can obtain similar results to the random coefficient models. We found that the set of marginal models accounting for extravariation through the additional dispersion submodel produced similar results with regards to inferences and predictions. Moreover, both marginal and conditional models demonstrated similar predictive power.

Keywords: Overdispersion, Generalized additive models, Generalized linear models, Bayes' hierarchical model, Logistic regression

* Correspondence: difang@uark.edu
[2]Department of Agricultural Economics and Agribusiness, University of Arkansas, Fayetteville, USA
Full list of author information is available at the end of the article

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

The standard logistic regression model is commonly used in the analysis of uncorrelated binary response observations with several covariates [1]. It is a member of the generalized linear models (GLM) where the canonical link is the logit, and the random component is binomial. Its ability to appeal to the odds of occurrence of an event has gained great interest and popularity in fitting binary data. However, when correlated observations are modeled with a standard logistic regression, the results are not necessarily efficient or reliable. In such cases, we can use the generalized estimating equation (GEE) model or a generalized linear mixed model (GLMM), or as emphasized in this paper, resort to the joint modeling of the mean and the dispersion. This additional modeling of the dispersion (beyond the generalized linear model for the mean) allows us to correct for the inflated standard errors due to the intraclass correlation and thereby provides more reliable and efficient estimates. The additional submodel has other benefits, as there are cases when one may be only interested in modeling the dispersion. Smyth and Verbyla [2] among others proposed joint modeling of the mean and dispersion parameters for certain distributions. Taylor and Verbyla [3], Wu, Zhang and Xu [4] proposed a unified procedure for selecting significant covariates in the mean and the dispersion sub models.

In survey research, one often encounters correlated data due to the designed nature of the study, longitudinal data or clustered data. Such data usually referred to collectively as repeated measures, often emanates in healthcare, education, psychology, sociology, and other related areas. As an example, in an obesity study among adolescents, students were sampled from within schools, and schools were sampled from within school districts. In such a situation, the observations were correlated due to the hierarchical structure of the design. Thus, to fit standard logistic regression models to such data will require one to violate of the usual assumption of independence. As such one may declare covariates as significant when in fact, they are not. In fact, it is not appropriate to model correlated data with any generalized linear model, as such models assume that the dispersion parameter is fixed [5]. In other words, the generalized linear model addresses location as a centering of the data, but the spread is not assumed to vary. However, in most of these cases when the distribution of response distribution belongs to the exponential family the spread is often related to the mean. On the other hand, it is well known that in most cases in the analysis of correlated data, that fixed dispersion or dispersion tied directly to the mean is less common than one assumes when confronted with survey data. However, the joint modeling of the mean and the dispersion is an approach where covariates are used to model both the mean and the dispersion simultaneously. In fact, the estimators of the regression parameters in the mean submodel is more efficient when the dispersion is correctly addressed [6].

In this paper, we modeled smoking based on a set of covariates. We compared different approaches to modeling these correlated data. These methods included marginal models based on the joint modeling of the mean and the dispersion and conditional models based on the hierarchical structure of the data with the use of random effects. The basis for the joint modeling is to directly address the correlation in the data [6]. In this paper, the joint modeling of the mean and the dispersion includes the generalized linear model (GLM) and generalized additive model (GAM). These are marginal models.

We fitted, compared, and validated several logistic regression models as we modeled smoking, as defined by having at least one cigarette every day for 30 days. Smoking is detrimental as it can cause lung cancer, COPD (chronic obstructive pulmonary disease), heart disease, stomach cancer, as well as sudden infant death syndrome for women. Studies related to smoking have found that race and gender accounted for a differential role in the regular smoking habit of adults, while other studies have found an interactional effect between smoking, drinking, and income [7]. One study found that advertisements through social media such as movies and television have been linked with adolescent smoking initiation [8]. Yet another study found that patients with mental illness have a higher incidence of smoking than the general population [9].

Data

We obtained 4,484 observations from the fourth wave of Add Health survey [10], which was initially designed based on a two-level clustered design with an in-home questionnaire survey. Respondents were selected randomly from 132 different communities and were followed from 1995 to 2009. We selected variables based on social demographics, health status, and psychological status. A description of chosen predictors is presented in Table 1.

