مجله دیابت و لیپید ایران. دو ماهنامه مهر- آبان ۱۳۹۱؛ دوره ۱۲ (شماره ۱) ۱۰-۱

آمار برای پزشکان: همبستگی و رگرسیون

محبوبه پارسائیان[۱]، حمیده موسی‌پور[۱*]، فرهاد حسین‌پناه[۲]

به عنوان یک پزشک پرمشغله که هر روز بیماران زیادی را در مطب شلوغ‌تان می‌بینید، سوالات زیادی نیز برایتان پیش می‌آید، اما گهگاه فرصت می‌کنید سری به ادبیات پزشکی تخصصی خود بزنید. شما به خوبی می‌دانید که توانایی استفاده از مطالب روا و به روز، می‌تواند نقش مهمی در طبابتتان در دنیای پیچیدهٔ فعلی بازی کند ولی به خوبی نیز می‌دانید که به کار بردن نتایج مقالات نیازمند درک مفاهیم آماری است.

امروزه آشنایی با تفکر آماری و درک اصول و مفاهیم پایه‌ای آن نه تنها برای داشتن یک رویکرد علمی در حیطهٔ تخصصی خود لازم است، بلکه تفکر آماری (که به ما اجازه می‌دهد از تجربهٔ خود با کمترین خطا بیاموزیم)، جزئی تفکیک ناپذیر از طبابت علمی است و پزشکان بیش از پیش برای فهم و نقد ادبیات پزشکی پیش رویشان به درک مفاهیم پایه و کاربردی آمار پزشکی نیازمندند.

«پزشکی مبتنی برشواهد» نهضت رایج نوپایی که امروزه می‌رود تا چهرهٔ غالب و رایج گفتمان پزشکی دنیا باشد، رویکرد عینی آن بیش از پیش پزشکان را نیازمند درک تفکر آماری و مفاهیم و اصول پایهٔ آن خواهد کرد تا آنجا که یکی از تعاریف Evidence Based Medicine (EBM) را مجموعهٔ علوم بالینی، متدولوژی و آمار بیان کرده‌اند.

ممکن است به خوبی ندانید که چگونه باید اطلاعات آماری ذکر شده در مقالات را تفسیر کنید. ما نیز با شما موافقیم که آمار مقوله‌ای بسیار فرار بوده و یادگیری آن برای افراد غیرمتخصص در این رشته مشکل است، به‌ویژه اگر قرار باشد بر فرمول‌ها و مفاهیم محض آن تأکید شود تا کاربردهای بالینی مفاهیم آمارپزشکی. مقاله حاضر جهت آشنایی شما با تفکر آماری برای استفاده در استدلال‌های بالینی و افزایش توانمندیتان در فهم و ارزیابی نقادانه مقالات پزشکی طراحی شده است. در این مقاله دو مفهوم آماری همبستگی و رگرسیون که کاربرد فراوان در ادبیات پزشکی دارند، انتخاب شده و در قالب مثال‌های بالینی ملموس در حوزه دیابت و لیپید توضیح داده شده است.

امید که این باور سنتی را که «آمار اساساً چیزی نیست که پزشکان بفهمند یا در طبابت بالینی روزانه بدردشان بخورد»، از مجموعه باورهایمان کنار بگذاریم.

۱- مرکز تحقیقات غدد/ پژوهشگاه علوم غدد و متابولیسم، دانشگاه علوم پزشکی تهران
۲- مرکز تحقیقات پیشگیری و درمان چاقی، پژوهشکده علوم غدد درون‌ریز و متابولیسم، دانشگاه علوم پزشکی شهید بهشتی

*نشانی: تهران، خیابان کارگر شمالی، بیمارستان شریعتی، طبقه پنجم، مرکز تحقیقات غدد، پژوهشگاه علوم غدد و متابولیسم دانشگاه علوم پزشکی تهران، کدپستی: ۱۴۱۱۴۱۳۱۳۷، تلفن: ۸-۸۸۲۲۰۰۳۷-۰۲۱، نمابر: ۸۸۲۲۰۰۵۲-۰۲۱، پست الکترونیک: dr_moosapour@yahoo.com

تاریخ دریافت: ۱۳۹۱/۰۶/۱۵ تاریخ درخواست اصلاح: ۱۳۹۱/۰۶/۲۰ تاریخ پذیرش: ۱۳۹۱/۰۶/۲۸

یکی از موضوعات مورد علاقه پزشکان درک و پیدا کردن ارتباط بین عوامل و متغیرهای مختلف است. به عنوان مثال اندوکرینولوژیست‌ها علاقه‌مندند که بدانند آیا 2-h PG[1] با عواملی چون FPG[2]، سطح کلسترول توتال و تری گلیسرید خون، فشارخون و... ارتباط دارد؟ چه اینترلوکین‌هایی با آترواسکلروزیس و نفروپاتی در دیابتی‌ها مرتبط هستند؟ و یا اینکه این ارتباط‌ها چقدر قوی است؟

البته ما معمولاً می‌خواهیم که پا را فراتر از این گذاشته و بتوانیم روابط علیتی را بین متغیرها پیدا کنیم ویا بتوانیم با داشتن یک سری اطلاعات، رخدادهایی را قبل از وقوع پیش‌بینی کنیم و یا از سیستم‌های نمره‌دهی[3] یا قانون‌های پیش‌بینی کننده[4] در تشخیص یا تعیین پیش آگهی بیماران استفاده کنیم.

