● مجلهٔ دانشگاه علوم پزشکی کرمان، دورهٔ بیست ویکم، شمارهٔ ۲، ص ۱۱۳-۱۰۳، ۱۳۹۳

مقالهٔ پژوهشی

پیش‌بینی دیابت نوع ۲ و تعیین میزان تأثیر عوامل خطر با استفاده از مدل رگرسیون لجستیک

محمد آرام احمدی[۱]، عباس بهرام‌پور*[۲]، حمید نجفی‌پور[۳]

خلاصه

مقدمه: بیماری دیابت یکی از بیماری‌های مزمن بوده که درمان قطعی ندارد و شایع‌ترین علت قطع اندام، نابینایی و نارسایی مزمن کلیویی و یکی از مهم‌ترین عوامل خطر در ایجاد بیماری‌های قلبی است. رگرسیون لجستیک یکی از روش‌های تحلیل آماری در امر پیش‌بینی می‌باشد و از جمله روش‌های آماری چند متغیره‌ای است که می‌تواند برای ارزیابی ارتباط بین متغیرهای مستقل هرچند مخدوش کننده و یک متغیر وابسته مورد استفاده قرار گیرد. هدف این مطالعه، تعیین متغیرهای تأثیرگذار و میزان تأثیر آن‌ها بر ابتلا به دیابت و برآورد یک مدل پیش‌بینی رگرسیون لجستیک بود.

روش: نمونه‌ها شامل ۵۳۵۷ نفر بود و دیابت نوع ۲ به عنوان متغیر پاسخ در نظر گرفته شد. متغیرهای مستقل شامل: وزن، قد، نمایه توده بدنی (Body mass index یا BMI)، محیط دور کمر، محیط دور باسن، نسبت کمر به باسن (Waist/hip ratio یا WHR)، کلسترول، سن، جنسیت، شغل، تحصیلات، استفاده از داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، میزان فشار خون سیستولی و دیاستولی، سطح HDL (High-density lipoprotein)، سطح LDL (Low-density lipoprotein)، وضعیت مصرف مواد مخدر، فعالیت‌هایی که منجر به بالا رفتن ضربان قلب می‌شود و تری‌گلیسرید بودند. جهت تعیین قدرت پیش‌بینی مدل رگرسیون لجستیک از شاخص‌های میزان حساسیت، ویژگی، دقت، ضریب Kappa و منحنی ROC (Receiver operating characteristic) استفاده گردید.

یافته‌ها: میزان حساسیت، ویژگی، دقت و ضریب توافق Kappa برای مدل به ترتیب ۰/۷۶۴، ۰/۷۲۵، ۰/۷۳۱ و ۰/۳۱۲ بود و همچنین مساحت زیر منحنی ROC، ۰/۸۲۲ به دست آمد. از بین متغیرهای موجود به ترتیب تأثیر بر متغیر وابسته (بر اساس نسبت شانس)، متغیر نسبت WHR، مصرف داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، جنسیت، سطح تحصیلات، محیط دور کمر، پیاده‌روی و دوچرخه سواری، وزن، BMI، فشار خون سیستولیک، سن، فشار خون دیاستولیک، محیط دور باسن، تری‌گلیسرید، کلسترول، سطح LDL، استفاده از مواد مخدر، قد، سطح HDL و داشتن فعالیت کاری شدید ۱۰ دقیقه‌ای معنی‌دار بودند.

نتیجه‌گیری: با توجه به ملاک‌های دقت و قدرت پیش‌بینی و نیز با در نظر گرفتن مساحت زیر منحنی ROC (۰/۸۲۲) که توانسته است دقت کلی آزمون را برای تشخیص دیابت در حد خوبی انجام دهد، می‌توان بیان داشت که رگرسیون لجستیک مدل مناسبی برای پیش‌بینی دیابت می‌باشد.

واژه‌های کلیدی: رگرسیون لجستیک، دیابت، جداسازی، توافق

۱- دانشجوی کارشناسی ارشد آمار زیستی، مرکز مدل‌سازی در سلامت و گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمان ۲- استاد آمار حیاتی، مرکز تحقیقات فیزیولوژی و گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمان ۳- استاد فیزیولوژی، مرکز تحقیقات فیزیولوژی، دانشگاه علوم پزشکی کرمان

* نویسنده مسؤول، آدرس پست الکترونیک: abahrampour@yahoo.com

دریافت مقاله: ۱۳۹۲/۱/۲۲ دریافت مقاله اصلاح شده: ۱۳۹۲/۳/۲۱ پذیرش مقاله: ۱۳۹۲/۴/۱۲

مقدمه

دیابت به عنوان یکی از شایع‌ترین بیماری‌های مزمن و از جمله بیماری‌های غدد درون‌ریز است که به دلیل اختلال در متابولیسم گلوکز بر سیستم‌های بدن تأثیرگذار می‌باشد (۱). این بیماری شایع‌ترین علت قطع اندام، نابینایی و نارسایی مزمن کلیوی و یکی از مهم‌ترین عوامل خطر در ایجاد بیماری‌های قلبی است. میزان وقوع جهانی دیابت به دلیل افزایش شیوع چاقی و کاهش میزان فعالیت بدنی در حال افزایش است (۲). شناخت عوامل خطر مؤثر در بروز دیابت یک اقدام اساسی برای برنامه‌های پیشگیری از دیابت در هر جامعه‌ای است؛ چرا که کاهش دادن این عوامل خطر باعث کاهش نرخ بروز دیابت خواهد شد. در این میان، یافتن معادله‌ای برای تعیین اثر عوامل خطر و شدت ارتباط آن‌ها با ابتلا به دیابت، دارای اهمیت فراوانی است (۳). از بین کل متغیرهای این مطالعه، متغیر پاسخ دیابت در نظر گرفته شده و با مدل‌سازی هدف، تعیین میزان تأثیر سایر متغیرها بر دیابت بود.