This sample consisted of respondents aged between 25 and 34 years, with 46.01% being male, and 32.4% being non-Caucasians. Approximately 45.2% of the respondents smoked cigarettes regularly (that is, at least one cigarette every day for 30 days), whereas 75% drank socially. 81% had both smoked and drunk at least once. Among those individuals who had both used alcohol or smoked, 20.9% had experienced smoking first, 23.1% had experienced alcohol first and 12.5% first experienced smoking and drink alcohol at the same time. On the health aspect, 37.5% of the sample had their last routine checks more than one year ago, 21.4% had no health

Table 1 Description of variables in our study

	Explanation
Demographic	Do you have degree above high school?
	What was the total household income before taxes and deductions?
	How many of these children are still living?
	Have you ever been Arrested?
	Gender
	Are you white?
Health	Do you have health insurance?
	How long ago did you last have a routine check-up?
	Do you have more than 5 drinks or drink alcohol more than 3 days a week during the past 12 months?
	Do you have prescription drug misuse?
	On the average, do you use a physical fitness or recreation center in your neighborhood per week?
	In the past seven days, how many hours did you watch television or videos, including VHS, DVDs or music videos?
	How old were you when you first had an Alcoholic drink?
Emotional	How important (if at all) is your religious faith to you?
	Before your 18th birthday, did a parent or other adult caregiver say things that really hurt your feelings or made you feel like you were not wanted or loved or hit you with a fist, kick you, or throw you down on the floor, into a wall, or down stairs more than 10 times?
	In the past 12 months, did you spend on volunteer or community service work?
	Are you optimistic about your future?
	Do you feel isolated from others?
	Do you satisfy with your current job?

insurance, and 30.6% had routine sports time. Regarding emotional satatus, 10.3% had psychological consulting experience, 13.5% had childhood mistreatment experience from their parents, 29.1% had negative or neutral attitude towards their future, and 28.8% felt isolated from the social life. Job satistfaction was average d at 27.7%.

Methods

We started with an unconditional hierarchical logistic regression model to obtain the intraclass correlation coefficient (ICC)[1]. The ICC measures the closeness of the individuals within a cluster as it pertains to smoking [11]. We found that ICC to be sufficiently large to warrant a correlated model [12]. We then conducted an analysis of smoking based on four (two marginal and two conditional) models appropriate for correlated binary data. We identified and addressed similarities and differences with these models as they addressed and accounted for the intraclass correlation. The two hierarchical models, one frequentist and one Bayes, and the two joint mean and dispersion models, one generalized linear model (GLM) and the other generalized additive model (GAM), were utilized. We compared the fit of these models as they pertained to their predictive ability on smoking. We stayed clear of addressing their mathematical or computational differences [13]. Our comparisons were made based on withholding a portion of the data for validation. We randomly selected 75% for the training dataset and 25% for validation dataset. We did this four different times, choosing a different 25% each time.

Hierarchical logistic regression models

We presented two conditional logistic regression or hierarchical logistic regression models for analysis. The dependency among observations was accounted for through adjustment to the systematic component with the addition of random intercepts and random slopes. This method of adjustment to the systematic component requires additional distributional assumptions as it now consists of random effects in addition to the fixed effects. This approach results in modeling the conditional mean rather than the marginal mean, thus, are referred to as subject-specific models [14]. Essentially, the subject-specific model consists of a product of a set of distributions, one set based on the conditional mean given the random effects and the other for the assumed distribution of the random effects. In our example, a subject -specific model tells us about the probability that the individual smoked given the community (random effect) [13].

Consider modeling smoking as the outcome as a Bernoulli random variable Y_{ij}, which takes on the value of one (if smoking) with probability P_{ij} and covariates X_{ijk} for the i^{th} individual in j^{th} community for the k^{th} covariate, i = 1,2,....I; j = 1,2,....J; k = 1,2,.....K. Let the overall differential cluster random effects be denoted as γ_{0j}(-random intercept in the model) and assumed to be distributed as a normal random variable with mean zero and variance δ_0^2. This addresses the unmeasured impact of the communities on the individual. The random intercept γ_{0j} represents the combined effect of all omitted covariates that causes individuals to be more prone to smoking. Our initial exploration of the data and some complimentary research suggested a differential effect across communities with regard to *arrest*. Thus, let the coefficient for *arrest* show a differential rate from community to community (random slope in the model). Random slope is denoted by γ_{1j} and assumed to be distributed normally with mean zero and variance δ_1^2. Finally, the hierarchical logistic