به عنوان مثال اندوکرینولوژیست‌ها علاقه‌مندند که بدانند به ازای هر سال افزایش در سن 2-h PG چقدر افزایش می‌یابد؟ آیا ابتلا به بیماری عروق کرونر با عوامل خطر متابولیکی چون دیس لیپیدمی و... به طور مستقل مرتبط است؟

قدرت ارتباط بین پدیده‌ها یا متغیرها را با همبستگی نشان می‌دهیم. رگرسیون نیز، تکنیک آماری دیگری است که در بررسی ارتباط بین متغیرها به خصوص برای پیش‌بینی کردن به کار می‌رود. رگرسیون و همبستگی دو روی یک سکه‌اند. در واقع در چنین تحلیل‌هایی ما برای هر فرد یک جفت متغیر داریم که ارتباط[5] آنها را با هم بررسی می‌کنیم.

بیایید یک مثال بالینی ملموس مثلا تعریف دیابت را در نظر بگیریم. همان طور که می‌دانید ADA(1997) (انجمن دیابت آمریکا) FPS بالاتر از ۱۲۶ را به عنوان معیار تشخیصی دیابت معرفی کرده است. در حالی که WHO (1999) توصیه به انجام OGTT نموده است و معیار تشخیص دیابت را FPG بالاتر از ۱۲۶ و یا ۲-hPG بالاتر از mg/dl 200 بیان کرده است.

Bando Y و همکاران در مطالعه‌ای مقطعی با استفاده از داده‌های بیمارانی که جهت تشخیص دیابت تحت OGTT قرارگرفته بودند، به بررسی ارتباط این دو مقدار پرداخته و بیان می‌کنند که این تعاریف حاکی از آن است که FPG برابر با ۱۲۶ معادل 2-h PG برابر با ۲۰۰ تخمین زده شده است. آنها در ادامه به این مطلب اشاره می‌کنند که اگر ارتباط FPG و 2-hPG خود تحت تاثیر عوامل دیگری باشد، معادل بودن این دو مقدارتحت تاثیر سطوح مختلف آن عوامل تغییر خواهد کرد. از سوی دیگر همان طور که می‌دانید شیوع IPH[6] در افراد مسن و خانم‌ها بیشتر است. این مطالعه بررسی کرده است که عوامل دیگری غیر از FPG تا چه میزان بر2-h PG موثر هستند. ما در ادامه، توضیح مباحث همبستگی و رگرسیون را در قالب مرور نتایج این مطالعه با هم دنبال خواهیم کرد.

همبستگی

دو کمیتسن و 2-h PG همبستگی نسبتاً قوی با هم نشان می‌دهند. 2-hPG با سطح کلسترول توتال خون نیز ارتباط ضعیفی دارد. معنی قدرت ارتباط بین دو متغیر چیست؟ ارتباط سن و 2-hPG قوی است زیرا بیمارانی که سن بالایی دارند 2-hPG بالایی نیز دارند و یا بیمارانی که سن متوسطی دارند مقدار متوسطی از 2-hPG دارند. اما در مقابل ارتباط سطح کلسترول توتال خون و 2-h PG ضعیف است، چراکه افرادی که سطح کلسترول توتال خون پایینی دارند با احتمال یکسانی مقادیر بالا یا پایینی از 2-hPG دارند.

برای اینکه درک شهودی از قدرت ارتباط پیدا کنیم، نموداری رسم می‌کنیم که نمودار پراکنش نام دارد. این نمودارمقدار متغیر اول و دوم را برای هر فرد نشان می‌دهد. اگر بتوانیم یک رابطه علیتی بین دو متغیر تصور کنیم، متغیری که تاثیر گذار باشد متغیر مستقل (کمکی یا پیش گویی کننده) و متغیر تاثیر پذیر را متغیر وابسته (پاسخ یا هدف) می‌نامیم. مقادیر متغیر مستقل را روی محور X و

1.2-h Post- 75-g Oral Glucose Load Glycemia
2.Fasting Plasma Glucose
3.Symptom Scores
4. Prediction Rules

۵. در بررسی رابطه بین متغیرهای کیفی(اسمی یا رتبه ای) از آزمون کای دو استفاده می‌شود. این آزمون وجود هرنوع ارتباطی را بین دو متغیر بررسی می‌کند. در تحلیل کای دو حتی به متغیرهای رتبه‌ای نیز مانند اسمی نگاه می‌شود و آزمون ماهیت رتبه‌ای آن را لحاظ نمی‌کند.

6. Isolated Post-challenging Hyperglycemia

مقادیر متغیر وابسته را روی محور Y نشان می‌دهیم. چنانچه رابطه‌ای علیتی قابل تصور نباشد، تفاوتی بین دو محور نخواهد بود.

شکل ۱ نمودار پراکنش مقادیر سن و 2-hPG در افرادی که FPG برابر با ۱۲۶ (mg/dl) دارند را نشان می‌دهد. در واقع هر نقطه در این نمودار نشان دهنده یک فرد است که دو جور اطلاعات به ما می‌دهد: سن و 2-hPG. دقت کنید که در این مورد هر دو متغیر کمی پیوسته هستند (یعنی بین مقادیر آن فاصله‌ای وجود ندارد). همان طور که مشاهده می‌کنید، نمودار در کل به ما نشان می‌دهد که بیمارانی که سن بالایی دارند گرایش بیشتری نشان می‌دهند که 2-hPG بالاتری داشته باشند و بیمارانی که سن کمتری دارند گرایش بیشتری نشان می‌دهند که 2-hPG کمتری داشته باشند. دقت کنید که این گرایش به صورت کلی می‌باشد. با این وجود شما می‌توانید موارد استثنایی را نیز پیدا کنید که هر چند سن بیشتری دارند، اما 2-hPG بالایی ندارند.

