دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد آمار زیستی

عنوان :

کاربرد مدل های رگرسیونی شمارشی در مدل بندی متغیر های پاسخ متورم در صفر

توسط : مریم جلالی

استاد راهنما : دکتر محمدرضا بانشی

استاد مشاور : دکتر علی‌اکبر حقدوست

سال تحصیلی : ۱۳۹۲-۱۳۹۱

Application of Count Regression Models for Modelling of Zero Inflated Outcomes

A Thesis

Presented to

The Graduate Studies

By

Maryam Jalali

In Partial Fulfillment

Of the requirements for the Degree

Master of Science in:

Biostatistics

Kerman University of Medical Sciences

may 2013

چکیده

مقدمه و هدف:

گروه هایی از افراد مثل افراد مصرف کننده الکل (در کشور هایی که مصرف الکل مجاز نمی باشد) گروه های پنهان نامیده می شوند. دسترسی و اطلاع از افراد این گونه گروه ها مشکل می باشد. از سویی دیگر اطلاع از تعداد افراد در این گروه ها و اینکه چه کسانی (ویژگی های جمعیتی) بیشتر افراد متعلق به این دسته ها را می شناسند برای تصمیم گیری ها و سیاستگذاری های سلامت مهم می باشند. برای تعیین اینکه کدام ویژگی های جمعیتی بر شناخت افراد متعلق به این گروه ها تاثیرگذار است باید از مدل های آماری مناسب این گونه داده ها استفاده کرد. داده هایی مثل تعداد افراد در گروه های پنهان، داده های شمارشی نامیده می شوند. داده های شمارشی به تعداد رویداد های مشخص در یک بازه زمانی اشاره دارد که شامل اعداد صحیح غیر منفی می باشد. بعضی از داده های شمارشی مثل داده هایی که از افراد سوالات حساس پرسیده می شود (مثلا تعداد الکلی هایی که افراد می شناسند) شامل تعداد زیادی صفر هستند. با در نظر گرفتن داده های شمارشی به عنوان متغیر پیوسته می توان مدل رگرسیون خطی را به داده ها برازش داد. ولی به دلیل اینکه توزیع داده های شمارشی با درصد زیاد صفر چوله هستند، نرمال بودن که یکی از فرضیات این مدل است برقرار نمی باشد، در نتیجه استفاده از این مدل برای این نوع داده مناسب نیست.

همچنین برازش مدل رگرسیون لجستیک به این نوع داده ها به دلیل اینکه درصد زیادی از جزییات در نظرگرفته نمی شود مناسب نمی باشد.

مدل رگرسیون پوآسن ، مدل استاندارد برای آنالیز داده های شمارشی می باشد، ولی این مدل به داده های شمارشی متورم در صفری که بیش پراکنش در آن ها موجود باشد، برازش ضعیفی دارد. بنابراین سوالی که باید مد نظر قرار گیرد این است که کدام یک از مدل های رگرسیونی شمارشی، مناسب داده های شمارشی متورم در صفرمی باشد و همچنین مساله دیگر این است که آیا ویژگی های جمعیتی مثل سن، تحصیلات و غیره بر پاسخ افراد از شناخت افراد متعلق به گروه های پنهان تاثیرگذار است یا خیر.

مواد و روش ها: در این مطالعه ۷۰۰ شرکت کننده شامل ۳۷۲ نفر(۵۳/۱ درصد) از استان فارس و ۳۲۸ نفر(۴۶/۹ درصد) از استان کرمان انتخاب شده و از آنها تعداد افراد متعلق به گروه های پنهان که می شناسند پرسیده شده است. این گروه ها شامل افراد مصرف کننده الکل، افراد مصرف کننده متادون و زنان تن فروش بوده اند. سه متغیر پاسخ در این مطالعه در نظر گرفته شده

است که عبارتند از: تعداد افراد الکلی، تعداد مصرف کننده های متادون و تعداد زنان تن فروشی که افراد شرکت کننده در مصاحبه می شناسند. متغیر های مستقل در این مطالعه عبارتند از: استان، جنسیت، سن، تحصیلات و وضعیت تاهل. پنج مدل رگرسیونی به این داده ها برازش داده شده اند که عبارتند از: لجستیک، پوآسن ،دوجمله ای منفی ، پوآسن متورم در صفر و دوجمله ای منفی متورم در صفر این مدل ها بر اساس آزمون نسبت درستنمایی، آزمون وانگ انجام شده و بر اساس معیارهای AIC و مجموع توان دوم خطا مقایسه شدند.