regression model with 19 covariates and random intercept and random slope is:

$$\text{logit}\left(p_{ij} \mid X_{ijk}\,\gamma_{0j},\,\gamma_{1j}\right) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_K X_{ijK} + \gamma_{0j} + \gamma_{1j} Z_{ij1} \quad (1)$$

where X_{ijk} represents the k^{th} covariate measure for the i^{th} person in the j^{th} cluster (community) β_i are the regression coefficients associated with covariate X_{ijk} β_0 is the overall fixed effects and Z_{ij1} is the covariate *arrest* in this case associated with the random slope coefficient of arrest, with conditional distribution $Y_{ij} \mid X_{ijk}\,\gamma_{0j}, \gamma_{1j} \sim Bin\,(n_i, p_{ij})$ where *Bin* denotes the binomial distribution, the random intercept for communities $\gamma_{0j} \sim N(0, \delta_0^2)$ and random slope $\gamma_{1j} \sim N(0, \delta_1^2)$ where N denotes the normal distribution and the correlation of the random effects intercept and slope parameters is $\sigma_{\gamma 0j,\gamma 1j}$. The correlation is due to the fact that observations in the same cluster share similar effects. The random slope γ_{1j} represents that there are differential rates of change with subjects in each cluster (community) as it relates to the covariate *arrest* Z_{ijk}. We referred to this as frequentist hierarchical logistic regression model as opposed to the Bayes' hierarchical logistic regression model with its additional set of prior distributions on the parameters. We fitted the frequentist hierarchical model with PROC GLIMMIX in SAS, which is presented in Additional file 1: Appendix.

Bayes' hierarchical logistic regression model

Bayesian hierarchical logistic regression model is presented as:

$$\text{logit}\left(p_{ij} \mid X_{ijk}\,\gamma_{0j},\,\gamma_{1j}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \gamma_{0j} + \gamma_{1j} Z_{ij1} \quad (2)$$

differs from the frequentist hierarchical model [1] merely in the set of prior distributional assumptions attached to the parameters in the Bayes' logistic regression model. The Bayes' logistic regression model requires that there are distributional assumptions specified for the unknown β_i parameters, as well as the covariance parameters δ_0^2 δ_1^2 and $\sigma_{\gamma 0j,\gamma 1j}$ associated with the distribution for the random effects γ_{0j}, $\gamma_{0j} \sim N(0, \delta_0^2)$ and γ_{1j} the random slope $\gamma_{1j} \sim N(0, \delta_1^2)$ respectfully. The β's are assumed to be normally distributed. The Bayes' model requires that we incorporate any prior information on the unknown parameters, $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_K$; δ_0^2 , δ_1^2 and $\sigma_{\gamma 0j,\gamma 1j}$ together with the information we obtained from the observed data. Thus we concentrate on the resulting posterior distribution from which we seek posterior modes. In fact, the prior knowledge is represented through the distributional assumptions. It allows updating the knowledge regarding the unknown parameter distribution as if its prior information is known. We used the prior information through the distribution $\Pr(\boldsymbol{\beta})$ which is based on initial beliefs to estimate and make inferences. It is customary to assume that the normal distribution is the most appropriate prior. Thus, we obtained the posterior distribution, as proportional ($\propto$) to the product of the likelihood function and the prior distribution on, which we concentrate. So that the probability is followed through the posterior,

$$\Pr\left[\boldsymbol{\beta}\;\gamma_{0j},\,\gamma_{1j},\,\delta_0^2,\delta_1^2 \mid Y\right] \propto \left[\Pr\left[Y \mid \boldsymbol{\beta},\,\gamma_{0j},\,\gamma_{1j},\,\delta_0^2,\delta_1^2\right] * \Pr\left(\boldsymbol{\beta},\,\gamma_{0j},\,\gamma_{1j},\,\delta_0^2,\delta_1^2\right)\right] \quad (3)$$

Since the data Y have distribution, Pr(Y) as constant relative to $\boldsymbol{\beta}$. If the prior distributions were chosen to be uniform instead of normal, then the estimates for equations [1] and [2] are equivalently both maximum likelihood estimates. We obtained estimates from the posterior distribution, [2.3]. Such posterior inference typically requires simulation techniques such as Gibbs sampling and the Metropolis Hasting sampling as a MCMC method to obtain estimates. In fact, it generates new values from a proposed distribution that determines how to select new parameter values based on the current values. The Bayesian procedure produces consistent and efficient estimates. It has also been shown that if we were to choose conjugate priors, it guarantees that a posterior distribution possesses the same property as the prior distribution [15]. We used MLWin to fit these Add Health data with Bayes' hierarchical logistic regression model. MLwiN is a powerful program designed for fitting multilevel models. It provides features such as *variance function window* to calculate the residual variance at any level.