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 داده‌های 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 می‌شود. در نظر گرفتن یک توزیع پیش‌فرض مانند توزیع نرمال برای متغیرهای پاسخ، خطی بودن رابطه پیشنهادی، یکسان بودن واریانس خطاها و... از جمله محدودیت‌های بعضی روش‌های کلاسیک هستند که هنگام استفاده عملی از این روش‌ها، اگر داده‌های واقعی شرایط مفروض مدل را نداشته باشند امکان‌پذیر نبوده یا با خطای قابل توجه همراه است (۴). در بین این روش‌ها رگرسیون لجستیک فرضیات زیادی را لازم را ندارد و از جمله روش‌های آماری چند متغیره‌ای است که می‌تواند برای ارزیابی ارتباط بین متغیرهای مستقل هرچند مخدوش کننده و یک متغیر وابسته (پیشامد طبقه‌ای) و پیش‌بینی ابتلا به بیماری بر اساس متغیرهای پیشگو در مدل مورد استفاده قرار گیرد.

در پژوهش‌هایی که پیش‌تر در این زمینه انجام شده است؛ Wilson و همکاران در مطالعه‌ای برای پیش‌بینی بروز دیابت در افراد بالای ۵۰ سال، عوامل خطر شامل سن بالا، دور کمر بالا، سابقه فامیلی دیابت، اختلال تحمل قند خون ناشتا، تری‌گلیسرید بالا و HDL (High-density lipoprotein) پایین را به عنوان متغیرهای پیش‌بینی کننده معرفی می‌کنند (۵). نتایج مطالعه Burke و همکاران نشان داد که نژاد، چاقی، بیماری‌های قلبی، تری‌گلیسرید بالا و اختلال تحمل قند خون ناشتا و دو ساعته به عنوان عوامل خطر مؤثر در بروز دیابت می‌باشند (۶). در مطالعه‌ای که توسط مرآثی و همکاران جهت بررسی شیوع و عوامل مرتبط با ابتلا به دیابت بر جمعیت عمومی شهر اصفهان انجام گرفت، شیوع دیابت در زنان دو برابر بیشتر از مردان بود. همچنین ارتباط میان سن، جنس، شاخص توده بدنی (Body mass index یا BMI) و سابقه فامیلی ابتلا به دیابت با ابتلا به دیابت معنی‌دار شد (۷).

در مطالعه Chae و همکاران برای پیش‌بینی عوامل مؤثر بر فشار خون توسط مدل رگرسیون لجستیک، مقادیر دقت پیش‌بینی، حساسیت و ویژگی به ترتیب ۶۳/۸۴، ۶۴/۳۶ و ۶۳/۳۳ درصد به دست آمد (۸). در مطالعه‌ای که توسط Su-juan در مورد کاربرد رگرسیون لجستیک در تحلیل ریسک اعتبار انجام شد، میزان دقت جداسازی مدل رگرسیون لجستیک در کل به ۹۹/۰۶ درصد رسید (۹).

با توجه به مطالعات مشابهی که پیش‌تر در زمینه علل ابتلا به دیابت و شناسایی عوامل خطر این بیماری انجام گرفته است، نتایج گوناگونی بر اساس نژادهای مختلف و رژیم‌های متفاوت به دست آمده‌اند. بدین معنی که نتایج این

مطالعه به ویژگی‌های نژادی و رژیم‌های غذایی خاص افراد بومی در پژوهش وابسته می‌باشد و تفاوت‌هایی بین این دو پارامتر در مکان‌های مختلف وجود دارد که البته در مطالعه‌ای که توسط Burke و همکاران صورت گرفته است نژاد یک عامل تأثیرگذار گزارش شد. یکی از دلایلی که ضرورت تحقیق حاضر را نشان می‌دهد، یافتن عوامل خطر احتمالی و میزان تأثیر آن‌ها بر ابتلا به دیابت (علاوه بر عوامل مشخص شده در مطالعات قبلی) بود. هدف از انجام این مطالعه علاوه بر تعیین عوامل خطر معمول در این بیماری، شناسایی و تعیین میزان تأثیر تک‌تک متغیرها بر ابتلا به دیابت و برآورد آن به کمک یک مدل پیش‌بینی رگرسیون لجستیک با توجه به خصوصیات نژادی و فرهنگی و رژیم افراد بومی بود.

رگرسیون لجستیک

رگرسیون لجستیک زمانی مورد استفاده قرار می‌گیرد که متغیر وابسته به صورت دوتایی، اسمی یا ترتیبی باشد و برای متغیرهای توضیحی یا مستقل هیچ محدودیتی وجود ندارد. در علوم پزشکی متغیر پیشامد به طور معمول حضور یا عدم آن از یک وضعیت بیان شده یا یک بیماری می‌باشد. مفهوم اصلی ریاضی که پایه و اساس رگرسیون لجستیک است، لوجیت یعنی لگاریتم طبیعی نسبت شانس می‌باشد. رگرسیون لجستیک بر اساس یک فرضیه که شامل یک ارتباط لجستیکی موجود بین احتمال عضویت گروه و یک یا چند متغیر پیش‌بینی کننده می‌باشد بنا نهاده شده است (۱۲-۱۰).