یافته ها: درصد صفر درشناخت افراد مصرف کننده الکل، مصرف کننده متادون و زنان تن فروشی که افراد شرکت کننده می شناسند به ترتیب عبارت است از: ٪۳۳/۹، ٪۴۹/۱ و٪ ۶۵/۳. در متغیر اول مدل های پوآسن و دوجمله ای منفی با استفاده از آزمون آزمون نسبت درستنمایی مقایسه شدند. P- مقدار کمتر از ۰/۰۰۱ نشان دهنده وجود بیش پراکنش در داده ها و در نتیجه برازش بهتر مدل دوجمله ای منفی به مدل به پوآسن می باشد. علاوه بر این به منظور مقایسه مدل های دوجمله ای منفی و دوجمله ای منفی متورم در صفر آزمون وانگ استفاده شده است که P- مقدار این آزمون ۰/۰۱ می باشد، در نتیجه مدل دوجمله ای منفی متورم در صفر به مدل دوجمله ای منفی ترجیح داده می شود. خطای استاندارد ضرایب متغیرها در مدل پوآسن و پوآسن متورم در صفر به طور قابل توجهی از مدل دوجمله ای منفی و دوجمله ای منفی متورم در صفر کمتر می باشد که دلیل آن در نظر نگرفتن بیش پراکنش در این مدل ها می باشد. متغیر های تاثیرگذار بر شناختن یا آشکار کردن تعداد افراد الکلی که افراد پاسخ دهنده می شناسند عبارت است از: استان، جنسیت، گروه سنی و سطح تحصیلات. بر اساس مدل دوجمله ای منفی متورم در صفرکرمانی ها ۲۵درصد کمتر از ساکنین فارس افراد مصرف کننده الکل می شناختند. ($IRR=$ ۰/۷۵ و $P =$ ۰/۰۲- مقدار). همچنین زنان ۷۰ درصد کمتر از مردان شناخت افراد الکلی را گزارش کرده اند. علاوه بر این افراد در گروه سنی کمتر از ۳۰ و کسانی که دارای تحصیلات پایین بوده اند تعداد افراد بیشتری از گروه الکلی ها می شناسند. افزایش سطح تحصیلات و سن با کاهش شناخت افراد الکلی ها مرتبط بوده است.

مدل دوجمله ای منفی بهترین برازش را به دو متغیر پاسخ دیگر داشته است. در متغیر دوم براساس آزمون آزمون نسبت درستنمایی مدل دوجمله ای منفی برازش بهتری (P- مقدار کمتر از ۰/۰۰۱) نسبت به مدل پوآسن به داده های این متغیر داشته است. به این معنی که بیش پراکنش در داده ها وجود داشته است. اما استفاده از آزمون وانگ برتری مدل دوجمله ای منفی متورم در صفر به دوجمله ای منفی را تایید نکرد ($P =$ ۰/۴۸- مقدار).

بر اساس نتایج مدل دوجمله ای منفی کرمانی ها ۲/۰۶ برابر افراد ساکن استان فارس بیشتر شناختن افراد مصرف کننده متادون را گزارش کرده اند. همچنین تحصیلات با شناخت افراد مصرف کننده متادون رابطه داشته به طوری که هر چه تحصیلات افراد کمتر بوده تعداد افرادی که از این دسته می شناختند بیشتر بوده است. به ویژه افراد دارای تحصیلات بالاتر از لیسانس نسبت به افراد زیر دیپلم ۸۰ درصد کمتر، شناخت افراد مصرف کننده متادون را گزارش کرده اند.

در مورد زنان تن فروش بر اساس آزمون آزمون نسبت درستنمایی مدل دوجمله ای منفی بر مدل پوآسن برتری داشته است (P- مقدار آزمون کمتر از ۰/۰۰۱). ولی آزمون وانگ برتری مدل دوجمله ای منفی متورم در صفر را به دوجمله ای منفی تایید نکرد (P- مقدار = ۰/۱۷). همچنین خطای استاندارد ضرایب در مدل های پوآسن و پوآسن متورم در صفر کمتر از خطای استاندارد این ضرایب در مدل دوجمله ای منفی بوده است در حالی که این خطای استاندارد ها در مدل دوجمله ای منفی متورم در صفر و دوجمله ای منفی به دلیل مشابه بودن برازش آن ها بر اساس آزمون های قراردادی تقریبا یکسان می باشد.

بر اساس مدل دوجمله ای منفی کرمانی ها ۰/۴۲کمتر از افراد ساکن استان فارس شناخت زنان تن فروش را گزارش کرده اند، همچنین زنان ۸۰درصد کمتر از مردان شناخت زنان تن فروش را گزارش کرده اند. بر اساس نتایج در همه مدل ها افراد در گروه سنی ۳۹-۳۰ و بالای ۴۰ سال و همچنین افراد با تحصیلات بالاتر شناخت کمتر زنان تن فروش را گزارش کرده اند. نتایج برازش مدل لجستیک به هیچ کدام از متغیرهای پاسخ رضایت بخش نبوده است.