Joint modeling of the mean and dispersion

The subject-specific models concentrated modeling the conditional mean through the use of random effects to address the correlation. However, there is added value to address the correlation through jointly modeling of the mean and the dispersion [7]. This joint modeling consists of two interlinked models, one for the mean and one for the dispersion [16]. This differs from the random coefficient models and its approach to addressing correlation. It is a reflection of the double exponential family that allows us to model the mean parameter while making use of a second parameter that addresses the variance but independently of the mean [17]. Each submodel is considered to be a generalized linear model with three components: random component, systematic component, and link component [18]. The joint modeling of the mean and dispersion consists of starting

initially with the mean submodel. Then, the residuals from the mean submodel or a function of them are used as the response (random component) in the dispersion submodel. The systematic component in the dispersion submodel is allowed if needed, to have the same or a subset of covariates as in the mean submodel or a totally different set of covariates not yet used. The link function in the dispersion submodel follows similar procedure though not necessarily the same function as the mean submodel. We used the same model checking techniques for both the mean and the dispersion submodel [19]. We considered two different joint modeling of both the mean and the dispersion submodel (Double Joint Modeling Mean and Dispersion), one based on GLM and the other based on GAM.

Double GLM Joint Modeling mean and dispersion

The joint modeling of the mean and the dispersion, also referred to as double generalized linear models (DGLM) consists of three parts: a function for the variance, a GLM submodel for the mean, and a GLM submodel for the dispersion [2]. Thus, we need to obtain an estimation of the mean function and estimation of the variance function [20].

In particular, let us assume that the binary random variable Y_i is Bernoulli with probability p_i, which depends on a set of covariates $\mathbf{X} = (X_1, X_2, \ldots X_K)$. The mean submodel is first fitted with logit link. The deviances from the mean submodel d_i have mean ϕ_i with variance $V_{di}(\phi_i)$ representing the random component. The systematic component of the dispersion submodel consists of the vector of covariates $\mathbf{Z} = (Z_1, Z_2, \ldots Z_t)$ with link component as log. In our model we have one covariate in the dispersion submodel as *arrest*. In summary, we have two interlinked generalized linear models consisting of the mean submodel, measuring smoking use $Yi \sim Ber(p_i)$, where p_i represents probability of smoking with logit link,

$$\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K \quad (4)$$

and the dispersion submodel based on d_i such that $d_i \sim D_d(\phi_i, V_{di}(\phi_i))$ and log link

$$n_{di} = \log(\phi_i) = \gamma_0 + \gamma_1 Z_1.$$

We used PROC QLIM and Macro HPGLIM in SAS to fit these models, as presented in Additional file 1: Appendix.

Double GAM Joint Modeling mean and dispersion

In the joint modeling of the mean and dispersion one can replace those GLM with GAM. As such we still present two submodels, one for mean and one for the dispersion, but in each we have a generalized additive model instead of GLM. A general additive model is similar to generalized linear model as it relates the mean of the random response variable Y and a set of covariates $X_i, \ldots, X_p$, [21]. The generalized additive model (GAM) distinguishes itself from the generalized linear model in that

$$E[Y] = \gamma_0 + \gamma_1(X_1) + \ldots + \gamma_K(X_K) \quad (5)$$

where $\gamma_i(X_i)$, $i = 1, \ldots, K$; are smooth functions. These smooth functions γ_i, are not given a parametric form but instead are estimated in a nonparametric or semiparametric form. The GAM consists of a random component with response Y_i, a systematic component that is additive in the function of covariates, and a link function relating the response in terms of the mean with a combination of the covariates. Whereas GLM has as its systematic component the linear predictor of the form $\Sigma\beta_i X_i$ the GAM instead, uses the additive component, as a sum of smooth functions $\Sigma\gamma_i(X_i)$. The function γ_i is determined or estimated based on a nonparametric technique usually influenced by the examination of actual plots. The GAM requires one to determine the smoothing parameter through the generalized cross validation (GCV) function in nonparametric regression methods [22]. The functions $\gamma(\cdot)$ and $f(\cdot)$ are additive smooth functions, unspecified linear functions or nonparametric methods, for both mean and dispersion submodels.