این داده‌ها نمایانگر ارتباط نسبتاً قوی بین سن و2-hPG می‌باشد. اگر بخواهیم این قدرت ارتباط را به صورت عددی نشان دهیم از ضریب همبستگی یا r[1] استفاده می‌کنیم. ضریب همبستگی در واقع برآوردی از میزان نزدیک بودن نقاط نمودار پراکنش در اطراف یک خط راست می‌باشد. ضریب همبستگی می‌تواند از مقدار ۱- (قوی‌ترین ارتباط معکوس ممکن که در آن بیمارانی با بیشترین نمره در یک کمیت، کمترین نمره را در کمیت دیگر دارند) تا ۱+ (قوی‌ترین ارتباط مستقیم ممکن که در آن بیمارانی با بیشترین نمره در یک کمیت، بیشترین نمره را در کمیت دیگر دارند). چنین نمودارهای پراکنشی داده‌ها را به صورت دانه‌های تسبیح در امتداد یک خط تصویر می‌کنند. ضریب همبستگی صفرنیز بیان می‌کند که هیچ گونه ارتباط خطی بین دو متغیر وجود ندارد یعنی بازه مقادیر کمیت دوم در بیمارانی که نمره بالایی از کمیت اول دارند، مشابه بیمارانی است که نمره پایینی از کمیت اول دارند و هیچ گرایش منظمی بین دو متغیر وجود ندارد. نمودار پراکنش داده‌هایی که ضریب همبستگی صفر داشته باشند، شبیه یک آسمان پر ستاره است (بدون هیچ گونه تجمعی در داده‌ها). در بررسی ارتباط بین متغیرها در حیطه پزشکی معمولاً مقادیر ضریب همبستگی بین ۰ و ۱± می‌باشدو روابط بین متغیرها خطی کامل نیستند. اما در یک نگاه روند خطی کلی را می‌توانیم در نمودار پراکنش ببینیم (شکل۲).

ضریب همبستگی یک خط مستقیم بین متغیرها فرض می‌کند، درحالی که ممکن است ارتباط بین متغیرها به شکل یک خط مستقیم نباشد. به عنوان مثال مقادیر دو متغیر ممکن است با هم افزایش پیدا کنند، اما یکی از آنها آرام‌تر از دیگری در مقادیر پایین و سریع‌تر از دیگری در مقادیر بالا افزایش یابد. حتی اگر یک ارتباط قوی نیز وجود داشته باشد این ارتباط خطی نیست و استفاده از ضریب همبستگی می‌تواند گمراه کننده باشد.[2]

در رابطه همبستگی بین 2-h PG و سن r=۰/۹۴ (p-۰/۰۰۱ value>) و در رابطه همبستگی بین 2-h PG و FPG r=۰/۶۴ (p-value<۰/۰۰۱) است. مشابه اینجا معمولاً همراه ضریب همبستگی یک p-value می‌بینید. این p-value درواقع مربوط به آزمونی است که فرض صفر آن بیان می‌کند که همبستگی بین دو متغیر وجود ندارد. p-value

---

۱- لازم به ذکر است که نام دیگرr، ضریب همبستگی پیرسون است و در مواردی استفاده می‌شود که دو متغیر پیوسته و دارای توزیع نسبتاً نرمال باشند. نوع دیگری از ضریب همبستگی به نام ضریب همبستگی اسپیرمن نیز داریم که به عنوان یک ضریب همبستگی غیر پارامتری در مواردی که حجم نمونه کوچک باشد، استفاده زیادی دارد. ضریب همبستگی اسپیرمن همچنین زمانی مناسب است که بخواهیم ارتباط متغیرهای رتبه‌ای مثل تحصیلات و سطح اقتصادی اجتماعی و یا متغیرهای پیوسته‌ای که پس از گروه‌بندی به مقادیر رتبه‌ای تبدیل شده‌اند را بررسی کنیم. چرا که در این صورت ضریب همبستگی اسپیرمن تحت تاثیر چولگی یا غیرنرمال بودن توزیع متغیری مثل سن (در مثال بالا) قرار نمی‌گیرد. در واقع ضریب همبستگی اسپیرمن یک رابطه همنوا را بین دو متغیر بررسی می‌کند. یعنی چنانچه مقدار یک متغیر افزایش یابد، مقدار متغیر دیگر افزایش یا کاهش می‌یابد. به عنوان مثال وجود رابطه همنوا مثبت بین تحصیلات و سطح اقتصاد اجتماعی به این معنی است که انتظار داریم هرچه تحصیلات بالاتر باشد، سطح اقتصادی اجتماعی نیز بالاتر باشد.

۲- یک مثال برای مواردی که دنبال رگرسیون غیرخطی می‌رویم، بررسی رابطه سن با بعضی متغیرهای زیستی است چراکه رابطه سن در سنین بالا با بعضی متغیرهای زیستی رابطه توان دوم دارد. درحالی که همین رابطه در سنین پایین‌تر خطی است.

بیان می‌کند، اگر همبستگی واقعی صفر باشد، چقدر احتمال دارد که رابطه مشاهده شده یا رابطه‌ای قوی‌تر از آن، شانسی اتفاق افتاده باشد. هرچه میزان p-value کوچک‌تر باشد، احتمال کمتری وجود دارد که شانس علت همبستگی مشاهده شده بین دو متغیر باشد.

البته توجه داشته باشید که p-value نه تنها به قدرت رابطه، بلکه به حجم نمونه نیز بستگی دارد. یک رابطه ضعیف می‌تواند بواسطه یک حجم نمونه به اندازه کافی بزرگ p-value کوچکی بدهد. برای مثال با یک حجم نمونه بزرگ ۵۰۰، یک ضریب همبستگی کوچک ۰/۱ نیز می‌تواند به حد آستانه معنی‌داری (P= ۰/۰۵) برسد.