از آن‌جا که احتمال پیش‌بینی شده باید بین اعداد ۰ و ۱ قرار گیرد، تکنیک‌های رگرسیون خطی ساده برای دستیابی به آن کفایت نمی‌کند؛ به این دلیل که آن‌ها به متغیر وابسته اجازه داده‌اند که از این محدودیت‌ها گذشته، نتایج ناسازگار تولید کنند. با تعریف P_0 به عنوان احتمال تعلق یک مشاهده به گروه غیر بیمار و P_1 احتمال تعلق یک مشاهده به گروه بیمار، مدل رگرسیون لجستیک به صورت زیر است:

$$z_i = \log(\frac{P_{i1}}{Pi0}) = b_0 + b_1X_{i1} + b_2X_{i2} + ... + b_kX_{ik}$$

که $\frac{P_{i1}}{P_{i0}}$ نسبت شانس نامگذاری شده است. b_j مقدار j امین ضریب که $j = 1, ..., k$ و x_{ij} مقدار i امین مشاهده از j امین متغیر پیش‌بینی کننده است. پارامترهای b_0 تا b_k از مدل لجستیک با استفاده از روش ماکزیمم درست‌نمایی برآورد می‌شوند. در معادله ذکر شده تبدیل Logit جهت مرتبط ساختن احتمالات عضویت گروه به یک تابع خطی از متغیرهای پیش‌بینی کننده مورد استفاده قرار می‌گیرد (۱۲، ۱۰).

مقادیر P_1 و P_0 احتمال اختصاص مشاهده به گروه‌های غیر دیابتی و مبتلا به دیابت بود و در نهایت جهت ساختن آماره z تحت لگاریتم طبیعی توسط نرم‌افزار به دست می‌آید.

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 محاسبه مستقیم نسبت شانس با استفاده از ضرایب مدل است (۴).

در این پژوهش هدف آن بود که با برازش مدل رگرسیون لجستیک بر داده‌هایی که از بانک اطلاعاتی (شامل ۵۹۰۰ نفر از افراد بالای ۱۵ سال شهر کرمان) اخذ شده است (۱۳)، توصیف ارتباط بین متغیر پاسخ (وابسته) و یک مجموعه از متغیرهای پیشگو (مستقل) به ترتیب اهمیت و میزان تأثیر هر یک از متغیرهای پیشگو بر دیابت مشخص گردد.

روش بررسی

با استفاده از بانک اطلاعاتی شامل داده‌های ۵۹۰۰ نفر از افراد بالای ۱۵ سال شهر کرمان که به منظور بررسی شیوع

عوامل خطر بیماری‌های قلبی- عروقی فراخوان و مورد مصاحبه و نمونه‌گیری قرار گرفته بودند و آزمایش‌های متفاوتی از آن‌ها به عمل آمده بود، متغیرهای ضروری جهت پیش‌بینی دیابت از بین کل متغیرها انتخاب و به دلیل وجود داده‌های گم شده در برخی از افراد، پس از حذف آنان، حجم نمونه به ۵۳۵۷ نفر کاهش یافت. از آن‌جا که نمونه‌گیری از کل شهر کرمان بود و حجم به نسبت بالایی داشت، توانسته بود نماینده خوبی از جامعه باشد. متغیر قند خون ناشتا (Fasting blood sugar یا FBS) پس از کددهی به دو گروه مبتلا به دیابت (دارای قند خون ناشتای بالای ۱۲۶ یا افرادی که دیابت آن‌ها پیش‌تر مشخص بود، یعنی داروی ضد دیابت یا انسولین دریافت می‌کردند) و غیر دیابتی (قند خون ناشتای زیر ۱۲۶) متغیر پاسخ در نظر گرفته شد و متغیرهای مستقل شامل: وزن، قد، نمایه توده بدنی، محیط دور کمر، محیط دور باسن، نسبت کمر به باسن (Waist/hip ratio یا WHR)، کلسترول، سن، جنسیت، شغل، تحصیلات، استفاده از داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، اندازه فشار خون سیستولی و دیاستولی، سطح HDL، سطح LDL (Low-density lipoprotein)، وضعیت مصرف مواد مخدر، فعالیت شدید در حدود ۱۰ دقیقه‌ای در حین کار، داشتن پیاده‌روی یا دوچرخه سواری ۱۰ دقیقه‌ای، داشتن فعالیت تفریحی یا ورزشی ۱۰ دقیقه‌ای با شدت زیاد که منجر به بالا رفتن ضربان قلب شود و تری‌گلیسرید بودند. با در نظر گرفتن اکثریت متغیرهای موجود در بانک اطلاعات و حذف متغیرهایی که به دلیل داشتن مشاهدات گم شده زیاد احتمال ریزش حجم داده‌ها را افزایش می‌داد، متغیرهای مناسب انتخاب شدند. همچنین برای بررسی وضعیت همبستگی بین متغیرهای مستقل، دو به دو توسط نرم‌افزار بررسی شدند و به استثنای متغیرهای سن و HDL، قد و فشار خون دیاستولیک همگی با هم همبستگی معنی‌داری داشتند. مدل رگرسیون لجستیک بر داده‌ها برازش داده شد و پیش‌بینی دیابت بر اساس این مدل انجام گردید. از نتایج به دست آمده میزان حساسیت، ویژگی، دقت، آماره Kappa و منحنی ROC (Receiver Operating Characteristic) برای اطلاع از قدرت پیش‌بینی مدل توسط نرم‌افزار SPSS نسخه ۲۰ (version 20, SPSS Inc., Chicago, IL) به دست آمد. آماره یا ضریب Kappa برای تعیین توافق بین مقادیر مشاهده شده و پیش‌بینی شده مورد استفاده قرار گرفت. با توجه به جدول ۱ حالت کلی طبقه‌بندی مقدار آماره Kappa محاسبه شد.