The GAM is known to be more flexible than the parametric methods as an exploratory analysis in identifying the relationship between the response and the covariates. Also it is more flexible as opposed to other models when detecting quadratic or higher order power in a piecemeal fashion [21]. In addition, the GAM can also be used more efficiently to uncover nonlinear effects among the covariates. While the generalized linear model is used for estimation and inferences for the regression parameters, the GAM is seen as an exploratory method based on a nonparametric approach while making use of an apparent response-covariate relationship. In summary, we have the systematic component for the mean additive submodel logit $(p_i) = \sum_{i=1}^{p}\gamma_i(X_i)$ and the systematic component for dispersion additive submodel $\log(\text{deviance}_i) = \sum_{i=1}^{k} f_i(X_i)$. The log (deviance) is assumed to be distributed as a normal distribution in the dispersion with identity link function [23]. We used PROC GAM in SAS to fit these models, as presented in Additional file 1: Appendix.

Results

We fitted two logistic regression models with random effects, frequentist hierarchical logistic regression model

and Bayes' hierarchical logistic regression model. We also fitted two joint modeling of the mean and dispersion models, one based on the generalized linear model, and the other on generalized additive modeling. Results of all models are presented in Table 2.

In general, we concluded that individuals who had higher income level were less likely to have smoked. People who had kids were about 1.23 times more likely to have smoked regularly. If one had alcohol or drug misuse, they were about 3.26 times more likely to smoke. Individuals who had been arrested or had spent more time watching TV or playing video games were about 1.96 times more likely to smoke than those with no such habits. Non-Caucasian were about 1.58 more likely to smoke than Caucasian. People not practicing religion were about 1.34 times more likely to smoke than those who were religious. People with lower education level were about more than 1.32 times likely to smoke. Individuals who drank alcohol early in life were more than 1.52 times more likely to be a regular smoker. People who experienced mistreatment from their parents during childhood were more than 1.70 times more likely to smoke regularly. Job satisfaction had no significant impact on smoking use. Finally, people who felt isolated from social activities had no impact.

Results of standard logistic regression

As a point of reference regarding the extra variation or overdispersion present in the data due to its hierarchical nature, we fitted a standard logistic regression with 19 covariates. The goodness- of- fit test for overdispersion (Hosmer-Lemeshow, $X^2 = 19.97$, p-value = 0.011) suggested that the model is not a good fit and overdispersion is clearly present. This may be in part due to ignoring the hierarchical structure of the design. Initially, we fitted an unconditional logistic regression model with random effects (community) and no covariates and obtained an ICC value of 0.146. Thus, the intraclass correlation is large enough to suggest using a correlated model to address the hierarchical structure [12].

Table 2 Parameter estimates and standard errors in hierarchical logistic regression