همان طور که در ارزیابی اثر درمان، اندازه اثر و فاصله اطمینان آن اطلاعات بیشتری نسبت به p-value در اختیار ما قرار می‌دهند، این موضوع در مورد ضریب همبستگی و فاصله اطمینان آن نیز صادق است. به عنوان مثال مطالعه‌ای گزارش می‌کند که r به دست آمده و فاصله اطمینان ۹۵٪ مربوطه به ترتیب برابر ۰/۵۰ و ۰/۵۵-۰/۴۵است.

یک اشتباه رایج در تفسیر همبستگی، استنباط رابطه علیتی از همبستگی است. در حالی که وجود همبستگی الزاماً به معنای وجود رابطه علیتی نیست و این تحلیل بیشتر در مرحله تولید یک فرضیه علیتی به کار می‌رود تا آزمون چنین فرضیه‌ای. به عنوان مثال، FPG و 2-hPG باهم همبستگی دارند اما رابطه علیتی بین این دو وجود ندارد. یعنی هردو در ارتباط باهم تغییر می‌کنند اما یکی علت دیگری نیست. در واقع برای استنباط یک رابطه علیتی بین دو متغیر غیر از همبستگی قوی به شرایط دیگری نیز نیازمندیم مثل مقدم بودن علت بر معلول و... (اصول Hill در اپیدمیولوژی). پس به عبارت دیگر برای استنباط علیتی، همبستگی لازم است. اما وجود همبستگی دال بر وجود رابطه علیتی نیست.

رگرسیون

اغلب پیش‌بینی کردن برای ما مهم است و می‌خواهیم بدانیم کدام یک از افراد بیمار خواهند شد و کدام یک نه. کدام بیماران بهبود می‌یابند و کدام نه و... . مثال تحلیل رگرسیون در پاسخ به چنین موضوعاتی مفید است. رگرسیون را با استفاده از داده‌های مثال قبل دنبال می‌کنیم.

مدل سازی و پیش‌بینی کردن

بیایید سوالی را که در بالا داشتیم را به شکل دیگری بیان کنیم. آیا FPG می‌تواند 2-hPG را پیش‌بینی کند؟ یا به چه میزان 2-hPG تحت تاثیر عواملی چون سن، جنسیت و... تعیین می‌شود؟ در چنین سوالاتی یک متغیر پاسخ مثل 2-hPG داریم که میزان آن توسط متغیرهای دیگری چون FPG، سن وجنسیت پیش‌بینی می‌شود که آنها را متغیرهای پیش‌بینی کننده می‌نامیم.

همان طور که می‌دانیم دامنه مقادیر 2-hPG در بیماران خیلی گسترده است. اگر هیچ اطلاعی راجع به یک بیمار خاص نداشته باشیم، بهترین حدسی که می‌توانیم در مورد 2-hPG او بزنیم، میانگین 2-hPG در همه بیماران است. اما قبول دارید که برای بیماران زیادی چنین حدسی مطابق با مقدار واقعی نیست. همان طور که بیان شد، ارتباط نسبتاً قوی بین FPG و2-hPG وجود دارد. یعنی مقداری از تغییرات 2-hPG با FPG مرتبط است. ما می‌توانیم مدلی بسازیم که به وسیله آن 2-hPG را براساس FPG پیش‌بینی کنیم. از آنجا که تنها یک متغیر مستقل وجود دارد، این معادله را رگرسیون ساده یا تک متغیره می‌نامیم.

در معادلات رگرسیونی به طور کلی به متغیر پیش‌بینی کننده X و به متغیر هدفY می‌گوییم. معادله، یک خط راست را بین X (مثلاً FPG) و Y (مثلاً 2-hPG) با دو پارامتر a و b فرض می‌کند. شکل کلی معادله به صورت زیر می‌باشد:

$$Y=a+bX$$

عرض از مبدا (a)، نقطه‌ای است که خط محور Y را قطع می‌کند و شیب (b) متناسب با زاویه‌ای است که خط با محورX ها می‌سازد. عرض از مبدا پیش‌بینی Y به ازای X=۰ است و شیب به این معنی است که به ازای هر واحد افزایش X چه مقدار Y زیاد می‌شود. شکل ۳ این مفاهیم را در یک معادله رگرسیونی فرضی Y=۲/۴+۱/۳۱X نشان می‌دهد.

مقدار پاسخی که از معادله بالا به دست می‌آید، میزان متغیر پاسخ تحت پیش‌بینی مدل است (مقدارY متناسب با X مورد نظر روی خط رگرسیون) که البته با مقدار واقعی پاسخ (مختصات خود نقطه) متفاوت خواهد بود. به عنوان مثال در شکل ۴ سه نقطه با X مساوی ۵۰ داریم که فقط یکی از آنها روی خط راست است و تنها برای این نقطه مقدار واقعی با مقدار پیش‌بینی شده (Y=۹۲۵) مساوی است و در دو نقطه دیگر بین مقدار واقعی (Y=۸۵۰ و Y=۹۵۰) و مقدار پیش‌بینی شده اختلاف وجود دارد. به این تفاوت، e_i یا باقیمانده می‌گویند. یعنی میزانی از تغییرات که مدل نتوانسته است آن را پیش‌بینی کند. کوچک بودن e_i به معنای پیش‌بینی بهتر مدل است.

ممکن است علاقه‌مند باشید که نرم‌افزار چگونه یک خط مستقیم به داده‌ها برازش می‌کند، به طوری که نقطه‌ها به بهترین شکل حول آن خط پراکنده باشند و معادله خط بهترین برآورد Y به ازای هر مقدار X را بدهد. همان طور که شکل ۵ نشان می‌دهد، هر چه در مجموع طول خط‌ها (فاصله عمودی هر نقطه تا خط برازش شده) کوچک‌تر باشد، برازش بهتری انجام شده است. در واقع نرم‌افزار a و b را طوری انتخاب می‌کند که مجموع مجذورات این فواصل عمودی کمترین مقدار شود[1].