توافق مورد انتظار
$$P_e = \left[\left(\frac{n_1}{n}\right)*\left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right)*\left(\frac{m_0}{n}\right)\right]$$

توافق مشاهده شده
$$P_o = \frac{a+d}{n}$$

$$Kappa, K = \frac{P_O - P_e}{1-P_e}$$

جدول ۱. *جدول محاسبه آماره Kappa*

	مقادیر مورد انتظار	نتایج		
مقادیر مشاهده شده		آری	خیر	جمع
نتایج	آری	a	b	m_1
	خیر	c	d	m_0
	جمع	n_1	n_0	n

نتایج

در بانک اطلاعاتی داده‌ها از کل ۵۳۵۷ فرد، ۴۶۱۷ نفر غیر دیابتی (۸۶/۲ درصد) و ۷۴۰ نفر مبتلا به دیابت (۱۳/۸ درصد) بودند که رقم ۱۳/۸ در نرم‌افزار SPSS به عنوان نقطه برش احتمال جهت محاسبه مدل رگرسیون لجستیک در نظر گرفته شد.

ابتدا مدل رگرسیون لجستیک را با در نظر گرفتن نقطه برش ۱۳/۸ برای تعیین احتمال افراد مبتلا به دیابت در نرم‌افزار اعمال نموده، سه متغیر طبقه‌ای که سطح تحصیلات، شغل و وضعیت مصرف مواد مخدر بود با تعیین سطح پایه آخر (پیش‌فرض نرم‌افزار) طبقه‌بندی شد. خروجی‌های این مدل به صورت زیر بود.

در جدول ۳ تمام متغیرها با توجه به کنترل اثرات متقابل تأثیرگذار بر آن‌ها در مدل تحلیلی رگرسیون لجستیک گزارش شد. با توجه به مقدار نسبت شانس (Odds ratio یا OR) برای متغیرهای مدل که در جدول ۳ ارایه شده است، برای متغیر وزن (β = ۰/۱۱۲) به ازای یک کیلوگرم افزایش وزن، ۱۱/۸ درصد احتمال ابتلا به دیابت افزایش می‌یابد. برای متغیر قد (β = -۰/۰۷۱) به ازای یک سانتی‌متر افزایش قد، ۷/۲۹ درصد احتمال ابتلا به دیابت کاهش می‌یابد. برای متغیر شاخص توده بدنی (β = ۰/۰۵۳) به ازای یک واحد افزایش آن، ۵/۴ درصد احتمال ابتلا به دیابت افزایش می‌یابد. برای متغیر محیط دور کمر (β = ۰/۱۲۰) به ازای یک سانتی‌متر افزایش، احتمال ابتلا به دیابت نیز ۱۲/۷ درصد افزایش می‌یابد. برای محیط دور باسن (β = ۰/۰۲۵) به ازای یک سانتی‌متر افزایش، احتمال ابتلا به دیابت ۲/۵ درصد افزایش می‌یابد. برای متغیر نسبت دور کمر به باسن (WHR) (β = ۲/۶۶۰) به ازای یک واحد افزایش این نسبت، ۱۴/۳۲۱ برابر احتمال ابتلا به دیابت افزایش می‌یابد.

برای متغیر سن (β = ۰/۰۴۶) به ازای یک سال افزایش، احتمال ابتلا به دیابت ۴/۷ درصد افزایش می‌یابد. برای متغیر جنسیت (β = ۰/۷۰۷) شانس ابتلا به دیابت برای زنان ۲/۰۲۸ برابر بیشتر از مردان است. نتایج برای متغیر شغل معنی‌دار نبود. برای متغیر سطح تحصیلات شانس ابتلا به دیابت در گروه با سطح سواد ابتدایی (β = ۰/۶۵۸)، ۹۳/۱ درصد بیشتر از سطح فوق لیسانس و بالاتر (سطح پایه) بود؛ در گروه با سطح راهنمایی (β = ۰/۶۷۲)، ۹۵/۸ درصد بیشتر از سطح فوق لیسانس و بالاتر بود و در گروه با سطح فوق دیپلم (β = ۰/۶۴۵)، ۹۰/۶ درصد بیشتر از سطح فوق لیسانس و بالاتر بود. لازم به ذکر است که این نتایج را طبق مقدار P حاصل شده می‌توان حداقل به ۸۰ درصد از جامعه تعمیم داد.

برای متغیر استفاده از داروی کاهش فشار خون طی دو هفته گذشته با تجویز پزشک (β = ۱/۲۷۹) در مقابل عدم مصرف آن، احتمال ابتلا به دیابت ۳/۵۹۲ برابر افزایش یافت. برای متغیر فشار خون دیاستولی (β = ۰/۰۴۳) به ازای یک واحد افزایش، ۴/۴ درصد و برای متغیر فشار خون سیستولی (β = ۰/۰۵۲) به ازای یک واحد افزایش، ۵/۴ درصد احتمال ابتلا به دیابت افزایش یافت.

جدول ۲. *مقادیر مشاهده شده و پیش‌بینی شده*

مشاهده شده	پیش‌بینی شده		
	دیابت دارد	دیابت ندارد	درصد صحیح
دیابت دارد	۳۳۵۱	۱۲۶۶	۷۲/۶
دیابت ندارد	۱۷۴	۵۶۶	۷۶/۵
درصد کل	-	-	۷۳/۱

دقت: ۰/۷۳۱، حساسیت: ۰/۷۶۴ و ویژگی: ۰/۷۲۵ به دست آمد که نشانگر برازش مناسب مدل بر داده‌های مورد مطالعه است