Estimates	Hierarchical logistic		Bayes' Hierarchical logistic		Joint GLM		Joint GAM	
	Estimates	*p*-value	Estimates	*p*-value	Estimates	*p*-value	Estimates	*p*-value
Intercept	0.451	0.057	0.25	0.055	0.3853	0.0003***	0.597	0.009**
Arrested	0.677	<.001***	0.709	<.001***	0.677	<.001***	0.606	<.001***
Race	0.461	<.001***	0.584	<.001***	0.63	<.001***	0.523	<.001***
Drug	1.221	<.001***	1.183	<.001***	1.235	<.001***	1.066	<.0001***
TV time	0.007	0.027*	0.007	0.005**	0.008	<.001***	0.008	0.008**
Mistreatment	0.309	0.011**	0.293	0.003**	0.33	<.001***	0.214	0.07
Religion	-0.299	0.007***	-0.293	<.001***	-0.31	<.001***	-0.281	0.001***
Alcohol	0.541	<.001***	0.544	<.001***	0.577	<.001***	0.420	<.001***
Kid	0.208	0.017***	0.208	0.002**	0.245	<.001***	0.231	0.006**
Public work	-0.27	0.002**	-0.28	<.001***	-0.302	<.001***	-0.268	0.002**
Starting age	-0.006	<.001***	-0.006	<.001***	-0.007	<.001***	-0.007	<.001***
Education	-0.473	<.001***	-0.509	<.001***	-0.497	<.001***	-0.484	<.001***
Income	-0.074	<.001***	-0.234	<.001***	-0.073	<.001***	-0.087	<.001***
Gender	-0.144	0.1149	-0.17	0.031	-0.18	<.001***	-0.166	0.063
Sports	-0.486	<.001***	-0.493	<.001***	-0.436	<.001***	-0.47	<.001***
Insurance	-0.15	0.177	-0.189	0.029*	-0.158	0.002**	-0.144	0.186
Routine check	-0.111	0.207	-0.103	0.113	-0.11	0.006**	-0.078	0.369
Attitude future	-0.152	0.095	-0.16	0.046*	-0.152	0.001***	-0.143	0.112
Social relation	0.065	0.482	0.102	0.062	0.096	0.024**	0.058	0.526
Job	0.102	0.274	0.082	0.098	0.078	0.070	0.12	0.195
Covariance Parameters								
var[Intercept]	0.099	0.023	0.009	<.001***	N/A		N/A	
var[Arrested]	0.116	0.025	0.118	<.001***	N/A		N/A	
cov[int & Slope]	0.028	0.027	0.03	<.001***	N/A		N/A	

*p<0.05, **p<0.05, ***p<0.05

Results of frequentist hierarchical and bayes' hierarchical logistic regression model

A frequentist hierarchical logistic regression model with random intercept and random slope based on the differential rate due to the covariate *arrested*, resulted in the estimate of the variance for the intercept as 0.099 and random slope of 0.116 with its standard errors of 0.062 and 0.071 respectively. The covariance estimate between intercept and slope is 0.028. Thus, resulting in a standardized value of 0.099/0.062 for p-value of 0.023 for intercept and 0.116/0.071 for p-value of 0.025 and suggesting that the rate of change as it pertains to arrest varied across communities. Covariates, alcohol, kid, arrested, race, drug, TV time, mistreatment and religion showed significant positive effects on the probability of smoking, while covariates religion, public work, starting age, education, income, gender, sports showed significant negative effects on the probability of smoking. Thus, one expected that the probability that an individual who smoked was impacted by factors, although unidentifiable, but were related to the communities and history with the law as it depended on community. Mistreatment, gender, attitude towards the future, social relation and whether or not they had a job had no impact on the smoking behavior.

We fitted a Bayes' hierarchical logistic regression model with random effects to measure in community random slope in arrested. A Bayes' hierarchical logistic regression model fitted with MLWin gave random effects in community and arrested as significant with $\widehat{\sigma^2_{intercept}} = 0.009$ and $\widehat{\sigma^2_{arrested}} = 0.118$ and $\widehat{\sigma_{int/arr}} = 0.030$ with standard errors 0.002, 0.035, and 0.009, respectively. This suggested that our random effects due to community and due to arrest across communities differed significantly.

Comparison of hierarchical models

The hierarchical models which tell about the conditional means (probabilities) gave similar results though for *social relation*, *routine check*, *insurance*, and *attitude towards the future* they differed. In particular, the models agreed with the set of covariates we categorized under demographics or health. When difference appeared they occurred mainly with Bayes' and double GLM joint modeling model. In those cases, the Bayes' hierarchical showed some marginal significance while the frequentist did not identify any trace of significance with these covariates. Bayes' results have a tendency to have smaller *p*-values. This may be due to the fact that the Bayes' method uses a prior distribution on the parameters.

Results of joint modeling of mean and dispersion

A joint modeling of the mean and the dispersion with a generalized linear model for each submodel with covariate *arrest* in the dispersion sub-model provided a good fit to the data. The mean sub-model consisted of covariates, alcohol, kid, arrested, race, drug, TV time, mistreatment, social relation and job showing significant positive effects on the probability of smoking. The covariates religion, public work, starting age, education, income, insurance, sports, routine check, attitude future showed significant negative effects on the probability of smoking, Table 2.

The joint modeling of the mean and the dispersion with generalized additive submodels was fitted to the data. The smoothing effects were significant ($p = 0.0001$) with smallest GCV value of 3.96, based on estimating the additive predictors by using a B-spline smoother with 3 degrees of freedom including a strong quadratic pattern.

