مجله تخصصی اپیدمیولوژی ایران؛ ۱۳۹٤، دوره ۱۱، شماره ۳: صفحات ٦۹-٦۲.

مقاله پژوهشی

مقایسه مدل‌های رگرسیون لجستیک با تحلیل جداسازی در پیش‌بینی دیابت نوع ۲

محمد آرام احمدی[۱]، عباس بهرامپور[۲]

[۱] فوق‌لیسانس آمار زیستی، مرکز مدل‌سازی در سلامت و گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت، دانشگاه علوم پزشکی کرمان

[۲] دکترای آمار زیستی، استاد مرکز مدل‌سازی سلامت پژوهشکده اینده پژوهی و گروه آمار زیستی و اپیدمیولوژی دانشکده بهداشت، دانشگاه علوم پزشکی کرمان

نویسنده رابط: عباس بهرامپور، نشانی: دانشکده بهداشت دانشگاه علوم پزشکی کرمان، تلفن: ۰۹۱۳۱۴۰۴۵۱۲، پست الکترونیک: abahrampour@yahoo.com

تاریخ دریافت: ۹۳/۱۰/۰۴؛ پذیرش: ۹۴/۰۳/۰۲

مقدمه و اهداف: بیماری دیابت از جمله بیماری‌های مزمن بوده، که درمان قطعی ندارد و شایع‌ترین علت قطع اندام، نابینایی و نارسایی کلیوی و از عوامل خطر در ایجاد بیماری‌های قلبی است. رگرسیون لجستیک و تحلیل جداسازی از مدل‌های تحلیل آماری برای امر پیش‌بینی و جداسازی چند متغیره می‌باشند. هدف تعیین متغیرهای تأثیرگذار بر دیابت نوع ۲ و مقایسه مدل‌های رگرسیون لجستیک و تحلیل جداسازی می‌باشد.

روش کار: داده‌ها شامل اطلاعات ۵۳۵۷ نفر از مرکز فیزیولوژی دانشگاه علوم پزشکی کرمان میباشد. متغیر پاسخ دیابت ومتغیرهای وزن، قد،(Body Mass Index) (BMI)، دور کمر، دور باسن، نسبت کمر به باسن (Waist hip Ratio) (WHR)، کلسترول و ... در مدل در نظر گرفته شدند. برای مقایسه از حساسیت، ویژگی، دقت، منحنی راک و شبیه سازی استفاده شد.

نتایج: حساسیت، ویژگی و دقت پیش‌بینی در مدل لجستیک و تحلیل جداسازی به ترتیب ۷۴، ۷۱/۱ و ۷۱/۵ و ۹۵/۲۲،۲۲/۴ و ۸۵/۳ بود. منحنی راک برای مدل رگرسیون لجستیک ۸۰/۳ درصد ۰ و تحلیل جداسازی ۸۰/۱ درصد به‌دست آمد. شبیه‌سازی نشان داد که حساسیت، ویژگی، دقت و منحنی راک در مدل لجستیک و مدل جداسازی به ترتیب ۹۹/۱۸، ۹۸/۴۹ ، ۹۸/۵۹ و ۹۹/۹ و ۹۲/۶۲ ، ۹۹/۱۹ ، ۹۸/۲۶ و ۹۹/۵۶ شدند.

نتیجه‌گیری: نتایج نشان داد که دور کمر، سن، جنس، مصرف داروی کاهنده فشار خون، اندازه فشار خون سیستولیک و سطح LDL مهم می‌باشند. حساسیت در مدل لجستیک بیشتر ، اما ویژگی و دقت پیش‌بینی در تحلیل جداسازی بالاتر بود. منحنی‌های راک نشان داد که مقادیر پیش‌بینی به طور مجانبی یکسان بودند.

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

مقدمه

دیابت به مجموعه اختلال متابولیسم کربوهیدرات، چربی و پروتئین که به وسیله یا فقدان ترشح انسولین یا کاهش حساسیت بافت‌ها به انسولین صورت می‌گیرد؛ اطلاق می‌شود. از میان انواع دیابت، دیابت نوع ۲ شایع‌ترین نوع است (۱). میزان وقوع جهانی دیابت به دلیل افزایش شیوع چاقی و کاهش میزان فعالیت بدنی در حال افزایش است (۲). شیوع چاقی در کشورهای صنعتی و توسعه یافته افزایش روزافزون دارد، به‌طوری‌که سازمان جهانی بهداشت، چاقی را اپیدمی جهانی گزارش کرده است (۳).

تشخیص الگوها و طبقه‌بندی، یکی از مهم‌ترین کاربردهای روش‌های آماری در علوم مختلف است. از جمله اهداف عمده طبقه‌بندی و مدل‌سازی در علوم آمار، پیش‌بینی بر اساس شواهد و متغیرها و اطلاعات موجود از یک موضوع خاص است. این امر در علوم آماری توسط روش‌هایی مانند رگرسیون، تحلیل ممیزی (جداسازی)، سری‌های زمانی، رده‌بندی، رگرسیون درختی و سایر روش‌ها انجام می‌گیرد. روش‌های کلاسیک آماری برای مدل‌سازی روابط بین متغیرها دارای تعدادی پیش‌فرض و محدودیت است. در نظر گرفتن یک توزیع پیش‌فرض مانند توزیع نرمال برای متغیر پاسخ، خطی بودن رابطه‌ی پیشنهادی، یکسان بودن واریانس خطاها و غیره از جمله محدودیت‌ها روش‌های کلاسیک هستند (۴،۵).

