دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

دانشگاه علوم پزشکی کرمان
جمعیت دانشجویی
کتابخانه مرکزی

پایان نامه مقطع کارشناسی ارشد رشته آمار زیستی

عنوان:

تاثیر نرخ داده گمشده و روش های برآورد داده های گمشده در برآورد اندازه ی گروه های در معرض خطر

توسط: نسیم دهدشتی

استاد راهنما: دکتر محمدرضا بانشی

استاد مشاور: دکتر علی اکبر حقدوست

سال تحصیلی: ۱۳۹۳- ۱۳۹۲

Impact of Missing Rate and Method of Imputation of Missing Data on the Size Estimation of Hidden Groups

A Thesis

Presented to

The Graduate Studies

By

Nasim Dehdashti

In Partial Fulfillment

Of the Requirements for the Degree

Masters

Biostatistics

Kerman University of Medical Sciences

Spring 2014

۹۲

چکیده

هنگامی که افراد تحت تأثیر مواد مخدر یا الکل قرار می گیرند، احتمال بیشتری دارد که مرتکب رفتارهای پرخطر نظیر تماس جنسی بدون رعایت بهداشت، استفاده از وسایل تزیق مشترک و... شوند. بنابراین می توان گفت اعتیاد به دلایل مختلف خطرناک است ولی ابتلا به ایدز یکی از مهمترین عواقب آن است. اغلب افراد فکر می کند که ابتلا به HIV فقط برای دیگران اتفاق می افتد. با این وجود بسیاری به HIV مبتلا شده و اکنون با آن زندگی می کنند. هنگامی که نام این ویروس به گوش می رسد این سوال که چه تعدادی از افراد جامعه مبتلا هستند یا اینکه چه تعدادی در معرض ابتلا قرار دارند، به ذهن خطور می کند.

در عین حال، اگر از هر کدام از ما در مورد اینکه آیا به این بیماری مبتلا هستیم یا خیر، سوال شود به دلایل گوناگون از جمله طرد شدن از اجتماع از پاسخ خوداری می کنیم. مشکل عدم پاسخ در سوالات حساس، پژوهشگر را بر آن می دارد تا سعی در استفاده از روش های غیر مستقیم و دقیق تر برای جمع آوری پاسخ ها نماید. در این راستا روش بسط شبکه ای ابزاری استاندارد، برای برآورد اندازه ی گروه های پنهان مانند شناسایی تعداد افراد در معرض خطر ابتلا به HIV است.

با وجود تمام راهکارهای پیشنهاد شده، باز هم برخی از سوالات بدلیل حساسیت قابل توجهشان، عدم اطمینان پاسخگو از محفوظ بودن اطلاعاتش و ترس از فاش شدنشان، بی جواب باقی می مانند. بروز این بی پاسخی مفهوم داده ی گمشده است. برای بدست آوردن نتایج قابل اعتماد و تعمیم پذیر، روش های متفاوتی برای برخورد با این داده ها وجود دارد. هدف ما از انجام این مطالعه بررسی تاثیر داده های گمشده و روش های برخورد با آنها در برآورد اندازه ی گروه های در معرض خطر است.

مواد و روش ها: در این راستا از ۹۹۷ نمونه مربوط به جمعیت عمومی ایران، شیوع سوء مصرف ده ماده ی مخدر محاسبه شد. سپس ۱۰٪، ۳۰٪ و ۵۰٪ از اطلاعات جمع آوری شده برای هر ماده، ۲۰۰ بار حذف شد. اندازه ی گروه ها با استفاده از آنالیز موارد کامل پیش بینی و بعد از آن برآورد داده های گمشده با بکاربردن روش های جایگذاری با میانه، رگرسیون خطی، رگرسیون دوجمله ای منفی و روش Expectation Maximation(EM) انجام شد. در هر بار ماده ای را که دارای داده ی گمشده بود به عنوان خروجی و سایر مواد را متغیرهای پیش بین در نظر گرفته و برآوردهای مورد نظر در روش های یاد شده، با استفاده از نرم افزار های SPSS، STATA و در نهایت نرم افزار Excel محاسبه شدند.

