دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایان نامه مقطع کارشناسی ارشد رشته آمار زیستی

عنوان:

تأثیر روش های مختلف برآورد داده های گمشده بر نتایج حاصل از مدل های رگرسیونی

توسط: سعیده حاجی مقصودی

استاد راهنما: دکتر محمدرضا بانشی

استاد مشاور: دکتر علی اکبر حق دوست

سال ۱۳۹۱

Impact of method of imputation of missing data on composition of regression models

A Thesis

Presented to

The Graduate Studies

By

Saiedeh Haji Maghsoudi

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in:

Biostatistics

Kerman University of Medical Sciences

June 2012

چکیده

مقدمه و هدف: وجود داده های گمشده مشکلی است که در اکثر مطالعات پزشکی وجود دارد، و باعث مشکلات عدیده ای هنگام تحلیل داده ها می شود. روش های مختلفی برای برخورد با داده های گمشده وجود دارد، که استفاده از هر یک از این روش ها مستلزم برقراری برخی فرضیات می باشد. در این مطالعه تأثیر این روش ها بر برآوردهای حاصل از مدل رگرسیونی لجستیک مورد بررسی قرار گرفته است. هنگامی که از روش انتساب چندگانه برای برآورد داده های گمشده استفاده می شود، چند مجموعه داده برآورد می شود، مشکلی که وجود دارد این است که هنگام استفاده از روش های انتخاب متغیر نظیر حذف رو به جلو یا حذف رو به عقب، متغیرهای معنی دار در مجموعه داده های مختلف با هم متفاوت هستند. این امر انتخاب بهترین مدل و تفسیر نتایج را با مشکل مواجه می سازد. روش روبین (RR) به عنوان روش استاندارد تحلیل داده ها، در چنین مواقعی کاربرد دارد. اما چون این روش پیچیده و طولانی می باشد، در این مطالعه به مقایسه چند روش ساده دیگر برای انتخاب متغیر و ساخت مدل هنگام استفاده از روش انتساب چندگانه پرداخته شده است.

مواد و روش ها: در این مطالعه از داده های بانک اطلاعاتی مراقبت رفتاری برنامه کشوری مراقبت سرولوژیکی و رفتاری HIV در زندانیان استفاده شده است. نمونه ای به حجم ۲۷۲۰ که یک نمونه داده کامل فاقد داده گمشده بود، از این مجموعه داده انتخاب گردید. متغیر سابقه مصرف مواد به صورت تزریقی (بلی/ خیر) به عنوان متغیر وابسته و ۲۱ متغیر دیگر به عنوان متغیرهای مستقل وارد مطالعه شدند. از این تعداد متغیر، سه متغیر به صورت کمی و بقیه متغیرها دوحالتی (۱۵ متغیر) یا رده بندی (سه متغیر) بودند. در این حالت نسبت تعداد رویدادها به ضرایب رگرسیونی براورد شده (EPV(Event Per Variable)) برابر ۲۵ می باشد. به منظور بررسی تأثیر حجم نمونه، نمونه هایی به حجم ۲۳۵۲ (EPV=10) و ۲۲۲۷ (EPV=5) به صورت تصادفی از مجموعه داده ۲۷۲۰ تایی انتخاب شد.

این سه مجموعه داده به عنوان مدل کامل در نظر گرفته شدند. و داده ها با استفاده از مدل رگرسیون لجستیک و روش انتخاب متغیر رو به عقب مورد تحلیل قرار گرفتند. نتایج تحلیل این داده ها به عنوان استانداردی برای مقایسه نتایج روش های مختلف در نظر گرفته شدند.

برای بررسی تأثیر داده های گمشده و برآوردهای حاصل از آن بر نتایج، داده های گمشده با میزان های ۵، ۱۰، ۳۰، ۵۰ و ۷۵ درصد ایجاد شدند. سپس این داده ها با روش های جایگزینی میانه، جایگزینی رگرسیونی، الگوریتم EM، انتساب چندگانه به روش MICE برآورد شدند. مجموعه داده های حاصل از روش های مختلف برآورد داده گمشده، تحلیل و با مدل کامل مقایسه شدند. همچنین داده ها با روش حذف موارد با داده ناکامل نیز مورد تحلیل و مقایسه با مدل کامل قرار گرفتند.