در بین این روش‌ها رگرسیون لجستیک فرضیه‌های قابل توجهی نمی‌خواهد، اما در تحلیل جداسازی برقراری فرض نرمال بودن چند متغیره برای متغیرهای پیش‌بینی کننده علاوه بر استقلال و تصادفی بودن آن‌ها و برقراری فرض تساوی ماتریس کوواریانس لازم و ضروری می‌باشد. روش‌های طبقه‌بندی آماری سنتی مانند تحلیل جداسازی خطی فیشر و رگرسیون لجستیک به طور

وسیعی در مسائل طبقه‌بندی پزشکی وقتی متغیر اصلی باینری (دو سطحی) باشد، مورد استفاده قرار می‌گیرد (۶). رگرسیون لجستیک و تحلیل جداسازی از جمله روش‌های آماری چند متغیره‌ای هستند که می‌توانند برای ارزیابی ارتباط بین متغیرهای مستقل هرچند مخدوش کننده و یک متغیر وابسته (دو یا چند سطحی) مورد استفاده قرار گیرد.

در مطالعه‌هایی که قبلاً در این زمینه انجام شده، ویلسون و همکاران در سال ۲۰۰۷ میلادی در مطالعه‌ای برای پیش‌بینی رخداد دیابت در افراد بالای۵۰ سال، عوامل خطرساز شامل سن بالا، دور کمر بالا، سابقه‌ی فامیلی دیابت، اختلال تحمل قند خون ناشتا، تری‌گلیسرید بالا و کلسترول HDL پایین را به‌عنوان متغیرهای پیش‌بینی کننده معرفی می‌کنند. در مطالعه‌ی دیگر یکهبارک و همکاران در سال ۱۹۹۸ میلادی انجام دادند، نژاد، چاقی، بیماری‌های قلبی، تری‌گلیسرید بالا و اختلال تحمل قند خون ناشتا و دو ساعته را به‌عنوان عوامل خطرساز مؤثر در رخداد دیابت معرفی کرده‌اند (۷). در مطالعه‌ای که توسط جیمز پرس و ساندرا ویلسون (۱۹۷۸) انجام گرفته دو روش رگرسیون لجستیک و تحلیل جداسازی خطی را بدون در نظر گرفتن پیش‌فرض توزیع نرمال چند متغیره با هم مقایسه نمودند و دریافتند که روش رگرسیون لجستیک برآورد و نتایج بهتری- حتی با نمونه‌های نه چندان بزرگ- نسبت به روش تحلیل جداسازی می‌دهد (۸).

در مطالعه‌ای که توسط ماجا پوهار و همکاران (۲۰۰۴) انجام شده در مقایسه دو روش رگرسیون لجستیک (LR)[۱] و تحلیل جداسازی خطی (LDA)[۲] به این نتایج رسیده‌اند که با وجود شرایط نرمال بودن برای LDA این روش نتایج بهتری از LR ارائه می‌دهد و به هر حال برای حجم نمونه‌های بزرگ نتایج دو روش خیلی به هم نزدیک می‌شوند و هم‌چنین دریافتند که در روش LDA هرچه تعداد سطوح متغیر وابسته بیش‌تر باشد (از ۵ به بالا) توان پیش‌بینی بالاتر خواهد بود و بهتر از LR می‌باشد، اما این قضیه هنگامی که تعداد طبقات کم باشد متفاوت بوده و در واقع LR تنها زمان انتخاب مناسبی است که توزیع به صورت باینری یا دو طبقه‌ای باشد (۹).

مرتضی سدهی و همکاران با مطالعه‌ای که برای پیش‌بینی سندرم متابولیک انجام داده‌اند (۱۳۸۸). با مقایسه دو مدل رگرسیون لجستیک وتحلیل جداسازی به این نتیجه رسیدند که میزان حساسیت برای مدل‌ها به ترتیب ۰/۴۸۳ و ۰/۶۷۷ بود. ویژگی به ترتیب ۰/۸۵۷ و ۰/۶۶ به‌دست آمد و مساحت زیر منحنی راک (ROC)[۳] برای دو مدل به ترتیب ۰/۷۴۹ و ۰/۷۳۹ بود (۴). در مطالعه‌ای که توسط جورج آنتونوجورجس و همکاران (۲۰۰۹) انجام گرفت، دو مدل رگرسیون لجستیک و تحلیل جداسازی خطی را برای ارزیابی فاکتورهای مرتبط با آسم مقایسه کردند و میزان دقت، حساسیت و ویژگی مدل رگرسیون لجستیک و تحلیل جداسازی خطی را به ازای نقطه برش‌های مختلف به‌دست آوردند، که درنهایت به این نتیجه رسیدند که به طور کلی تابع جداسازی خطی زمانی که فرضیه‌های نرمال بودن برقرار باشد، روشی بهتر از رگرسیون لجستیک می‌باشد و اختلاف‌های بین دو روش زمانی که حجم نمونه به اندازه‌ی کافی بزرگ باشد، بسیار ناچیز است (۱۰).