جدول ۳. *نتایج میزان تأثیرات متغیرها بر ابتلا به دیابت*

متغیرها	β	P	OR	CI (۹۵ درصد)	
				حد بالا	حد پایین
وزن	۰/۱۱۲	< ۰/۰۰۱	۱/۱۱۸	۱/۱۷۵	۱/۰۶۴
قد	۰/۰۷۱	< ۰/۰۰۱	۰/۹۳۲	۰/۹۶۹	۰/۸۹۷
شاخص توده بدنی	۰/۰۵۳	۰/۳۸۰	۱/۰۵۴	۱/۱۸۶	۰/۹۳۷
دور کمر	۰/۱۲۰	۰/۰۱۰	۱/۱۲۷	۱/۲۳۵	۱/۰۲۹
دور باسن	۰/۰۲۵	۰/۵۰۵	۱/۰۲۵	۱/۱۰۲	۰/۹۵۳
نسبت کمر به باسن	۲/۶۶۲	۰/۴۵۴	۱۴/۳۲۱	۰/۱۵۱۹	۰/۰۱۴
سن	۰/۰۴۶	< ۰/۰۰۱	۱/۰۴۷	۱/۰۵۸	۱/۰۳۶
جنسیت	۰/۷۰۷	< ۰/۰۰۱	۲/۰۲۸	۲/۱۱۹	۱/۸۹۷
تحصیلات ابتدایی(سطح پایه = فوق لیسانس و بالاتر)	۰/۶۵۸	۰/۱۷۳	۱/۹۳۱	۲/۲۰۵	۱/۷۸۷
تحصیلات راهنمایی (سطح پایه = فوق لیسانس و بالاتر)	۰/۶۷۲	۰/۱۶۹	۱/۹۵۸	۲/۲۱۶	۱/۷۹۵
تحصیلات فوق دیپلم (سطح پایه = فوق لیسانس و بالاتر)	۰/۶۴۵	۰/۱۹۵	۱/۹۰۶	۱/۹۸۶	۱/۶۸۳
مصرف داروی کاهش فشار خون	۱/۲۷۹	< ۰/۰۰۱	۳/۵۹۲	۴/۳۵۴	۲/۹۶۳
فشار خون دیاستولیک	۰/۰۴۳	۰/۰۱۹	۱/۰۴۴	۱/۰۸۲	۱/۰۰۷
فشار خون سیستولیک	۰/۰۵۲	< ۰/۰۰۱	۱/۰۵۴	۱/۰۷۷	۱/۰۳۰
HDL	=۰/۰۷۸	< ۰/۰۰۱	۰/۹۲۵	۰/۹۵۲	۰/۸۹۹
LDL	۰/۰۰۱	۰/۷۶۸	۱/۰۰۱	۱/۰۰۶	۰/۹۹۶
کلسترول	۰/۰۰۳	۰/۱۵۴	۱/۰۰۳	۱/۰۰۷	۰/۹۹۹
تری‌گلیسرید	۰/۰۱۰	< ۰/۰۰۱	۱/۰۱۱	۱/۰۱۵	۱/۰۰۶
استفاده از مواد مخدر: هرگز (سطح پایه = ترک کرده)	=۰/۰۴۴	۰/۸۷۷	۰/۹۵۷	۰/۹۸۶	۰/۹۲۰
استفاده از مواد مخدر: اکنون (سطح پایه = ترک کرده)	۰/۰۱۸	۰/۹۵۰	۱/۰۱۹	۱/۲۱۰	۰/۹۵۶
فعالیت شدید کاری	=۰/۵۰۷	۰/۱۰۱	۰/۶۰۲	۰/۷۶۶	۰/۴۹۸
پیاده‌روی و دوچرخه سواری	۰/۱۳۶	۰/۱۶۰	۱/۱۴۶	۱/۲۵۰	۱/۱۱۳
فعالیت شدید ورزشی- تفریحی	۰/۰۸۰	۰/۷۳۰	۱/۰۸۳	۱/۱۵۳	۱/۰۰۵

OR: Odds ratio; CI: Confidence intervals, HDL: High-density lipoprotein; LDL: Low-density lipoprotein

برای متغیر سطح HDL (β = ۰/۰۷۸) به ازای یک واحد افزایش، ۸/۱ درصد احتمال ابتلا به دیابت کاهش می‌یابد. برای متغیر سطح LDL (β = ۰/۰۰۱) به ازای یک واحد افزایش در مقدار آن احتمال ابتلا به دیابت ۰/۱ درصد افزایش می‌یابد. برای متغیر کلسترول (β = ۰/۰۰۳) به ازای یک واحد افزایش در مقدار آن احتمال ابتلا به دیابت ۰/۳ درصد افزایش می‌یابد. برای متغیر تری‌گلیسرید (β = ۰/۰۱۰) به ازای یک واحد افزایش در آن احتمال ابتلا به دیابت به اندازه ۱/۱ درصد افزایش می‌یابد. برای متغیر وضعیت استفاده از مواد مخدر با در نظر گرفتن سطح پایه آخر

(پیش‌تر مصرف کرده و حالا ترک کرده‌اند) افرادی که هرگز مواد مخدر مصرف نکرده‌اند (β = -۰/۰۴۴) نسبت به افرادی که پیش‌تر مصرف نموده‌اند و اکنون ترک کرده‌اند به اندازه ۴/۴۹ درصد احتمال ابتلا به دیابت کمتر بود و افرادی که اکنون مواد مخدر مصرف می‌کنند (β = ۰/۰۱۸) نسبت به افرادی که پیش‌تر مصرف نموده‌اند و اکنون ترک کرده‌اند به اندازه ۱/۹ درصد احتمال ابتلا به دیابت بیشتر بود. برای متغیر فعالیت شدید کاری (β = -۰/۵۰۷) در مقابل عدم داشتن فعالیت احتمال ابتلا به دیابت ۶۶/۱ درصد افزایش می‌یابد. برای متغیر داشتن پیاده‌روی و دوچرخه سواری (β = ۰/۱۳۶)، در کسانی که فعالیت ندارند احتمال ابتلا به دیابت ۱۴/۶ درصد بیشتر از افرادی است که فعالیت دارند. برای متغیر داشتن فعالیت تفریحی- ورزشی (β = ۰/۰۸۰)، احتمال ابتلا به دیابت در کسانی که فعالیت ندارند ۸/۳ درصد بیشتر از افرادی هستند که فعالیت دارند.