در قسمت دیگری از این مطالعه، روش های مختلف انتخاب متغیر پس از برآورد داده های گمشده با روش انتساب چندگانه مورد مقایسه قرار گرفت. در روش های S1، S2 و S3 متغیرهایی که حداقل در یک داده، حداقل در پنج داده و معنی دار در تمام داده ها بودند برای مدل نهایی انتخاب شدند. همچنین داده های تولید شده ترکیب و یک مجموعه داده به دست آمد که با روش های رگرسیون وزنی تحلیل و با سایر روش ها مقایسه شدند. روش رگرسیون وزنی W1 وزن یک دهم و روش W2 وزنی متناسب با نسبت داده های گمشده را به متغیرها اعمال می کنند.

یافته ها: هنگامی که میزان داده گمشده کم بود اریبی در نتایج نیز کم بود اما با افزایش میزان داده گمشده به ویژه در روش حذف موارد با داده ناکامل اریبی نیز افزایش یافت. در EPV برابر ۳۵، تنها در میزان داده گمشده ۵۰ و ۷۵ درصد اریبی مشاهده شده است. و روش حذف موارد با داده گمشده تنها روشی است که در همه متغیرها دارای اریبی می باشد.

هنگامی که EPV برابر ۱۰ می باشد، روش حذف موارد با داده گمشده در همه میزان های داده گمشده دارای اریبی بوده است و در میزان داده گمشده ۵۰ و ۷۵ درصد علاوه بر اریبی در ضرایب و خطای معیار متغیرها، منجر به حذف اشتباه به ترتیب دو و چهار متغیر معنی دار از مدل شده است. و در ۵۰ درصد داده گمشده یک متغیر را به اشتباه وارد مدل کرده است. غیر از روش حذف موارد با داده گمشده، روش جایگزینی میانه در میزان داده گمشده ۲۰ و ۷۵ درصد دو متغیر را به اشتباه معنی دار تشخیص داده است. نتایج بقیه روش ها کمابیش مشابه می باشد.

در EPV برابر ۵، غیر از روش حذف موارد با داده ناکامل که در همه میزان ها جز ۱۰ درصد داده گمشده، دارای اریبی می باشد، روش جایگزینی رگرسیونی در میزان داده گمشده ۵۰ درصد و جایگزینی میانه در ۷۵ درصد داده گمشده در یک متغیر دارای اریبی می باشند. نتایج روش های انتساب چندگانه با روش روبین و الگوریتم EM در میزان داده گمشده کمتر از ۷۵ درصد مشابه مدل کامل می باشد.

با در نظر گرفتن معیارهای AIC و سطح زیر منحنی راک و همچنین مقایسه متغیرهای باقیمانده در مدل در مقایسه با مدل کامل، نتایج نشان داد که روش انتساب چندگانه به عنوان یک روش استاندارد، نسبت به سایر روش ها بهتر عمل کرده است.

استفاده از روش انتساب چندگانه برای برآورد داده های گمشده وقتی میزان داده گمشده افزایش می یابد، منجر به نتایج واقع بینانه تری نسبت به سایر روش ها می شود. روش های الگوریتم EM و جایگزینی رگرسیونی پس از روش انتساب چندگانه نتایج نزدیک به واقعیت (در مقایسه با مدل کامل) را ارائه کرده اند. روش جایگزینی میانه در میزان های ۵ و ۱۰ درصد داده گمشده مشابه بقیه روش های برآورد داده گمشده بوده است اما در میزان های بیشتر داده گمشده منجر به معنی داری اشتباه برخی متغیرها در مدل شده است.

مقایسه روش های مختلف انتخاب متغیر پس از انتساب چندگانه نشان داد که هنگامی که EPV برابر ۲۵ بوده است در میزان داده گمشده ۵، ۱۰ و ۲۰ نتایج همه روش ها مشابه مدل کامل بوده است. در میزان داده گمشده ۵۰ و ۷۵ درصد ، همه روش های انتخاب متغیر غیر از روش حذف موارد با داده ناکامل، به طور مشابه یک متغیر را به اشتباه به عنوان یک متغیر معنی دار وارد مدل کرده اند.