هدف از این مطالعه تعیین متغیرهای تأثیرگذار بر ابتلا به دیابت نوع ۲ و مقایسه توانایی مدل‌های رگرسیون لجستیک و تحلیل جداسازی برای پیش‌بینی دیابت در نمونه‌ای از افراد که از قبل داده‌های آن از مرکز تحقیقات فیزیولوژی دانشگاه علوم پزشکی کرمان جمع‌آوری شده‌اند، می‌باشد.

روش کار

رگرسیون لجستیک

فرمی از رگرسیون می‌باشد و زمانی مورد استفاده قرار می‌گیرد، که متغیر وابسته به صورت دو یا چندسطحی بوده و متغیرهای مستقل از هرنوع دیگر باشند. در علوم پزشکی متغیر پیشامد معمولاً حضور یا غیاب یک وضع بیان شده یا یک بیماری می‌باشد (۱۰). برای این مدل اگر دو گروه وجود داشته باشد، رگرسیون لجستیک باینری مورد استفاده قرار می‌گیرد و چنان‌چه سه و بیش‌تر از سه گروه موجود باشد بین رگرسیون لجستیک اسمی و ترتیبی باید انتخابی صورت گیرد (۱۱).

از آن‌جایی‌که احتمال پیش‌بینی شده باید بین اعداد ۰ و ۱ قرار گیرد، روش‌های رگرسیون خطی ساده برای دست‌یابی به آن کفایت نمی‌کند، به این دلیل که آن‌ها به متغیر وابسته اجازه می‌دهند که از این محدودیت‌ها گذشته و نتایج ناسازگار تولید کنند. با تعریف P به عنوان احتمال تعلق یک مشاهده به گروه

[۱] Logistic Regression; LR
[۲] Linear Discriminant Analysis; LDA
[۳] Receiver OperatingCharacteristic; ROC

بیمار و P-۱ احتمال تعلق یک مشاهده به گروه غیر بیمار، مدل رگرسیون لجستیک به صورت زیر است:

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = \alpha + \sum_{i=1}^{k} \beta_i . x_i$$

که p احتمال تخصیص مشاهده‌ای به سطحی از متغیر وابسته (گروه بیمار)، x_i امین متغیر مستقل و β_i ضریب برآورد شده مدل برای i امین متغیر مستقل و β_0 یک عرض از مبدأ می‌باشد. پارامترهای β_0 تا β_k از مدل رگرسیون لجستیک توسط روش حداکثر درست‌نمایی برآورد شده‌اند (۱۰،۴).

در تساوی بالا تبدیل logit برای مرتبط ساختن احتمالات عضویت گروه به یک تابع خطی از متغیرهای پیش‌بینی کننده مورد استفاده قرار گرفته است (۱۱).

روش حداکثر درست‌نمایی (ML)[1] برای حداکثر کردن تابع درست‌نمایی داده‌های گرفته شده از برآورد پارامترها طراحی شده است. این الگوریتم‌ها با کم‌ترین مربعات رگرسیونی تفاوت دارد؛ زیرا داده‌های ورودی متغیر وابسته از نوع ترتیبی، طبقه‌ای و باینری هستند (۱۲،۱۳).

از مزایای استفاده از مدل رگرسیون لجستیک علاوه بر مدل‌سازی مشاهده‌ها، امکان پیش‌بینی احتمال تعلق هر فرد به هر یک از سطوح متغیر وابسته و هم‌چنین امکان محاسبه‌ی مستقیم نسبت شانس[2] با استفاده از ضرایب مدل است (۴).

تحلیل جداسازی

از جمله روش‌های چند متغیره است که با تفکیک کردن مجموعه‌های متمایز مشاهده‌ها و نیز تخصیص مشاهده‌های جدید به گروه‌های تعریف شده که همان سطوح یا طبقات متغیر وابسته می‌باشند؛ سروکار دارد. معیاری که در این‌جا مطرح است، به وجود آوردن یک قانون یا تابع تشخیص بر مبنای اندازه‌های حاصل از مشاهده‌ها می‌باشد. مشاهده‌های جدیدی را که معلوم نیست از کدام گروه هستند به وسیله این تابع به‌دست آمده، می‌توان به یکی از سطوح یا گروه‌های متغیر وابسته نسبت داد. از معروف‌ترین توابع مورد استفاده در تحلیل جداسازی خطی می‌توان به تابع جداسازی فیشر اشاره کرد (۴).

تحلیل جداسازی خطی روی ارتباط بین متغیرهای مستقل چندگانه و یک متغیر وابسته چندسطحی توسط تشکیل ترکیبی از متغیرهای مستقل متمرکز است و می‌تواند یک مدل طبقه‌بندی را برای پیش‌بینی مشاهده‌های جدید به گروه‌های از پیش تعیین شده از آن به‌دست آورد. در ساده‌ترین نوع تحلیل جداسازی خطی یعنی تحلیل جداسازی خطی دو گروهی، یک تابع جداسازی خطی[3] که از میان نقاط ثقل (مراکز هندسی) دو گروه می‌گذرد؛ می‌تواند برای جداسازی بین دو گروه مورد استفاده قرار گیرد (۱۰).