شاخصی که به صورت کمی به ما نشان دهد که با برازش مدل چقدر از تغییرات متغیر پاسخ را می‌توان پیش‌بینی کرد، R^2 یا "ضریب تعیین[2]" است که در حالت رگرسیون تک متغیره مجذور همان r (ضریب همبستگی) می‌باشد و R^2-۱ درصدی از تغییرات متغیر پاسخ است که مدل نتوانسته پیش‌بینی کند. پس هرچقدر مقدار r یا قدرت رابطه بیشتر باشد، تخمین مدل به مقدار واقعی نزدیک‌تر است و بالعکس.

در بررسی همبستگی 2-hPG و FPG ضریب همبستگی مساوی ۰/۶۴ و p-value<۰/۰۰۱ است و احتمال آنکه رابطه شانسی باشد کم است. بنابراین اینگونه نتیجه می‌گیریم که FPG رابطه معنی‌داری با 2-hPG دارد. در این مثال R^2 مساوی ۴۱/۶ ٪ است. یعنی ۴۱/۶ درصد تغییرات 2-hPG با FPG پیش‌بینی می‌شود. معادله رگرسیونی نیز بدین شکل است:

$$Y= - ۸۴/۵ +۲/۲۱۲X$$

طبق این معادله رگرسیونی برای FPG مساوی با ۱۲۶مقدار پیش‌بینی 2-hPG برابر ۱۹۴ است. فاصله اطمینان ۹۵٪ به علت بزرگ بودن انحراف معیار 2-hPG در این نقطه (۳۵/۱) دامنه گسترده‌ای از مقادیر 2-hPG را شامل می‌شود (۲۶۴/۲-۱۲۳/۸). این پراکندگی حاکی از آنست که مقدار 2-hPG تحت تاثیر عوامل دیگری غیر از FPG نیز قرار دارد.

در این مطالعه هشت متغیر سن، جنس، فشار خون سیستولیک و دیاستولیک، سطح کلسترول توتال و تری‌گلیسرید سرم، BMI و FPG در تحلیل رگرسیونی ساده یا تک متغیره رابطه معنی‌داری با 2-hPG نشان دادند. سوالی که در این مرحله پیش می‌آید این است که آیا ارتباط تمام این متغیرها و نقش‌شان در تبیین 2-hPG مستقل است یا اینکه همبستگی و پیش‌بینی کنندگی بعضی از آنها وابسته به دیگری است؟ آیا هر یک از معادلات تک متغیره سهم مستقلی در توضیح تغییرات دارند یا خیر؟ برای پاسخ به این سوال، هر هشت متغیر را به طور هم‌زمان وارد مدل می‌کنیم. مدل ریاضی که برای در نظر گرفتن اثر هم‌زمان متغیرهای پیش‌بینی کننده پاسخ (2-h PG) بکار می‌رود را رگرسیون چندگانه یا چندمتغیره می‌نامیم. نتایج بیان می‌کند زمانی که BMI وارد مدل می شود، متغیر جنس ارتباط معنی دار خود را از دست می دهد که این دال بر وابسته بودن ارتباط جنس و2-h PG به BMI می‌باشد. در واقع اگر ما دو متغیر جنس و BMI را به عنوان متغیر مستقل در نظر بگیریم، هر دو ارتباط معنی‌داری با 2-h PG دارند. اما از آنجا که یک همبستگی قوی بین این دو متغیر وجود دارد، غیرمحتمل است که بتوانند سهم مستقلی داشته باشند. به طور مشابه دو متغیر دیگر، فشار خون دیاستولیک و سطح کلسترول توتال خون در تحلیل چند متغیره کنار گذاشته شده‌اند. هرچند این سه متغیر در تحلیل تک متغیره معنی‌دار بودند، در تحلیل چند متغیره معنی‌دار نمی‌باشند. در این مثال از پنج متغیر

۱- مطالب بالا به روش حداقل مربعات خطا (Least Square Error) اشاره دارد که یکی از معمول‌ترین روش‌های برآورد پارامترهای مدل‌های رگرسیونی می‌باشد.

2.Coefficient of determination

باقیمانده که سهم مستقلی در پیش‌بینی 2-h PG داشتند استفاده شد.

با استفاده از برازش معادله رگرسیون می‌توان نسبت تغییرات متغیر پاسخ (2-h PG) را که مربوط به هریک از متغیرهای مستقل و یا مربوط به کل متغیرهای مستقل باشد را محاسبه کرد. در این مثال، FPG اولین متغیر به کار رفته در مدل است و ۴۱/۶٪ تغییرات را تبیین می‌کند. وقتی متغیرهای بعدی به مدل اضافه شده است، ۱/۸٪ دیگر نیز به قدرت تبیین مدل اضافه شده و در مجموع مدل ۴۳/۴٪ تغییرات 2-h PG را بیان می‌کند. اطلاع از اینکه چند متغیر به طور مستقل قسمتی از تغییرات متغیر دیگر را توضیح می‌دهند، اطلاعی در مورد قدرت پیش‌بینی مدل رگرسیون چندگانه نمی‌دهد و برای ارزیابی قدرت پیش‌بینی مدل باید به R^2 مدل چندگانه توجه کنیم. به عنوان مثال هرچند متغیرهای FPG، سن، BMI، فشار خون سیستولیک و تری گلیسرید سرم به شکل معنی‌داری با متغیر 2-h PG در ارتباط هستند، اما در بررسی پیش‌بینی این مدل با استفاده از R^2 مشاهده می‌کنیم که تنها ۴۳/۴٪ تغییرات 2-h PG توسط این متغیرها توضیح داده می‌شود. بنابراین به نظر می‌رسد عوامل موثر دیگری نیز وجود دارند که باید در نظر گرفته شوند.