نتیجه گیری: با وجود اینکه مدل رگرسیون پوآسن اولین انتخاب برای مدل بندی داده های شمارشی می باشد، این مدل ممکن است در برازش داده های شمارشی کافی نباشد و به مدل های شمارشی دیگر برازش بهتری به داده ها داشته باشند. هنگام وجود بیش پراکنش و یا صفرهای زیادی، استفاده از این مدل رگرسیونی منجر به کوچک شدن خطای استانداردها می شود. در نتیجه متغیر هایی که اثری روی متغیر پاسخ ندارند ممکن است معنی دار شوند. همچنین ویژگی های جمعیتی به عنوان عوامل تاثیرگذار بر شناخت یا فاش کردن افراد متعلق به گروه های پنهان شناخته شد.

واژگان کلیدی: مدل های رگرسیونی شمارشی، متورم در صفر، بیش پراکنش، دوجمله ای منفی، لجستیک، پوآسن

Abstract

***Background**: A group such as 'number of alcoholics' (in counties where alcohol is totally banned), are known as 'hidden'. Accessing to the members of these groups is difficult. On the other way determining the count and demographic characteristics of respondents which influence knowing people of these hidden groups are difficult. However, information about their size is important for health planning.*

To determine the effective demographic characteristics of respondents on their responses to the question "number of people they knew from hidden groups" suitable statistical models should be used.

The data like number of people in hidden groups are called count data. A count refers to the number of specified events that occur in a given interval of time. By definition, count data consist of only nonnegative integers.

In the case of sensitive questions such as number of alcoholics known, majority of respondents might give an answer of zero. Considering the data as a continuous variable, one can fit the linear regression to the data. But due to the Skewness of the data with high preponderance of zero the normality assumption which is one of the hypotheses of this model dose not met. So this model is not suitable for these data.

Also fitting logistic regression is not suitable because this model doesn't consider the details of the data.

Poisson regression model (P) is the standard tool to analyze count data. However, P provides poor fit in the case of zero inflated counts, when over-dispersion exists. Therefore, the questions to be addressed are to compare performance of alternative count regression models; and to investigate whether characteristics of respondents affect their responses.

***Methods**: A total of 700 participants were asked about number of people they know in hidden groups; alcoholics, methadone users, and Female Sex Workers (FSW). Five regression models were fitted to these outcomes: Logistic, P, Negative Binomial (NB), Zero Inflated Poisson (ZIP), and Zero Inflated Negative Binomial (ZINB). Models were compared in terms of Likelihood Ratio Test (LRT), vuong, AIC and Sum Square of Error (sse).*

***Results**: Percentages of zero were 35% for number of alcoholics, 50% for methadone users, and 65% for FSWs. ZINB provided the best fit for alcoholics, and NB provided the best fit*

for other outcomes. For the alcohol, LRT was used to compare NB and P. NB was superior to P, indicating existence of over-dispersion (P<0.0001). Applying vuong test, the ZINB model reflected the observed data better than NB (P=0.01).

In addition, we have seen that SEs of all variables in P and ZIP models were considerably smaller than ZINB. This is mainly due to the fact that P and ZIP models do not consider the over-dispersion parameter.

Variables affected respondents to reveal the number of known alcoholics were province, gender, age groups and education levels.. Based on ZINB model, Kermaninans' were about 25% less likely to know alcoholics than those from Fars (OR=0.74, P-value=0.02). Furthermore, women were 70% less likely to know alcoholics than male. Respondents in less than 30 years old and those with low education were more likely to know or reveal alcoholics. Increasing in age and education level was associated with reduction of revealing information about the number of alcoholics, people knew.

The NB had the best fit for the other two outcomes. For the second outcome the NB explained the observed data better than P (P-value<0.001) indicating over-dispersion. However, the ZINB model was not preferred over the NB according to the vuong test (P=0.48). Based on NB model, Kermaninans, relative to respondents in Fars, was 2.06 times more likely to know or reveal information about the number of methadone users they knew. Education level was associated with the outcome as well, where less educated people knew more methadone users than those in the other categories. In particular those in more than 16 years education group, relative to those in lower than 12 years education category, were about 80% less likely to reveal or know methadone users.

Regarding the FSW, the LRT of over-dispersion comparing the NB to the P yielded a P-value <0.0001. However, ZINB model did not preferred over the NB (P=0.17). The SEs for the P and ZIP were smaller than that of NB. However, results of NB and ZINB were nearly equal because they provided the same fit to the data according the formal tests. Based on the NB model, people in Kerman were about 42% less likely to know FSWs than those in Fars.

Females were about 80% less likely to know FSW than male (in NB model). The result of logistic regression was satisfying in any models.

***Conclusion**: Although P is the first choice for modeling of count data in many cases, it seems because of over-dispersion of zero inflated counts data in the case of sensitive questions,*

other models, specifying NB and ZINB, might have better goodness of fit and using this model in these condition may lead to small sses and significancy of variables that do not affect on the outcome variable. Also demographic characteristics were effective factors on knowing or revealing people of hidden groups.

Keywords: *count regression, zero inflated, over-dispersion, Negative Binomial, ZINB*