تابع جداسازی خطی به صورت زیر نمایش داده می‌شود:

$$LDF = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

که β_0 عرض از مبدأ و β_1 تا β_k ضرائب رگرسیونی k متغیر موجود هستند.

برای محاسبه‌ی احتمال این‌که یک مشاهده داده شده به یک گروه تعلق داشته باشد، از فاصله ماهالانوبیس استفاده می‌شود. فاصله‌ی ماهالانوبیس، فاصله‌ی مشاهده از مرکز ثقل گروه در فضای چند بعدی تعریف شده توسط متغیرهای پیش‌بینی کننده است. فاصله ماهالانوبیس وقتی که متغیرهای فضای چند بعدی همبسته هستند؛ معیاری مناسب از فاصله می‌باشد. اگر متغیرهای پیش‌بینی کننده ناهمبسته باشند؛ از فاصله‌ی اقلیدسی استفاده می‌شود (۱۱).

وقتی که گروه‌ها (سطوح متغیر وابسته) بیش از دو گروه باشند، به «تعداد گروه‌ها منهای یک» تابع برای طبقه‌بندی یک مشاهده به میان آن‌ها مورد نیاز است. برای هر یک از گروه‌ها، تحلیل جداسازی خطی متغیرهای توضیحی را با توزیع نرمال و ماتریس کوواریانس‌های مساوی فرض می‌گیرد. برای هر مشاهده، ضریب برآورد شده برای یک متغیر مستقل در مقدار مشاهده‌ی آن متغیر ضرب می‌شود. این حاصل‌ضرب‌ها با یک مقدار ثابت جمع می‌شوند و نتیجه یک نمره ترکیبی خواهد بود که امتیاز جداسازی برای آن مشاهده می‌باشد (۱۰).

تکنیک‌های روش‌های تحلیل جداسازی و رگرسیون لجستیک دو مقوله متمایز از هم هستند و همان‌طور که انتظار می‌رود اگرچه آن‌ها باهم مرتبطاند، اما به طور کلی راه‌حل‌های پیشنهاد شده برای یکی متفاوت از دیگری است (۸).

هدف رگرسیون لجستیک پیدا کردن بهترین برازش و به صرفه‌ترین مدل برای توصیف ارتباط بین متغیر پیشامد (پاسخ یا وابسته) و مجموعه‌ای از متغیرهای مستقل (پیش‌بینی کننده یا توضیحی) است. روشی نسبتاً قوی انعطاف‌پذیر و با کاربردی آسان

[1] Maximum Likelihood; ML
[2] Odds Ratio; OR
[3] Linear Discriminant Function; LDF

است. در رگرسیون لجستیک برخلاف تحلیل جداسازی خطی هیچ فرضیه‌هایی در توزیع متغیرها لازم نیست.

تحلیل جداسازی خطی در هر داده منحصر به‌فردی واریانس نسبت بین و داخل دو گروه را حداکثر می‌کند. بدین‌سان تفکیک‌پذیری حداکثری تضمین می‌شود. استفاده از تحلیل جداسازی خطی برای طبقه‌بندی داده‌ها برای مسأله طبقه‌بندی در بحث تشخیص به‌کار برده شده است (۱۴).

هدف تحلیل جداسازی خطی برای تعیین این‌که متغیر بین دو یا چند کلاس جداسازی شود، مورد استفاده قرار می‌گیرد و نیز برای به‌دست آوردن مدل طبقه‌بندی برای پیش‌بینی عضویت گروه‌ها برای مشاهده‌های جدید به‌کار می‌رود. به این معنی که مشاهده‌های جدید به کدام گروه (سطح متغیر وابسته) تخصیص یابند. برای هرکدام از گروه ها، تحلیل جداسازی خطی متغیرهای توضیحی را با توزیع نرمال و ماتریس کوواریانس مساوی در نظر می‌گیرد (۹).

از تحلیل راک به عنوان یک مقیاس اندازه‌گیری توانایی جداسازی یک مدل استفاده می‌شود که بیش‌ترین ناحیه‌ی زیر منحنی نشان دهنده‌ی توانایی پیش‌بینی بهتر برای مقایسه مدل‌هاست. در نهایت هرچه پیش‌بینی‌ها به واقعیت نزدیک‌تر باشد، مبنای تصمیم‌های صحیح‌تری قرار خواهند گرفت (۱۵،۱۶).