سوال بعدی که پیش می‌آید این است که r یا R^2 چقدر باشد تا قدرت ارتباط یا قدرت پیش‌بینی مدل را کافی بدانیم؟ پاسخ چنین سوالی به کاربرد آن و زمینه بالینی مربوطه بستگی دارد. به عنوان مثال اگر دنبال این باشیم که تستی را (که معمولاً ساده‌تر اندازه‌گیری می‌شود) جایگزین تست دیگری کنیم، r=۰/۵ کفایت نمی‌کند و یک ضریب همبستگی ۰/۸ یا بیشتر جایگزین مناسبی خواهد بود. اگر میزان ضریب همبستگی پایین باشد، خطر زیادی وجود دارد که بیماری با یک نمره بالا از تست اول نمره متوسط یا کمی از تست دوم داشته باشد و بالعکس. پس صرفاً یک معنی‌داری آماری دال بر کاربرد و اهمیت بالینی نیست. به عنوان مثال با توجه به اهمیت پروگنوستیک IPH در مباحث دیابت، ممکن است r=۰/۶۴ برای توجیه جایگزینی FPG برای 2-h PG مناسب نباشد.

همان طور که در مدل‌های بالا مشاهده کردید متغیر مستقل در یک معادله رگرسیونی می‌تواند یک متغیر دوحالته مثل جنسیت یا یک متغیر پیوسته مثل FPG باشد. در این مثال‌ها متغیر پاسخ پیوسته بود و یک خط مستقیم بین دو متغیر فرض شده بود. اما با تحلیل رگرسیونی می‌توان ارتباطات غیر خطی (نمایی، چند جمله‌ای، لجستیک و...) را نیز بین دو متغیر بررسی کرد. در مثال بعدی متغیر پاسخ دوحالتی است و از آنجا که این مدل‌ها برمبنای معادلات لگاریتمی می‌باشند، به آنها مدل‌های "رگرسیون لجستیک " می‌گویند.

پیش‌بینی ابتلا به بیماری عروق کرونر بر اساس عوامل خطر مربوطه

در این قسمت با استفاده از نتایج مطالعه Western Collaborative Group Study، که به پیش‌بینی ابتلا به بیماری عروق کرونر براساس عوامل خطر مربوطه پرداخته، به بیان رگرسیون لجستیک می‌پردازیم. در این مطالعه متغیر پاسخ، ابتلا به بیماری عروق کرونر (یک متغیر دوحالته)، و متغیرهای مستقل شامل تیپ شخصیتی، قد، وزن، فشار خون سیستولیک و دیاستولیک، کلسترول، تعداد سیگار مصرفی و سن می‌باشند. جدول ۱ شامل قسمتی از نتایج مطالعه فوق بوده که تحلیل رگرسیون لجستیک تک متغیره عوامل خطر مورد بررسی را نشان می‌دهد.

جدول ۱- نتایج برازش مدل‌های رگرسیون لجستیک تک متغیره در داده‌های مطالعه Western Collaborative Group

متغیر	برآورد ضریب رگرسیونی (b_j)	خطای معیار برآورد رگرسیونی	p-value	OR
تیپ شخصیتی	۰/۸۶۴	۰/۱۴۰	<۰/۰۰۱	۲/۳۷۳
قد	۰/۰۲۷	۰/۰۲۵	۰/۲۹	۱/۰۲۸
وزن	۰/۰۱۰	۰/۰۰۳	<۰/۰۰۱	۱/۱۳۰
فشار خون سیستولیک	۰/۰۲۷	۰/۰۰۴	<۰/۰۰۱	۱/۰۲۷
فشار خون دیاستولیک	۰/۰۳۴	۰/۰۰۶	<۰/۰۰۱	۱/۰۳۴
کلسترول	۰/۰۱۲	۰/۰۰۱	<۰/۰۰۱	۱/۰۱۳
تعداد سیگار مصرفی	۰/۰۲۳	۰/۰۰۴	<۰/۰۰۱	۱/۰۲۳
سن	۰/۰۷۴	۰/۰۱۱	<۰/۰۰۱	۱/۰۷۷

به جز متغیر قد (p-value=۰/۲۹) سایر متغیرها ارتباط معنی‌داری با ابتلا به بیماری عروق کرونر نشان می‌دهند. همان طور که در رگرسیون خطی نیز گفته شد، سوالی که در اینجا مطرح می‌شود این است که آیا همه این عوامل خطر به طور مستقل با ابتلا به بیماری عروق کرونر مرتبط هستند یا اینکه بعضی از آنها ارتباط مستقل با ابتلا به بیماری ندارند اما به علت ارتباط با سایر متغیرها چنین ارتباط کاذبی را نشان می‌دهند. جهت پاسخ به چنین سوالی و برای در نظر گرفتن اثر توام این متغیرها از یک رگرسیون لجستیک چندگانه استفاده شده است (جدول ۲).