در EPV برابر ۱۰، همه روش های انتخاب متغیر غیر از روش حذف موارد با داده ناکامل، در میزان داده گمشده ۱۰، ۲۰، ۵۰ و ۷۵ درصد نتایج مشابه مدل کامل ارائه کرده اند. در میزان داده گمشده ۵۰ درصد غیر از روش رگرسیون وزنی W2 و روش Single بقیه روش ها در برآورد ضریب یک متغیر دارای اریبی بوده اند.

در EPV برابر ۵، در میزان داده گمشده ۵، ۱۰ و ۲۵ نتایج همه روش ها غیر از روش حذف موارد با داده ناکامل، مشابه مدل کامل بوده است. در میزان داده گمشده ۵۰ درصد، تنها روش های حذف موارد با داده گمشده و S3 دارای اریبی بوده اند. در میزان داده گمشده ۷۵

درصد روش حذف موارد با داده گمشده در همه متغیرها دارای اریبی بوده است و بقیه روش ها به طور مشابه معنی داری یک متغیر را از دست داده اند. روش Single علاوه بر این، اریبی در ضریب یک متغیر را نتیجه داده است.

به طور کلی، در بین روش های انتخاب متغیر بعد از برآورد داده های گمشده با روش انتساب چندگانه، روش S3 منجر به حذف اشتباه یک متغیر معنی دار از مدل شده است در حالی که نتایج انتخاب متغیر بقیه روش ها شبیه روش RR می باشد و از لحاظ اریبی هم نتایج کمابیش به هم نزدیک است. روش های S1، S2 و روش های وزنی نتایجی مشابه روش RR را نتیجه داده اند. نتایج روش رگرسیون وزنی W2 مشابه روش RR بوده است و در یک مورد که بقیه روش ها حتی RR اریبی داشته اند. این روش مشابه مدل کامل عمل کرده است، یعنی فاقد اریبی بوده است.

نتیجه گیری: استفاده از روش های برآورد داده گمشده منجر به نتایج واقع بینانه تری در مقایسه با روش حذف موارد با داده گمشده می شود. روش های الگوریتم EM و جایگزینی رگرسیونی می توانند به عنوان جایگزین های مناسبی برای روش انتساب چندگانه در نظر گرفته شوند. روش حذف موارد با داده گمشده وقتی که نرخ داده گمشده پایین است می تواند به عنوان یک روش ساده مورد استفاده قرار گیرد، ولی با افزایش نرخ داده های گمشده این روش منجر به اریبی می شود. روش رگرسیون وزنی W2 که تنها از یک مجموعه داده استفاده می کند به عنوان یک روش ساده می تواند جایگزین مناسبی برای روش RR باشد. همچنین عملکرد روش های S1 و S2 مشابه RR است که نشان دهنده این موضوع است که می توان با یک غربالگری اولیه متغیرهایی را که حداقل در پنجاه درصد داده ها معنی دار هستند کاندید حضور در مدل چند متغیره کرد.

در کل، هنگامی که در یک تحلیل چند متغیره درصد داده های گمشده زیاد می باشد اما این میزان در هر متغیر به تنهایی مقدار کوچکی (کمتر از ۱۰ درصد) باشد، مشکل چندانی در برآوردها ایجاد نخواهد کرد. مگر در مواردی که از روش حذف موارد با داده گمشده استفاده می شود.

Abstract

Introduction:

Missing data are common in medical studies, and cause severe problems when the data are analyzed. There are different methods for dealing with missing data; the using of these methods needs some assumptions. In this study the impact of these methods on estimates of the logistic regression model has been studied.

When we use multiple imputation methods for estimating missing data, several data sets is estimated. Therefore, there is a problem when using variable selection methods such as forward or backward selection. This is because significant variables are different in multiply imputed data sets.