داده‌های پژوهش حاضر از یک نوع مطالعه آینده‌نگر در کرمان بین سال‌های ۹۰-۱۳۸۸ به‌دست آمد. متغیر دیابت (قند خون ناشتا) به عنوان متغیر وابسته در نظر گرفته شد. متغیرهای مستقل پس از حذف هم‌خطی و همبستگی بین برخی متغیرها شامل قد، محیط دور کمر، سن، جنس، شغل، تحصیلات، مصرف داروهای کاهنده فشار خون، اندازه فشار خون سیستولیک، HDL، LDL، وضع مصرف مواد مخدر، فعالیت‌هایی که باعث بالا رفتن ضربان قلب شود و تری‌گلیسرید بودند. اثر متقابل بین HDL و تری‌گلیسرید در فرایند تحلیل کنترل شدند. دقت، حساسیت، ویژگی و منحنی راک برای مقایسه‌ی قدرت پیش‌بینی مدل‌های تحلیل جداسازی و رگرسیون لجستیک محاسبه شدند. فرایند شبیه‌سازی به این صورت انجام گرفت که، توزیع هر متغیر را با توجه به پارامترهای آن که به‌طور عمده با پارامتر توزیع هایگاما و لگ نرمال نزدیک بود، توسط نرم‌افزار EASYFIT نسخه ۵/۵ برازش یافته، سپس از طریق نرم‌افزار MINITAB نسخه ۱۶ داده‌های مورد نظر با توزیع تعیین شده، تولید و شبیه‌سازی می‌شدند.

در کل برای محاسبه‌های یاد شده، ساختن فرایند شبیه‌سازی و ترکیب منحنی‌های راک دو مدل از نرم‌افزارهای MINITAB نسخه ۱۶، EASYFIT نسخه ۵/۵، SPSS نسخه ۲۰، و STATA نسخه ۱۱ استفاده شدند.

یافته‌ها

متغیر وابسته به دو سطح قند خون ناشتای کم‌تر و مساوی ۱۲۶ (غیر دیابتی) و قند خون ناشتای بیش‌تر از ۱۲۶ هم‌چنین افرادی که داروی دیابت مصرف می‌کنند (دیابتی) تقسیم شدند. فراوانی‌ها نشان داد که برای همه ۵۳۵۷ نفر، ۴۶۱۷ نفر غیر دیابتی (۸۶/۲ درصد) و ۷۴۰ نفر دیابتی (۱۳/۸ درصد) بودند که عدد ۰/۱۳۸ به عنوان نقطه‌ی برش احتمال برای محاسبه‌ی مدل‌های تحلیل جداسازی و رگرسیون لجستیک به عنوان احتمال پیشین در نرم‌افزار SPSS تعریف شد. در قدم نخست، مدل‌های رگرسیون لجستیک و تحلیل جداسازی به داده‌ها برازش داده شدند و متغیرهای تحصیلات، وضع شغل و وضع مصرف مواد با در نظر گرفتن سطح آخر به عنوان سطح مبدأ طبقه‌بندی شدند و مدل کامل ساخته شد. در مدل کامل با کنترل اثر متقابل متغیرهای HDL و تری‌گلیسرید، متغیرهای معنی‌دار با مقدار p کمتر یا مساوی از ۰/۰۵ استخراج شدند که شامل قد، محیط دور کمر، سن، جنس، اندازه فشار خون سیستولیک، مصرف داروی کاهنده فشار خون و LDL بودند.

پس از استخراج این متغیرهای معنی‌دار و ساختن مدل کاهش یافته، اثرات اصلی متغیرها و عوامل خطر معنی‌دار دیابت در جدول شماره ۱ نشان داده شده است.

برای مدل تحلیل جداسازی کاهش یافته با اعمال این مدل به متغیرهای معنی دار استخراج شده از مدل کامل، یک متغیر به نام «اسکور» به‌دست می‌آید. متغیر امتیاز از طریق ایجاد یک تابع تحلیل جداسازی ساخته می‌شود و می‌تواند ویژگی همه‌ی مشاهده‌ها باشد. بنابراین با به‌کار بستن مدل رگرسیون لجستیک به عنوان مدل تک‌متغیره بین امتیاز و متغیر وابسته، اثر امتیاز بر متغیر وابسته به صورت ویژه‌ای ساخته می‌شود.

ضریبی که از مدل تحلیل جداسازی به‌دست می‌آید با مدل رگرسیون لجستیک متفاوت خواهد بود. نتایج در جدول شماره ۲ گزارش شده‌اند.

جدول طبقه‌بندی به‌دست آمده در جدول شماره ۳، برای رگرسیون لجستیک حساسیت ۷۴، ویژگی ۷۱/۱، و دقت ۷۱/۵ درصد و برای تحلیل جداسازی حساسیت ۲۲/۴، ویژگی ۹۵/۴، و

دقت ۸۵/۳ درصد به‌دست آمد.

مقادیر منحنی راک در مدل رگرسیون لجستیک و تحلیل جداسازی به ترتیب ۸۰/۳ و ۸۰/۱ درصد بود.

شبیه‌سازی

با اطلاعات اصلی در ابتدای مطالعه و تمامی متغیرهای آن، پس از تعیین توزیع مورد نظر برای تولید داده‌ها توسط نرم‌افزار EASYFIT، داده‌ها که شامل ۱۰ سری با حجم ۲۰۰۰۰ نفر بودند، توسط نرم‌افزار MINITAB شبیه‌سازی شدند. برای هر مجموعه از مقادیر حساسیت، ویژگی، دقت و منحنی راک برای دو مدل محاسبه شدند. با میانگین‌گیری از نتایج ۱۰ مجموعه داده، نتایج در جدول شماره ۴ نشان داده شده است.