جدول ۲- نتایج برازش مدل‌های رگرسیون لجستیک چندگانه در داده‌های مطالعه Western Collaborative Group

متغیر	برآورد ضریب رگرسیونی (b_j)	خطای معیار برآورد رگرسیونی	p-value	OR
تیپ شخصیتی	۰/۶۵۴	۰/۱۴۵	<۰/۰۰۱	۱/۹۲۴
وزن	۰/۰۰۹	۰/۰۰۳	۰/۰۰۶	۱/۰۰۹
فشار خون سیستولیک	۰/۰۱۸	۰/۰۰۶	۰/۰۰۶	۱/۰۱۸
فشار خون دیاستولیک	-۰/۰۰۱	۰/۰۱۷	۰/۹۵	۰/۹۹۹
کلسترول	۰/۰۱۱	۰/۰۰۲	<۰/۰۰۱	۱/۰۱۱
تعداد سیگار مصرفی	۰/۰۲۱	۰/۰۰۴	<۰/۰۰۱	۱/۰۲۱
سن (سال)	۰/۰۶۴	۰/۰۱۲	<۰/۰۰۱	۱/۰۶۷

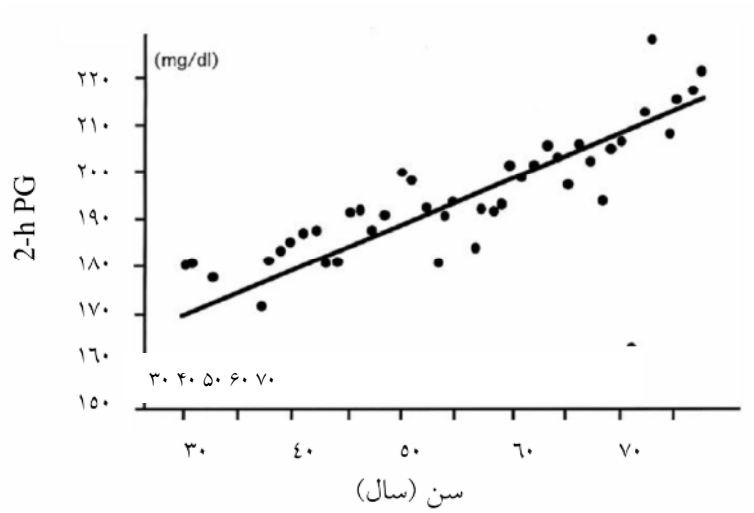
به جز فشارخون دیاستولیک (p-value=۰/۹۵) ارتباط تمامی متغیرها در ابتلا به بیماری عروق کرونر با استفاده از رگرسیون لجستیک چندگانه معنی‌دار می‌باشند. به طور مشابه در این مثال نیز فشارخون دیاستولیک با سایر عوامل خطر مرتبط بوده و ارتباط مستقلی با متغیر پاسخ ندارد.

معنی OR ها در جدول چیست؟ همان طور که در رگرسیون خطی افزایش یک واحد در متغیر پیش‌بینی کننده متغیر پاسخ را به اندازه b افزایش می‌دهد، در رگرسیون لجستیک یک واحد افزایش در متغیر پیش‌بینی کننده شانس رخداد متغیر پاسخ را به اندازه OR افزایش می‌دهد. به عنوان مثال OR=۱/۰۶۷ بیان می‌کند به ازای هرسال افزایش سن، شانس ابتلا به بیماری عروق کرونر ۱/۰۶۷ برابر می‌شود.

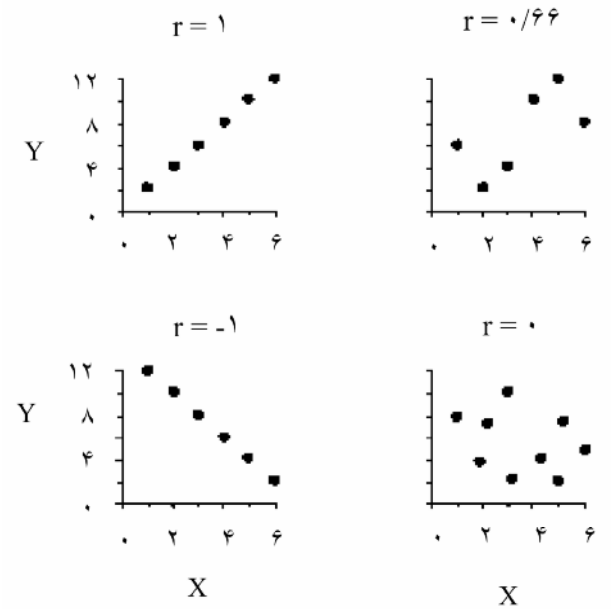
نتیجه‌گیری نهایی

همبستگی به بررسی قدرت ارتباط بین دو متغیر می‌پردازد و هیچ کدام از این دو متغیر، متغیر هدف در نظر گرفته نمی‌شوند و الزاماً رابطه علیتی باهم ندارند. رگرسیون به

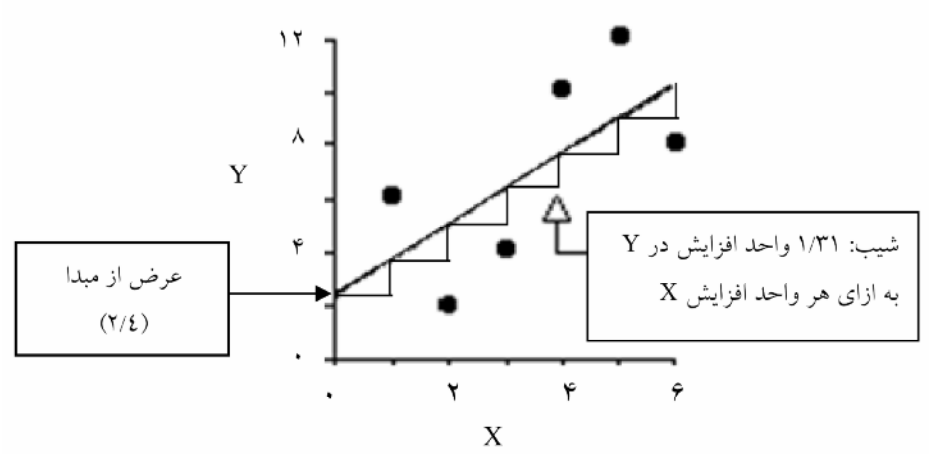
بررسی قدرت ارتباط بین یک یا چند متغیر پیش‌بینی کننده و متغیر پاسخ می‌پردازد و می‌تواند در ارائه مدل‌های پیش‌بینی مانند مثال‌های بالا مفید باشد. چنین مدل‌هایی می‌توانند در تصمیم‌گیری‌های بالینی بسیار مهم باشند. در اینجا بهتر است مجدداً تاکید کنیم که p-value معنی‌دار، اطلاعات زیادی در مورد قدرت همبستگی به ما نمی‌دهد و به مقدار ضریب همبستگی نیز نیاز داریم. در برازش مدل‌های رگرسیون نیز، نه تنها معنی‌داری ارتباط، بلکه قدرت رابطه و یا درصدی از تغییرات متغیر پاسخ که به وسیله متغیرهای پیش‌بینی کننده بیان می‌شود باید مورد توجه قرار گیرد.



