دانشگاه علوم پزشکی کرمان

دانشکده بهداشت

پایانامه مقطع کارشناسی ارشد آمار زیستی

عنوان:

مقایسه دو مدل رگرسیون لجستیک و تحلیل جداسازی برای پیش بینی ابتلا به دیابت نوع ۲

توسط : محمدآرام احمدی

استاد راهنما : دکتر عباس بهرام پور

استاد مشاور : دکتر حمید نجفی پور

سال تحصیلی ۱۳۹۲ - ۱۳۹۱

Kerman University of Medicine Sciences

Faculty of Health

Master part thesis of biostatistics

Title:

Comparison of logistic regression model and discriminant analysis to prediction of type 2 diabetes

By: Mohammad Aram Ahmadi

Supervisor: Dr. Abbas Bahrampour

Advisor: Dr. Hamid Najafipour

2013

چکیده:

مقدمه : رگرسیون لجستیک یکی از پرکاربردترین روش های تحلیل آماری در علوم پزشکی و در امر پیش بینی می باشد. تحلیل جداسازی روشی برای جداسازی مشاهدات بر حسب تعداد سطوح متغیر وابسته بوده که می توان پس از تشکیل توابع جداسازی، هر مشاهده جدیدی را نیز به این سطوح متغیر وابسته اختصاص داد. رگرسیون لجستیک و تحلیل جداسازی از جمله روش های آماری چندمتغیره ای هستند که می توانند برای ارزیابی ارتباط بین متغیرهای مستقل هرچند مخدوش کننده، اثرات متقابل متغیرها و یک متغیر وابسته(پیشامد چندحالته) مورد استفاده قرار گیرد.

بیماری دیابت یکی از شایع ترین بیماری های مزمن و از جمله بیماری های اندوکرین است. بیماری دیابت درمان قطعی ندارد و می تواند عوارض کشنده ای ایجاد نماید. این بیماری شایع ترین علت قطع اندام، نابینایی و نارسایی مزمن کلیویی و یکی از مهمترین عوامل خطر در ایجاد بیماری های قلبی است. میزان وقوع جهانی دیابت به دلیل افزایش شیوع چاقی و کاهش میزان فعالیت بدنی در حال افزایش است.

هدف از این مطالعه مقایسه توانایی مدل های رگرسیون لجستیک و تحلیل جداسازی برای پیش بینی دیابت نوع ۲ در نمونه ای از افراد می باشد که از قبل داده های آن جمع آوری شده است.

مواد و روش ها : داده ها شامل ۵۳۵۷ نفر بوده و متغیر دیابت به عنوان متغیر پاسخ درنظر گرفته شد و متغیرهای مستقل شامل: وزن، قد، نمایه توده بدنی (BMI)، محیط دور کمر، محیط دور باسن، نسبت کمر به باسن (WHR)، کلسترول، سن، جنسیت، شغل، تحصیلات، استفاده از داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، اندازه فشار خون سیستولیک و دیاستولیک، سطح HDL ، سطح LDL ، وضعیت مصرف مواد مخدر، فعالیت هایی که منجربه بالارفتن ضربان قلب شود و تری گلیسرید به عنوان متغیرهای مستقل در نظر گرفته شدند. بعد از بررسی همبستگی خطی بین متغیرها و حذف متغیرهای همبسته، مدل های رگرسیون لجستیک و تحلیل جداسازی بر داده ها برازش داده شدند و پیش بینی دیابت بر اساس این مدل ها انجام شد. سپس در مرحله بعد متغیرهای معنی دار بدست آمده را از مدل کامل استخراج کرده و مدلسازی را دوباره بر داده های جدید با عنوان مدل کاهش یافته اعمال نموده و با مدل کامل مقایسه شدند. همچنین برای تعیین تاثیر افزایش حجم نمونه با فرآیند شبیه سازی حجم داده ها را با متغیرهای مستقل اولیه مطالعه به ۲۰۰۰۰ نفر رسانده و ۱۰ سری داده با این حجم تولید شد و سرانجام دو مدل مقایسه شدند. نهایتا از نتایج بدست آمده جهت محاسبه میزان حساسیت، ویژگی، دقت و منحنی راک برای مقایسه قدرت پیش بینی مدل ها و مراحل شبیه سازی و رسم نمودارها از نرم افزارهای SPSS(V20) ، MINITAB(V16) ، STATA(V11) و EASYFIT(V5.5) استفاده گردید.

یافته ها : در مدل کامل، میزان حساسیت برای مدل رگرسیون لجستیک و برای تحلیل جداسازی به ترتیب ۷۷/۷ ٪ و ۲۴/۶ ٪ بود. میزان ویژگی نیز به ترتیب ۷۲ ٪ و ۹۶ ٪ بدست آمد. دقت پیش بینی در مدل

رگرسیون لجستیک ۷۲/۸ ٪ و در مدل تحلیل جداسازی ۸۶/۱۵ ٪ بود. همچنین مساحت زیر منحنی راک برای مدل رگرسیون لجستیک و تحلیل جداسازی به ترتیب ۸۱/۸ ٪ و ۸۱/۵ ٪ بود.

در مدل کاهش یافته با وجود متغیرهای معنی دار استخراج شده از مدل کامل، برای مدل رگرسیون لجستیک میزان حساسیت، ویژگی، دقت و ناحیه زیر منحنی راک به ترتیب ۷۴ ٪ ، ۷۱/۱ ٪ ، ۷۱/۵ ٪ و ۸۰/۳ ٪ و برای مدل تحلیل جداسازی به ترتیب ۳۲/۴ ٪ ، ۹۵/۴ ٪ ، ۸۵/۳ ٪ و ۸۰/۱ ٪ بودند. نتایج شبیه سازی نهایی برای دو مدل با متغیرهای اولیه مطالعه نیز برای معیارهای حساسیت، ویژگی، دقت و منحنی راک در مدل رگرسیون لجستیک به ترتیب : ۹۹/۱۸ ، ۹۸/۴۹ ، ۹۸/۵۹ و ۹۹/۹ درصد بود و در مدل تحلیل جداسازی به ترتیب ۹۲/۶۲ ، ۹۹/۱۹ ، ۹۸/۲۶ و ۹۹/۵۶ درصد گزارش شدند.

نتیجه گیری : یافته ها در دو مدل کامل نشان دادند که حساسیت مدل رگرسیون لجستیک بیشتر بوده ولی ویژگی و دقت پیش بینی مدل تحلیل جداسازی بالاتر از رگرسیون لجستیک بدست آمد. مقایسه منحنی راک در دو مدل نشان داد که مقدار برآورد شده پیش بینی برای دو مدل بسیار بهم نزدیک است. همچنین با افزایش حجم نمونه توسط فرآیند شبیه سازی، نتایج بدست آمده بصورت مدل های مجانبی به طور چشمگیری نزدیک بهم بدست آمده اند.

واژگان کلیدی : رگرسیون لجستیک، تحلیل جداسازی، دیابت، حساسیت، ویژگی، دقت، منحنی راک