Comparative dispersion submodel

In both the mean and the dispersion submodel for both the GLM and GAM, covariate arrest had a significant impact ($p < 0.001$) on the dispersion (see Table 3).

In the next section, we examined the predictive probability for each of these models based on the validation set and the training set in four different datasets. We used this in our comparisons as we took a closer look at the predictive ability of the models.

Accuracy comparison of four models

We used the area under the Receiver Operating Characteristic (ROC) curves for both the training dataset and the validation dataset as a measure of fit for our four logistic regression models. We compared the four models based on their predicted probabilities in the training dataset and the validation dataset (see Table 4). We repeated the process four times, with a different 25% of the data omitted for validation each time. The joint modeling of the mean and the dispersion with GAMs had the best predicted probabilities among validation dataset. However, the four models did not show marked differences in their predicted probabilities.

Discussion

We fitted four binary logistic regression models, frequentist hierarchical, Bayes' hierarchical model, GLM joint modeling of mean and dispersion, and GAM joint modeling mean and dispersion to model smoking. Two of these models (the joint mean and dispersion submodel) addressed the marginal probabilities while the other

Table 3 Estimates of coefficients in dispersion submodel

	Joint GLMs		Joint GAMs	
Parameter	Estimate	*p*-value	Estimate	*p*-value
Intercept	-2.016	<.001	-0.519	<.001
Arrested	0.551	<.001	1.051	<.001

Table 4 Accuracy Comparison of four models based on 4-fold

Method	ROC							
	Testing dataset	Validation dataset	Testing dataset	Validation dataset	Testing dataset	Validation dataset	Testing dataset	Validation dataset
Frequentist Hierarchical	0.7957	0.7375	0.7757	0.7608	0.7875	0.74	0.7795	0.7658
Bayes' Hierarchical	0.7527	0.7279	0.7501	0.7377	0.7447	0.7228	0.7587	0.7500
Joint GLMs	0.7613	0.7378	0.7528	0.7613	0.7596	0.7411	0.75	0.7691
Joint GAMs	0.7769	0.7516	0.7671	0.7765	0.773	0.7594	0.7679	0.7769

two hierarchical models (Frequentist and Bayes') addressed the conditional probabilities. The conditional probability models found *arrest* as a covariate with differential rates across communities while the marginal model also found that *arrest* is a key variable in explaining the dispersion. Overall, the models did not perform markedly different and gave similar results based on the training dataset and the validation dataset. The marginal models were similar with their predictability in the validation datasets and the training datasets.

We acknowledge the limitation of using a single dataset and encourage future research to conduct similar studies on a variety of correlated data. In the meantime, we are confident about the applicability of our conclusion to other research domains. Even though we did not perform a simulation study, we found similar patterns with hierarchical models with the UCLA dataset and recent research [12]. Similar results were also seen in the fit of the conditional models. The consistency of the models and the fit with validation and training datasets led us to believe in the generalizability of this paper to broader receivers.

Conclusion

We fitted both conditional and marginal models to study smoking behavior for adolescents who had eventually become adults. While these models are addressing two different questions, we did not detect significant differences in both model performances based on training or validation. Further investigation showed that most communities were alike with a few showing that different random effect. By wave 4 these adolescents were moved away from their original communities so over time the community impact was negligent or independent of the smoking behavior.

Endnote

[1]We define the ICC as the ratio of the variance component due to clusters to the total variance for individuals, such that $\mathrm{ICC} = \frac{\delta^2_{\mathrm{cluster}}}{\delta_{\mathrm{individual}}{}^2 + \delta^2_{\mathrm{cluster}}}$.

Additional file

Additional file 1: Appendix: SAS Code Used to Perform Models. (DOCX 150 kb)

Abbreviations

Add Health: The National Longitudinal Study of Adolescent to Adult Health (Add Health); COPD: Chronic obstructive pulmonary disease; DGLM: Double generalized linear models; GAM: Generalized additive model; GCV: Generalized cross validation; GEE: Generalized estimating equation; GLMM: Generalized linear mixed model; ICC: Intraclass correlation coefficient; MADAM: Mean and dispersion additive model; ROC: Receiver operating characteristic

Acknowledgements

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

Funding

There is no funding sources for this work.