جدول شماره ۱- مقادیر ضرایب، نسبت شانس، معنی‌داری و فواصل اطمینان (۹۵ درصد) برای متغیرهای استخراج شده مدل کاهش یافته رگرسیون لجستیک

مدل	رگرسیون لجستیک (مدل کاهش یافته)			
متغیرها	**B**	OR	*p*- value	CI (OR) 95%
قد	۰/۰۱٤	۱/۰۱٤	۰/۰۵۳	(۱/۰۲۸ و ۱/۰۰۰)
محیط دور کمر	۰/۰۳۵	۱/۰۳٦	۰/۰۰۰۱	(۱/۰٤۳ و ۱/۰۲۸)
سن	۰/۰٤٦	۱/۰٤۷	۰/۰۰۰۱	(۱/۰۵٦ و ۱/۰۳۹)
جنسیت	۰/٤۹۹	۱/٦٤۸	۰/۰۰۰۱	(۲/۲۱٦ و ۱/۲۷۷)
مصرف داروی کاهنده فشار خون	۰/٦۹۰	۱/۹۹٤	۰/۰۰۰۱	(۲/٤٤ و ۱/٦۳)
سیستولیک	۰/۰۰۷	۱/۰۰۷	۰/۰۰۲	(۱/۰۱۲ و ۱/۰۰۲)
LDL	-۰/۰۰۳	۰/۹۹۷	۰/۰۱۲	(۰/۹۹۹ و ۰/۹۹۵)

جدول شماره ۲ – مقادیر ضریب، نسبت شانس، معنی‌داری و فواصل اطمینان (۹۵ درصد)برای متغیر امتیاز به‌دست آمده در مدل کاهش یافته تحلیل جداسازی

مدل	تحلیل جداسازی (مدل کاهش یافته)			
متغیر	β	OR	*p*- value	CI (OR) 95%
امتیاز	۰/۹۷۳	۲/٦٤٦	۰/۰۰۰۱	(۲/۸۵۸ و ۲/٤۵۱)

جدول شماره ۳- جدول طبقه‌بندی مدل کاهش یافته برای رگرسیون لجستیک و تحلیل جداسازی (داخل پرانتز)

		پیش‌بینی شده		
	مشاهده شده	دیابت		جمع
		خیر	بلی	
دیابت	خیر	۳۲۸۴ (۴۴۰۷)	۱۳۳۳ (۲۱۰)	۴۶۱۷
	بلی	۱۹۲ (۵۷۴)	۵۴۸ (۱۶۶)	۷۴۰
جمع				۵۳۵۷

جدول شماره ٤-نتایج نهایی فرایند شبیه‌سازی برای دو مدل تحلیل جداسازی و رگرسیون لجستیک

	تحلیل جداسازی				رگرسیون لجستیک			
	ROC	دقت	ویژگی	حساسیت	ROC	دقت	ویژگی	حساسیت
میانگین‌های ۱۰ سری داده	۹۹/۵۶	۹۸/۲۶	۹۹/۱۹	۹۲/۶۲	۹۹/۹	۹۸/۵۹	۹۸/۴۹	۹۹/۱۸

بحث

متغیرهای پیشگوکننده‌ی معنی‌دار در این مطالعه شامل قد، محیط دور کمر، سن، جنس، اندازه فشار خون سیستولیک، مصرف داروی کاهنده فشار خون و LDL بودند که بر اساس ویژگی‌های هر متغیر بر دیگر متغیرها فعالیت‌های شدیدی داشته و روی دیابت مؤثرتر بوده‌اند.

با مقایسه متغیرهای معنی‌دار این مطالعه و مطالعه‌های قبلی، مشخص می‌شود که عوامل خطر و جدی در دیابت می‌تواند به طور مشترک شامل چاقی، تری‌گلیسرید، سطح LDL بالا و سن باشدکه اثرهای اصلی و دیگر عوامل خطرآمیز شامل قد، محیط دور کمر بالا، اندازه فشار خون سیستولیک در این مطالعه سهم بیش‌تری به خود اختصاص دادند.

در مطالعه‌ای که توسط جیمز پرس و ساندرا ویلسون (۱۹۷۸) انجام شد که دو تحلیل رگرسیون لجستیک و تحلیل جداسازی، بدون درنظر گرفتن توزیع نرمال چندمتغیره پیش‌فرض مقایسه شدند و دریافتند که استفاده از تحلیل رگرسیون لجستیک توسط روش برآورد حداکثر درست‌نمایی برای برآورد تحلیل جداساز حتی بدون نمونه‌های خیلی بزرگ نتایج بهتری می‌دهد. در مطالعه‌ای دیگر که توسط ماجاپوهار و همکاران (۲۰۰۴) انجام شد، مقایسه‌ای بین رگرسیون لجستیک و تحلیل جداسازی خطی نشان داد که با وجود این‌که حالت‌های نرمال برای تحلیل جداسازی خطی در نظر گرفته شد، این روش نتایج بهتری از رگرسیون لجستیک ارایه داد. نتایج دو روش وقتی که حجم نمونه بزرگ بودند، به‌هم نزدیک بوده و تفاوت‌های اصلی می‌تواند در نمونه‌های کوچک مشاهده شود (۸و۹). با در نظر گرفتن نتایج به‌دست آمده از این پژوهش، با استفاده از داده‌های واقعی، دقت تحلیل جداسازی نسبتاً بهتر از رگرسیون لجستیک بود هرچند در فرایند شبیه‌سازی با افزایش حجم نمونه، نتایج دو روش به‌هم نزدیک شدند.