از تحلیل ROC به عنوان مقیاس اندازه‌گیری توانایی جداسازی یک مدل با بیشترین ناحیه که نشان دهنده توانایی پیش‌بینی بهتر برای مقایسه انجام مدل‌ها به کار می‌رود، استفاده می‌شود (۱۴).

مقدار سطح زیر منحنی ۰/۸۲۲ به دست آمد که می‌توان گفت دقت کلی آزمون در تشخیص دیابت به طور تقریبی ۸۲/۲ درصد می‌باشد و این به معنی خوب بودن در برآورد مدل رگرسیون می‌باشد.

لازم به ذکر است که دقت بالای ۹۰ درصد عالی، بین ۹۰-۸۰ درصد خوب، بین ۸۰-۷۰ درصد قابل قبول، بین ۷۰-۶۰ درصد ضعیف و چنان‌چه دقت آزمونی ۶۰-۵۰ درصد گردد، غیر قابل قبول می‌باشد.

در این مطالعه مقدار ضریب Kappa برابر با ۰/۳۱۲ به دست آمد که می‌توان گفت توافق متوسطی بین مقادیر مشاهده شده و پیش‌بینی شده وجود دارد.

آماره Hosmer-Lemeshow جهت تعیین این‌که مدل به طور کافی داده‌ها را توصیف می‌کند یا خیر به کار می‌رود که برای این تحقیق آماره مورد نظر HR = ۱۵/۹۸ (Hazard ratio) با مقدار P برابر با ۰/۰۴۳ بود که نشان دهنده کفایت معنی‌دار مدل برای توصیف از داده‌ها می‌باشد.

مدل لجستیکی که در این مطالعه بررسی شد با متغیرهایی که به ترتیب اهمیت معنی‌داری ذکر شده‌اند می‌توانند شانس ابتلا به دیابت را تغییر دهند؛ متغیر نسبت دور کمر به باسن (WHR)، مصرف داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، جنسیت، سطح تحصیلات، محیط دور کمر، پیاده‌روی و دوچرخه سواری، وزن، شاخص توده بدنی (BMI)، فشار خون سیستولی، سن، فشار خون دیاستولی، محیط دور باسن، تری‌گلیسرید، کلسترول، سطح LDL، استفاده از مواد مخدر، قد، سطح HDL و داشتن فعالیت کاری شدید ۱۰ دقیقه‌ای.

بحث

با توجه به اهمیت بیماری دیابت و بالا بودن درصد ابتلا به آن در جوامع مختلف، امر پیش‌بینی این بیماری بسیار حایز اهمیت می‌باشد و در این راستا انتخاب مدل آماری مناسبی که بتواند مشاهدات را به درستی برای داشتن دیابت پیش‌بینی کند و یا این‌که آیا فرد در آینده شانس ابتلا به دیابت را دارد، دارای اهمیت بسزایی است. در این مطالعه برای پیش‌بینی دیابت، مدل رگرسیون لجستیک اعمال شد و متغیرهای مؤثر و معنی‌دار تشخیص داده شدند و همچنین شاخص‌هایی مانند حساسیت، ویژگی، دقت و ناحیه زیر منحنی ROC برای سنجش میزان قدرت و دقت مدل محاسبه گردید.

در مطالعاتی که پیش‌تر در این زمینه انجام شده است؛ Wilson و همکاران در مطالعه‌ای برای پیش‌بینی بروز دیابت در افراد بالای ۵۰ سال، عوامل خطر شامل سن بالا، دور کمر بالا، سابقه فامیلی دیابت، اختلال تحمل قند خون ناشتا، تری‌گلیسرید بالا و مقدار HDL پایین را به عنوان متغیرهای پیش‌بینی کننده معرفی کردند (۵). در مطالعه دیگری که

Burke و همکاران انجام دادند؛ نژاد، چاقی، بیماری‌های قلبی، تری گلیسرید بالا و اختلال تحمل قند خون ناشتا و دو ساعته به عنوان عوامل خطر مؤثر در بروز دیابت معرفی شدند (۶). در مطالعه‌ای که توسط مرآتی و همکاران برای بررسی شیوع و عوامل مرتبط با ابتلا به دیابت بر جمعیت عمومی اصفهان انجام گرفت؛ شیوع دیابت ۶/۶ درصد برآورد شد که در زنان دو برابر بیشتر از مردان بود و همچنین ارتباط میان سن، جنس، شاخص توده بدنی و سابقه فامیلی ابتلا به دیابت با ابتلا به دیابت معنی‌دار بود (۷).

متغیرهای پیش‌بینی کننده و معنی‌داری که در این مطالعه به دست آمده‌اند، به ترتیب اولویت معنی‌داری و تعمیم به حداقل ۸۰ درصد جامعه، شامل متغیر نسبت دور کمر به باسن (WHR)، مصرف داروی کاهش فشار خون در دو هفته گذشته با تجویز پزشک، جنسیت، سطح تحصیلات (راهنمایی، ابتدایی و فوق دیپلم)، پیاده‌روی و دوچرخه سواری، محیط دور کمر، وزن، شاخص توده بدنی بود که مرآتی و همکاران نیز در نتایج خود اثر این عوامل را معنی‌دار گزارش کرده‌اند (۷). فشار خون سیستولی و سن با نتایج مطالعه Wilson و همکاران مشابهت دارد (۵). فشار خون دیاستولی، کلسترول و تری گلیسرید با نتایج مطالعه Burke و همکاران همخوانی می کند (۶). سطح LDL، استفاده از مواد مخدر، محیط دور باسن، قد، سطح HDL و داشتن فعالیت کاری شدید ۱۰ دقیقه‌ای نیز در مطالعات دیگر گزارش شده‌اند (۷-۵).