برای مقایسه ی این روش ها نسبت خطا، یعنی تفاضل مقدار برآورد شده از مقدار واقعی، تقسیم بر مقدار واقعی در نظر گرفته شد. سپس برای نسبت خطای شدید مثبت(SRB+)، مقدار بیش از ۱۰٪ و برای نسبت خطای شدید منفی (SRB-)، مقدار کمتر از ۱۰٪- بعنوان خطای شدید

(SRB) تعریف شد. بعلاوه روشی بهتر تلقی می شود که کمترین تعداد این نوع خطا را به خود اختصاص داده است. همچنین از رگرسیون لجستیک استفاده کرده و روش EM را به عنوان گروه رفرنس و دو متغیر دوتایی +SRB و -SRB، متغیرهای خروجی در نظر گرفته شدند. سپس عملکرد هر کدام از روش ها را، در هر مقدار داده ی گمشده، بدست آوردیم.

یافته ها: در نرخ ۱۰٪ و ۳۰٪ داده ی گمشده تفاوت سهم MED و EM برای ایجاد SRBs بترتیب، ۳۵٪ (۴۱٪ در برابر ۶٪) و ۱۰٪ (۲۵٪ در مقابل ۱۰٪) می باشد. اما با افزایش ۵۰٪ داده گمشده، این درصد برای روش های گوناگون، تفاوت اندکی داشت، به طوری که بازه ی تغییرات، (۲۲٪، ۱۸٪) بوده است. همچنین برای روش میانه (MED)، اکثریت نسبت خطاها منفی بوده است و این در حالیست که اکثر SRBs در روش های رگرسیون خطی و رگرسیون دوجمله ای منفی، مثبت بوده اند. این بدان معناست که روش میانه در اکثر موارد سبب کم برآوردی و روش های رگرسیون خطی و دوجمله ای منفی در بیشتر موارد سبب بیش برآوردی اندازه ی گروه ها شده اند. بعلاوه روش EM در نرخ های متفاوت همواره کمترین فراوانی SRB را به خود اختصاص داده است.

در نرخ ۱۰٪ داده ی گمشده، همه ی روش ها نسبت به EM دارای احتمال بیشتری برای تولید +SRB بودند و با افزایش نرخ داده ی گمشده به ۳۰٪ و ۵۰٪، برتری EM نسبت به سایر روش ها کاهش یافته، به طوری که در نرخ ۵۰٪، تنها روش Reg با روش EM از لحاظ عملکرد بهتر تفاوت معنی داری داشته است. همچنین عملکرد روش ها در ایجاد -SRB به گونه بوده که در نرخ ۱۰٪ داده گمشده، همه روش ها بجز روش NB تفاوت معنی داری داشته اند و با افزایش درصد داده گمشده، سایر روش ها از لحاظ عملکرد مناسب، خود را به EM رسانده اند. اما روش MED در همه نرخ ها به بدترین شکل عمل کرده است

در این مطالعه روش میانه و رگرسیون خطی، عملکرد ضعیفی از خود نشان دادند. در ۱۰٪ داده ی گمشده، روش EM تاحدی خطا را کاهش داده است. این در حالیست که در نرخ متوسط، هیچ یک از روش ها رضایت بخش نبوده اند.

کلمات کلیدی: ایدز، گروه پنهان، داده گمشده، بسط شبکه ای، برآورد اندازه

Abstract

Network Scale-Up is a standard tool in size estimation of hidden groups. Our aim is to address the impact of missing data and imputation methods on its results. Recruiting 997 Iranian from general population, the prevalence of misuse of ten drugs was calculated. Then 10%, 30%, and 50% of data were deleted 200 times. Size of groups were predicted analyzing complete case (CC); and after imputation by median replacement (MED), linear and Negative Binomial regression (NB), and Expectation Maximum (EM). For positive Relative Biases (RB), values>10% were defined as Severe Relative Bias (SRB+). For negative RBs, SRB- was defined as values <-10%. At 10% and 30% missing rates, difference between contribution of MED and EM to create SRBs were 35% (41% versus 6%) and 10% (25% versus 10%). For MED, majority of SRBs happened was SRB-. However, majority of SRBs seen in linear and NB regression were SRB+. At 10%, relative to EM, all methods were more likely to produce SRB. By increase in missing rate superiority of EM over other methods reduced. MED and linear regression imputations were the poorest methods. At 10% missing, EM partially reduced bias. However, at moderate missing rate, performance of no method was satisfying.

Keywords: AIDS, Hidden Group, Missing Data, Network Scale Up, Size Estimation