Availability of data and materials

Public-use data are available from three different sources: The Odum Institute at UNC, the Inter-University Consortium for Political and Social Research (ICPSR) and Sociometrics. The authors downloaded the datasets from: https://dataverse.unc.edu/dataverse/addhealth.
No permission or contract was required to access the public use dataset. The Public-Use dataset contains all the data from the In-home Interview, just a smaller sampling. A smaller sample was released through the public-use to limit deductive disclosure risk. Public-Use data doesn't contain ID numbers of friends, siblings or romantic partners, so the data cannot be linked.

Authors' contributions

JRW conceived the statistical theory of this study. JP carried out all computations and JP and DF drafted the manuscript. JRW and DF critically reviewed and made substantial contributions to the manuscript. All authors commented on and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Add Health participants provided written informed consent for participation in all aspects of Add Health in accordance with the University of North Carolina School of Public Health Institutional Review Board guidelines that are based on the Code of Federal Regulations on the Protection of Human

Subjects 45CFR46: http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html.
The Add Health data files do not contain respondent identifiers or any links to identifiers. The data do contain ID numbers which are necessary to allow researchers to link data across the waves and to friends and partners (also de-identified).

Author details
[1]School of Mathematical and Statistical Science, Arizona State University, Tempe, USA. [2]Department of Agricultural Economics and Agribusiness, University of Arkansas, Fayetteville, USA. [3]Department of Economics, Arizona State University, Tempe, USA.

Received: 29 October 2016 Accepted: 28 January 2017
Published online: 03 February 2017

References

1. Wilson JR, Lorenz KA. Modeling binary correlated responses using SAS, SPSS and R. New York: Springer; 2015.
2. Smyth GK, Verbyla AP. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics. 1999;10(6):695–709.
3. Taylor J, Verbyla A. Joint modelling of location and scale parameters of the t distribution. Stat Model. 2004;4(2):91–112.
4. Wu L, Zhang Z, Xu D. Variable Selection for Joint Mean and Dispersion Models of the Lognormal Distribution. Hacet J Math Stat. 2012;41:2.
5. Nelder, JA, Baker RJ. Generalized linear models. In Encyclopedia of Statistical Sciences. New York: Wiley; 2004.
6. Smyth GK. Generalized linear models with varying dispersion. J Royal Stat Soc Series B (Methodological) 1989;51(1)47–60.
7. Office of the Surgeon General (US), Office on Smoking and Health (US): 2004.
8. Auld MC. Smoking, drinking, and income. J Hum Resour. 2005;40(2):505–18.
9. Dalton MA, Sargent JD, Beach ML, Titus-Ernstoff L, Gibson JJ, Ahrens MB, Tickle JJ, Heatherton TF. Effect of viewing smoking in movies on adolescent smoking initiation: a cohort study. Lancet. 2003;362(9380):281–5.
10. Harris KM. The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009 [machine-readable data file and documentation]. Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill; 2009. doi:10.3886/ICPSR27021.v9.
11. Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study. Outcome and mediating variables, by sex and ethnicity. Am J Epidemiol. 1996;144(4):425–33.
12. Irimata KE, Wilson JR: Identifying Intraclass Correlations Necessitating Hierarchical Modeling. Journal of Applied Statistics, *accepted*
13. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiol. 1998;147(7):694–703.
14. Hu F. Bias from Exposure Suspicion in Case–Control Studies. Wiley StatsRef: Statistics Reference Online.
15. Zellner A. New information-based econometric methods in agricultural economics: Discussion. Am J Agric Econ. 1999;81(3):742–6.
16. Efron B. Double exponential families and their use in generalized linear regression. J Am Stat Assoc. 1986;81(395):709–21.
17. Nelder JA, Lee Y. Joint modeling of mean and dispersion. Technometrics. 1998;40(2):168–71.
18. Dobson AJ, Barnett A. An introduction to generalized linear models. London: CRC press; 2008.
19. McCullagh P, Nelder JA. Generalized linear models. London: CRC press; 1989.
20. Carroll RJ, Ruppert D. Transformation and weighting in regression. London: CRC Press; 1988.
21. Hastie T, Tibshirani R. Generalized additive models. Statistical science 1986: 297–310.
22. Xiang D. Fitting generalized additive models with the GAM procedure. SAS Institute, Paper P256–26 2001.
23. Rigby R, Stasinopoulos D. A semi-parametric additive model for variance heterogeneity. Stat Comput. 1996;6(1):57–65.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central