در محاسبه‌هایی که برای دو مدل در این مطالعه انجام شد، طبق جدول شماره ۳ دقت، حساسیت، ویژگی و منحنی راک برای رگرسیون لجستیک به ترتیب ۷۱/۵، ۷۴، ۷۱/۱، و ۸۰/۳ درصد و برای تحلیل جداسازی به‌ترتیب ۸۵/۳۶، ۲۲/۴۳، ۸۵/۳۶، و ۸۰/۱ درصد به‌دست آمدند که در مقایسه با مطالعه‌ای که توسط مرتضی سدهی و همکاران (۱۳۸۸) انجام گرفته بود، میزان حساسیت برای مدل‌های رگرسیون لجستیک و تحلیل جداسازی به ترتیب ۰/۴۸۳ و ۰/۶۷۷ بود. ویژگی به ترتیب ۰/۸۵۷ و ۰/۶۶ به‌دست آمد و مساحت زیر منحنی راک (ROC) برای دو مدل به ترتیب ۰/۷۴۹ و ۰/۷۳۹ بود (۴) که با وجود مقداری تفاوت در دیگر شاخص‌ها، اما در پیش‌بینی منحنی راک، مطالعه حاضر دارای دقت بالاتری می‌باشد.

هم‌چنین در مطالعه‌ای که توسط آنتونوجرجوس و همکاران (۲۰۰۹) انجام شد در نهایت به این نتیجه رسیدند که به طور کلی تابع جداسازی خطی زمانی که فرضیه‌های نرمال بودن برقرار باشد، روشی بهتر از رگرسیون لجستیک می‌باشد و اختلاف‌های بین دو روش زمانی که حجم نمونه به اندازه کافی بزرگ باشد، بسیار ناچیز است (۱۰)، که این مطلب نیز با ادعای مطالعه حاضر هم‌خوانی دارد و همان‌طور که گزارش شد با افزایش حجم نمونه در فرایند شبیه‌سازی دقت پیش‌بینی مدل تحلیل جداسازی بالاتر رفته و به مدل رگرسیون لجستیک نزیک شد که اگر فرضیه‌های نرمال بودن نیز برقرار باشد چه بسا از مدل رگرسیون لجستیک نیز بهتر نتیجه دهد.

در وضع فعلی جامعه و زندگی نوین اجتماعی و هم‌چنین ویژگی‌های رژیم‌های غذایی می‌توان با تغذیه‌ای مناسب به دور از فست‌فودها و غذاهای چاق کننده و هم‌چنین برنامه‌ریزی برای فعالیت‌های فیزیکی مستمر از جمله نرمش‌های بدنی و هوازی از افزایش وزن و بالارفتن کلسترول و اندازه فشار خون جلوگیری کرد، که این مسأله می‌تواند با توجه به نتایج مطالعه حاضر و مطالعه‌های گذشته به شکلی منجر به کاهش رشد ابتلا به دیابت در جامعه شود.

نتیجه‌گیری

بر اساس نتایج به‌دست آمده از داده‌های شبیه سازیدر مدل رگرسیون لجستیک دقت، حساسیت و منحنی راک بیش‌تر از تحلیل جداسازی بود، اما در تحلیل جداسازی ویژگی بهتر از رگرسیون لجستیک به‌دست آمد.

در نهایت، نتایج در مدل‌های کاهش یافته نشان داد که مقدار حساسیت در رگرسیون لجستیک بهتر از تحلیل جداسازی بود، اما مقادیر ویژگی و دقت در تحلیل جداسازی بهتر از رگرسیون لجستیک بود و منحنی راک در دو روش تقریباً یکسان بود.

در این پژوهش مشخص شد که مقدار حساسیت در مدل تحلیل جداسازی با افزایش حجم نمونه بهبود یافت و به طور کلی

می‌توان نشان داده شود که در فرایند شبیه‌سازی هرچه حجم نمونه بزرگ باشد، نتایج مدل‌های رگرسیون لجستیک و تحلیل جداسازی به‌هم نزدیک‌تر خواهد بود.

تشکر و قدردانی

نگارندگان برخود لازم می‌دانند که از مرکز تحقیقات فیزیولوژی دانشگاه علوم پزشکی کرمان برای در اختیار گذاردن داده‌ها تشکر به‌عمل آورند. این مقاله بخشی از پایان‌نامه کارشناسی ارشد آمار زیستی بوده و بخشی از آن با حمایت مرکز تحقیقات مدل‌سازی در سلامت دانشگاه علوم پزشکی کرمان انجام شده است.