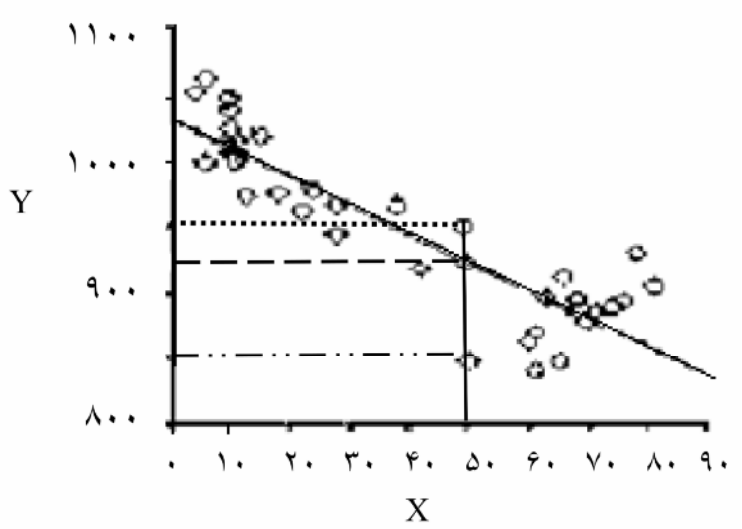
شکل ۱- نمودار پراکنش 2-h PG در مقابل سن برای افرادی که FPG برابر ۱۲۶ (mg/dl) دارند.



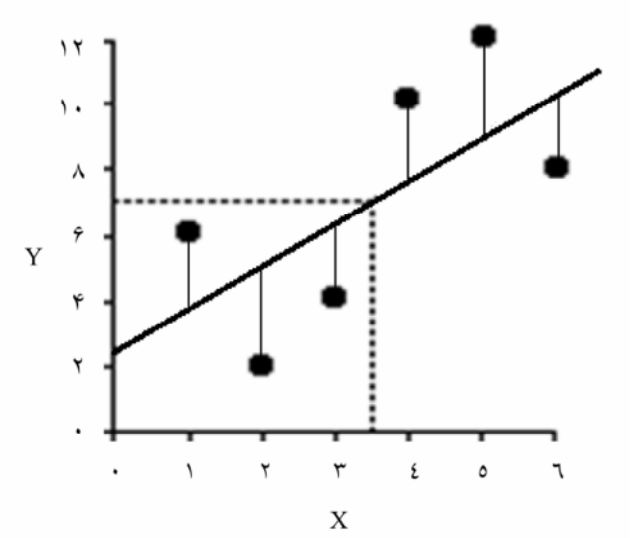
شکل ۲- نمودارهای پراکنش به ازای مقادیر متفاوت فرضی از ضرایب همبستگی



شکل ۳- عرض از مبدا و شیب در معادله رگرسیونی Y=۲/۴+۱/۳۱X



شکل ۴- نمایش مقادیر واقعی و پیش‌بینی شده بر اساس خط رگرسیون به ازای نقطه X=۵۰



شکل ۵- نمایش مقادیر باقیمانده (طول خطوط عمودی بین مقدار واقعی و مقدار پیش‌بینی شده)

مأخذ

1. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians: 4. Correlation and regression. *CMAJ* 1995; 152(4):497-504.
2. Bando Y, Ushiogi Y, Okafuji K, Toya D, Tanaka N, Fujisawa M. The relationship of fasting plasma glucose values and other variables to 2-h postload plasma glucose in Japanese subjects. *Diabetes care* 2001; 24(7):1156-60.
3. Jewell PN, Statistics for Epidemiology, CHAPMAN&HALL/CRC, London New York Washington, 2004.
4. Michael HK, NachtsheimCJ, Neter J, LiW.Applied linear statistical methods. Fifth edition. McGraw-Hill; 2005.
5. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. *Radiology* 2003; 227(3):617-22.
6. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. *Critical care* (London, England) 2003; 7(6):451-9.
7. Williams AC, Bower EJ, Newton JT. Research in primary dental care Part 6: Data analysis. *British dental journal* 2004; 197(2):67-73.
8. Petrie A, Bulman JS, Osborn JF. Further statistics in dentistry. Part 6: Multiple linear regressions. *British dental journal* 2002; 193(12):675-82.
9. Gareen IF, Gatsonis C. Primer on multiple regression models for diagnostic imaging research. *Radiology* 2003; 229(2):305-10.
10. Kleinbaum D, Klein M. Logistic Regression: A Self Learning Text. 2nd edition. Verlag New York Inc: Springer; 2002.
11. Montgomery D, Peck E, Vining G, Krueger D. Introduction to Linear Regression Analysis. 2nd edition, John Wiley & Sons; 2002.
12. Rosenman RH, Friedman M, Straus R,Wurm M, Kositchek R, Hahn W, Werthessen N(1964). A predictive study of coronary heart disease: The Western Collaborative Group Study. *Journal of the American Medical Association* 189, 15-22.