با مقایسه‌ای در مورد متغیرهای معنی‌دار در مطالعه حاضر و مطالعات پیشین درمی‌یابیم که از جمله عوامل خطر و جدی در ابتلای به دیابت به طور مشترک می‌توان به چاقی، دور کمر بالا، تری گلیسرید بالا، سن بالا، جنس و شاخص توده بدنی اشاره کرد که البته در این مطالعه عوامل مهم و هشدار دهنده دیگری از جمله سطح تحصیلات، نسبت دور کمر به باسن، فشار خون دیاستولی و سیستولی بالا و قد نیز در ابتلا به دیابت نقش داشته‌اند که می‌توان از جمله مشخصه تحقیق حاضر دانست.

Chae و همکاران در پیش‌بینی عوامل مؤثر بر فشار خون توسط مدل رگرسیون لجستیک، مقادیر دقت پیش‌بینی، حساسیت و ویژگی به ترتیب ۶۳/۸۴، ۶۴/۳۶ و ۶۳/۳۳ درصد به دست آمد (۸). در مطالعه‌ای توسط Su-juan در مورد کاربرد رگرسیون لجستیک در تحلیل ریسک اعتبار انجام شده، میزان دقت جداسازی مدل رگرسیون لجستیک در کل به ۹۹/۰۶ درصد رسید و نشان دهنده این است که دقت جداسازی مدل لجستیک به طور چشمگیری بالا می‌باشد (۹).

یافته‌های به دست آمده از مطالعه حاضر از نظر مقادیر حساسیت، ویژگی و دقت به ترتیب ۰/۷۶۴، ۰/۷۲۵ و ۰/۷۳۱ بود که می‌توان گفت این مدل پیش‌بینی مناسبی از مشاهداتی که به گروه‌های مربوطه تخصیص داده شده‌اند را اعمال کرده است. در واقع درصد افراد مبتلا به دیابت که توسط این مدل به درستی تشخیص داده شده‌اند، ۷۶/۴ درصد بود و نیز درصد افراد غیر دیابتی که توسط مدل به درستی پیش‌بینی شده‌اند، ۷۲/۵ درصد بوده است. همچنین با توجه ضریب توافق Kappa با مقدار ۰/۳۱۲، میزان توافق در حد متوسطی بود. در مقایسه با نتایجی که توسط محرابی و همکاران در مورد داده‌های مشابه با این مطالعه انجام گرفت، برای میزان حساسیت و ویژگی به ترتیب ۷۵ و ۸۲ درصد به دست آمد که نزدیک به نتایج این مطالعه و مطالعات ذکر شده قبلی می‌باشد. همچنین در این مطالعه ناحیه زیر منحنی ROC مقدار ۰/۸۲۲ به دست آمد که نشانگر مناسب بودن مدل است و در مقایسه با منحنی ROC که در مطالعه محرابی و همکاران انجام گرفت، این ناحیه ۰/۸۳۹ بود که به طور تقریبی یکسان می‌باشد (۳). بنابراین می‌توان گفت مدل رگرسیون لجستیک مدل مناسبی است که می‌تواند دقت و قدرت پیش‌بینی را به خوبی برآورد کرده، علاوه

بر آن نسبت خطر ابتلای به بیماری را به طور مستقیم تبیین کند.

نتیجه‌گیری

در این مطالعه مدل رگرسیون لجستیک برازش شده با توجه به ملاک‌های دقت و قدرت آن از جمله حساسیت، ویژگی و دقت پیش‌بینی، مدل مناسبی برای پیش‌بینی دیابت بود. همچنین نتیجه به دست آمده از مساحت زیر منحنی ROC نشان دهنده خوب بودن دقت کلی آزمون در تشخیص دیابت بود. ضریب توافق Kappa بین مقادیر مشاهده شده و مورد انتظار در حد متوسطی برآورد گردید. همچنین متغیرهای معنی‌داری که در این پژوهش گزارش شده‌اند و نیز متغیرهایی مانند سطح تحصیلات، نسبت دور کمر به باسن، فشار خون دیاستولی و سیستولی بالا و قد که در پژوهش‌های پیشین به صورت جدی به آن‌ها پرداخته نشده بود، در این مطالعه معنی‌دار گزارش شده‌اند که می‌توان با نظارت بر این موارد جدید علاوه بر سایر عوامل خطر در کنترل و پیش‌بینی ابتلا به دیابت گام بزرگی برداشت.

سپاسگزاری

پشتیبانی معاونت تحقیقات و فن‌آوری دانشگاه علوم پزشکی کرمان از پروژه حاضر شایان قدردانی می‌باشد.