Rubin method (RR) as the standard method of data analysis is used in these situations. But this method is complicated and time demanding. In this study, different variable selection methods as well as weighted regression models are compared with standard RR so as to enhance the value of statistical modeling

Material and Methods:

Information of national HIV Bio-Behavioral Surveillance Survey (BBSS) among prisoners in 2009 was used. Information of 2720 subjects was analyzed. The dependent variable was history of drug injection (yes/ no question). A total of 21 variables were included as independent variables. Among them, 15 variables were binary, and 3 variables were categorical, and 3 variables were quantative. In this case EPV (Event Per Variable) was equal to 25. For determination of the effect of sample size, sample size of 2352 (EPV = 10) and 2227 (EPV = 5) selected randomly from the data set.

Logistic regression model was fitted to each of these three data sets. Results were considered as gold standard.

To investigate the effect of missing data we generated, missing data rates of 5, 10, 20, 50 and 75 percent rates. Then these data were estimated with the Median Substitution, Regression Imputation, Expectation Maximum (EM) algorithm and Multiple Imputation via Chain Equations (MICE) method. Data sets obtained were analysis and compared with the gold standard. Also data analyzed with Complete Case (C-C) method.

In another part of this study, different methods of variable selection after Multiple Imputation were compared. In total 10 data sets were imputed. In S1, only variables retained significant at least in one imputed data set was candidate for the multifactorial model, in S2 and S3 variables retained in more than 5, and in all 10 data sets were selected to be offered to the multifactorial model. Two weighting schemes were implemented (W1 and W2). In W1 weight of 0.1 was used. In W2 weight (1-f) /10 was applied where f is the mean of fraction of missing rate for all variables. Also RR (Rubin's Rule) was applied as a standard method for this data.

Results:

In the case of Complete-Case (C-C) analysis, we have seen the bias was low at low missing rate. However, increase in missing rate was associated with increases in bias. In particular, at EPV of 25, all methods had bias but C-C method was the only method that had bias on all variables.

At EPV of 10, C-C method had bias on all variables as well. In addition at 50 and 75% missing rate C-C had removed 2 and 4 significant variables respectively. And it had entered one false variable to the model at 50% missing rate. At 20 and 75% missing rate Median Substitution had entered two variables to the model wrongly. Other methods had more or less similar result.

Results at EPV of 5 were as follows: at all missing rate except 10% missing rate C-C method provided biased estimates. At 50 and 75% missing rate respectively Regression Imputation and Median Substitution had bias in one variable.

EM algorithm and Multiple Imputation were similar to full model. With considering AIC and area under ROC curve and comparison the remaining variable in model at missing rate of less than 75%, result showed (that) Multiple Imputation was better than other methods.

At high missing rate Multiple Imputation provides the best fit and performance. EM algorithm and Regression Imputation are good alternative for Multiple Imputation.

At 5 and 10% missing rate Median Substitution was similar to other methods, but use of this ad hoc method at higher missing rates led to biased estimates.

Comparison of variable selection after Multiple Imputation showed that at EPV of 25 all of methods were similar to full model in 5, 10 and 20% missing rate. At 50 and 75% missing rate, all methods except C-C had entered wrongly one variable in the model.

When EPV was 10, at 10, 20, 50 and 75% missing rate all methods except C-C was similar to full model in variable selection. At 50% missing rate all methods except W2 and single had bias.

When EPV was 5, at 5, 10 and 25% missing rate all methods except C-C was similar to full model. At 50% missing rate C-C and S3 had bias in variable selection. At 75% missing rate C-C had bias on all variables and other methods had bias in one variable (one variable removed wrongly). In addition Single method had bias on one variable. And S3 removes one variable wrongly. But other methods were similar to RR. In conclusion, we have seen that S1, s2 and weighting method was similar to RR.

Discussion:

Result of missing data estimation leads to more realistic results than Complete Case. EM algorithm and Regression Imputation are suitable alternative for Multiple Imputation.

Removal of cases with missing data when the missing data rate is low can be used as a simple method, but with increased rates of missing data this method does not work.

W2 weighted regression method that uses only one data set as a simple method can be a good alternative to the RR.

With low missing rate, C-C analysis can be used as a good approximation for MI analysis. However, with higher missing rate this method does not work. Even with high rate of missing rate performance of S1 and S2 methods was similar to that of MI. This indicates that screening step can be used to select candidate variables for. W2 method uses one data set and it can be simple alternative for RR.