منابع

1. Meraci M, Feizi A, BagherNejad M.Investigating the Prevalence of High Blood Pressure, Type 2 Diabetes Mellitus and Related Risk Factors According to a Large General Study in Isfahan- Using Multivariate Logistic Regression model. Journal of Health Systems Research. 1391; 8: 193-203.
2. Esmaeilnasab N, Afkhamzadeh A, Ebrahimi A. Sectionalstudy of risk factors in type 2 diabetes in the Diabetes Center Sanandaj. Iran Research Journal of Epidemiology. 1389; 6: 39-45.
3. Zabetian A, Hadaeq F, Harati H and Azizi F. Anthropometric indices forecasts of incident type 2 diabetes in adults: Tehran Lipid and Glucose Study. Journal of Diabetes and Lipid Disorders. 1384; 5: 143-51.
4. Sedehi M, Mehrabi Y, Kazemnejad A, Hadaegh F. Comparison of artifitial neural network,logistic regression and discriminant analysis methods in prediction of metabolic syndrome. Iranian journal of endocrynology and metabolism (IJEM). 2010; 11: 638-46.
5. Osborne J, Waters E. Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation. 2002; 8: 1-9.
6. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I and Mendonça A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of Linear Discriminant Analysis, Logistic Regression, Neural Networks, Support Vector Machines, Classification Trees and Random Forests. BMC Research Notes 2011, 4:299.
7. Mehrabi Y, Khadem-Maboudi A, Sarbakhsh P and Hadaegh F. Prediction of Diabetes Using Logic Regression. Iranian Journal of Endocrinology and Metabolism. 2010; 12: 16-24.
8. Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association. 1978; 699-705.
9. Pohar M, Blas M and Turk S. Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodolski Zvezki. 2004; 1: 143-61.
10. Antonogeorgos G, Panagiotakos DB, Priftis KN and Tzonou A. Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10-to 12-years-old children: divergence and similarity of the two statistical methods. International Journal of Pediatrics. Volume 2009, Article ID 952042, 6 pages.
11. Worth AP, Cronin MTD. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. Journal of Molecular Structure: THEOCHEM. 2003; 622: 97-111.
12. Peng CYJ, Lee KL and Ingersoll GM. An introduction to logistic regression analysis and reporting. The Journal of Educational Research. 2002; 96: 3-14.
13. Harris DV, Pan G. Mineral favorability mapping: a comparison of artificial neural networks, logistic regression, and discriminant analysis. Natural Resources Research 1999; 8: 93-109.
14. Balakrishnama S, Ganapathiraju A. Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing. 1998
15. AbdolmalekiP, Yarmohammadi M and Gity M. Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings. Iranian Journal of Radiation Research. 2004; 1: 217-28.
16. Makian N, Almodaresi M, Karimi T. Comparison of artificial neural network models with logistic regression and discriminant analysis methods in predict of Companies Bankruptcy. Journal of Economic Research. 1389; 10: 141-61.

Iranian Journal of Epidemiology 2015; 11(3): 62-69.
Original Article

Comparison of Logistic Regression and Discriminant Analysis in Predicting Type 2 Diabetes

Aram Ahmadi M[1], Bahrampour A[2]
1- Department of Biostatistics and Epidemiology, Faculty of Health, kerman University of Medical Sciences, kerman, Iran
2-Research Center for Modeling in Health, Institue for Future Studies in Health, Department of Biostatistics and Epidemiology, Faculty of Health, Kerman University of Medical Sciences, Kerman, Iran
Corresponding author: Bahrampour A, abahrampour@yahoo.com

Background and Objectives: Diabetes is a chronic and common metabolic disease which has no curative treatment. Logistic regression (LR) is a statistical model for the analysis and prediction in multivariate statistical techniques. Discriminant analysis is a method for separating observations in terms of dependent variable levels which can allocate any new observation after making discriminating functions. The aim of this study was to compare and determine the effective variables in type 2 diabetes.

Methods: The data included *5357* persons obtained through a cohort study in Kerman, southeastern Iran, in 2009-11. Diabetes was considered the response variable. The independent variables after deleting colinearity and correlated variables included height, waist circumference, age, gender, occupation, education, drugs, systolic blood pressure, HDL, LDL, drug abuse, activities, and triglyceride. Sensitivity, specificity, accuracy, and ROC curve were applied for determining and comparing the prediction power of the models.

Results: The results in the reduced model with extracted significant variables from the full model, the sensitivity of the LR model and DA was 74% and 22.4%, the specificity of the LR model and DA was 71.1 % and 95.4 %, the prediction accuracy of the LR model and DA was 71.5% and 85.3%, and the ROC curve of the LR model and DA was 80.3% and 80.1%, respectively . Simulation showed the sensitivity, specificity, accuracy, and ROC curve was 99.18%, 98.49%, 98.59%, and 99.9% for the LR model and 92.62%, 99.19%, 98.26%, and 99.56% for DA, respectively.

Conclusion: The results showed that the risk factors of diabetes in the logistic regression reduced model were waist circumference, age, gender, LDL level, systolic pressure, and drugs. Also, the sensitivity of the LR model was more than DA while DA had a higher specificity and prediction accuracy. Comparison of the ROC curve showed that the prediction estimated values were rather similar in both models, but the two models were the same asymptotically.

Keywords: Sensitivity, Specificity, Logistic regression, Discriminant analysis, ROC curve