References

1. Shakibi M, Atapoor J, Kalantari B, Namjoo B. Prevalence and risk factors for upper extremity soft tissue rheumatism in patients with diabetes in 2001 in Kerman. *Sci J Hamadan Nurs Midwifery Fac* 2003; 10(3): 20-6. [In Persian].
2. Esmailnasab N, Afkhamzadeh A, Ebrahimi A. Effective Factors on Diabetes Control in Sanandaj Diabetes Center. *Iran J Epidemiol* 2010; 6(1): 39-45. [In Persian].
3. Mehrabi Y, A Khadem-Maboudi A, Hadaegh F, Sarbakhsh P. Prediction of Diabetes Using Logic Regression. *Iran J Endocrinol Metab* 2010; 12(1): 16-24. [In Persian].
4. Sadehi M, Mehrabi Y, Kazemnejad A, Hadaegh F. Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Methods in Prediction of Metabolic Syndrome. *Iran J Endocrinol Metab* 2009; 11(6): 638-46. [In Persian].
5. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. *Arch Intern Med* 2007; 167(10): 1068-74.
6. Burke JP, Haffner SM, Gaskill SP, Williams KL, Stern MP. Reversion from type 2 diabetes to nondiabetic status. Influence of the 1997 American Diabetes Association criteria. *Diabetes Care* 1998; 21(8): 1266-70.
7. Meraci MR, Feizi A, Bager Nejad 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. *J Health Syst Res* 2012; 8(2): 193-203. [In Persian].
8. Chae YM, Ho SH, Cho KW, Lee DH, Ji SH. Data mining approach to policy analysis in a health insurance domain. *Int J Med Inform* 2001; 62(2-3): 103-11.
9. Su-juan P. An Application of Logistic Regression Model in Credit Risk Analysis. Mathematics in Practice and Theory 2006; 36(9): 129-37.

10. Antonogeorgos G, Panagiotakos DB, Priftis KN, 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. *Int J Pediatr* 2009; 2009: 952042.

11. Joanne Peng CY, Lida Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research* 2002; 96(1): 3-14.

12. 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(1-2): 97-111.

13. Najafipour H, Mirzazadeh A, Haghdoost A, Shadkam M, Afshari M, Moazenzadeh M, et al. Coronary Artery Disease Risk Factors in an Urban and Peri-urban Setting, Kerman, Southeastern Iran (KERCADR Study): Methodology and Preliminary Report. *Iran J Public Health* 2012; 41(9): 86-92.

14. Abdolmaleki P, Yarmohammadi M, Gity M. Comparison of logistic regression and neural network models in predicting the outcome of biopsy in breast cancer from MRI findings. *International Journal of Radiation Research* 2004; 1(4): 217-28.

Predicting Type Two Diabetes and Determination of Effectiveness of Risk Factors Applying Logistic Regression Model

Aram-Ahmaddi M., B.Sc.[1] Bahrampour A., Ph.D.[2*], Najafipour H., Ph.D.[3]

1. MSc Student of Biostatistics, Research Center for Modelling in Health Institute for Future Studies in Health and Epidemiology & Biostatistics Department, School of Health Medical Sciences, Kerman, Iran

2. Professor of Bostatistics, Physiology Research Center and Epidemiology & Biostatistics Department, School of Health, Kerman University of Medical Sciences, Kerman, Iran

3. Professor of Physiology, Physiology Research Center, Kerman University of Medical Sciences, Kerman, Iran

* Corresponding author; e-mail: abahrampour@yahoo.com

(Received: 11 April 2013 Accepted: 3 July 2013)

Abstract

Background & Aim: Diabetes is one of the chronic diseases with no curative treatment; also, it is the most common cause of amputation, blindness and chronic renal failure and the most important risk factor of heart diseases. Logistic regression is one of the statistical analysis models for predicting that can be used to find out the relationship between dependent and predictor independent variables and control of the confounding variables. The aim of this study was to determine the rate of effective variables on diabetes and estimation of the logistic regression model for predicting.

Methods: 5357 persons in Kerman city, Iran, were enrolled. Diabetes considered as the response variable and weight, height, body mass index (BMI), waist circumference, hip circumference, waist-to-hip ratio (WHR), age, gender, occupation, education, drugs, drug abuse, activities, systolic and diastolic blood pressure, and levels of total cholesterol, the high-density lipoprotein (HDL), the low-density lipoprotein (LDL), and triglycerides were considered as independent variables in the model. Measures of sensitivity, specificity, accuracy, Kappa measure of agreement and ROC (receiver operating characteristic) curve was applied for determining the power of test.

Results: The Sensitivity, specificity, accuracy rate, Kappa measure of agreement and area under the ROC curve for the model were 0.764, 0.725, 0.731, 0.312 and 0.822, respectively. The following variables were significant according to their impact and their importance, respectively: WHR ($\beta = 2.66$, OR=14.32), anti-hypertensive drug ($\beta =1.279$, OR= 3.59), sex ($\beta =0.707$, OR= 2.028), level of education, walking and cycling ($\beta = 0.136$, OR= 1.146), waist circumference ($\beta =0.12$, OR= 1.127), weight ($\beta = 0.112$, OR= 1.118), BMI ($\beta = 0.053$, OR= 1.054), systolic blood pressure ($\beta =0.052$, OR= 1.054), age ($\beta =0.046$, OR= 1.047), diastolic blood pressure ($\beta =0.043$, OR= 1.044), total cholesterol ($\beta = 0.003$, OR= 1.003), triglycerides ($\beta =0.01$, OR= 1.011), LDL ($\beta = 0.001$, OR= 1.001), hip circumference ($\beta = - 0.025$, OR= 1.025), height ($\beta = -0.071$, OR= 0.932), HDL ($\beta = -0.078$, OR= 0.925), an intense 10-minute work activities ($\beta = -0.507$, OR=0.602).

Conclusion: According to the criteria of accuracy and power of prediction, and considering ROC curve value (0.822) which could perform test accuracy as well for the diagnosis of diabetes, the logistic regression model was an appropriate model for the prediction of diabetes in this study.

Keywords: Logistic regression, Diabetes, Discrimination, Calibration

Journal of Kerman University of Medical Sciences, 2014; 21(2): 103-113