

# به کارگیری متغیرهای پنهان در مدل رگرسیون لجستیک برای حذف اثر هم خطی چندگانه در تحلیل برخی عوامل مرتبط با سرطان پستان

محمد امین پورحسینقلی<sup>۱</sup>، یدا... محرابی<sup>۲</sup>، حمید علوی مجد<sup>۳</sup>، پروین یاوری<sup>۴</sup>

<sup>۱</sup> کارشناس ارشد آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی، تهران.

<sup>۲</sup> دانشیار آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی، تهران.

<sup>۳</sup> استادیار آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی، تهران.

<sup>۴</sup> استاد اپیدمیولوژی، دانشگاه علوم پزشکی شهید بهشتی، تهران.

نویسنده‌ی رابط: دکتر یدا... محرابی، گروه پزشکی اجتماعی و بهداشت دانشکده پزشکی، دانشگاه علوم پزشکی شهید بهشتی، تهران، اوین، کد پستی ۱۹۳۹۵، تلفن: ۸-۲۳۸۷۲۵۶۷-۲۱-۰،

نمبر: ۲۱-۲۲۴۱۴۱۰۸-۰، پست الکترونیک: mehrobi@sbmu.ac.ir ; ymehrobi@gmail.com

تاریخ دریافت: ۸۴/۱۰/۲۶، پذیرش: ۸۵/۲/۳

**مقدمه و اهداف:** رگرسیون لجستیک یکی از کاربردی‌ترین مدل‌های خطی تعمیم‌یافته برای تحلیل رابطه‌ی یک یا چند متغیر توضیحی بر متغیر پاسخ رسته‌ای است. زمانی که بین متغیرهای توضیحی همبستگی‌های نسبتاً قوی وجود داشته باشد هم خطی چندگانه ایجاد شده، ممکن است به کاهش کارایی مدل منجر شود. هدف این تحقیق استفاده از متغیرهای پنهان برای کاهش اثر هم خطی چندگانه در تحلیل یک مطالعه مورد - شاهدهی است.

**روش کار:** داده‌های مورد استفاده در این تحقیق متعلق به یک مطالعه مورد - شاهدهی است که در آن ۳۰۰ نفر زن مبتلا به سرطان پستان با ۳۰۰ زن شاهد از نظر عوامل خطر مورد مقایسه قرار گرفتند. برای بررسی اثر هم خطی، پنج متغیر کمی که بین آن‌ها همبستگی بالایی وجود داشت، در نظر گرفته شدند. ابتدا مدل لجستیک به متغیرهای فوق برازش داده شد. سپس به منظور حذف اثر هم خطی، دو متغیر پنهان با استفاده از هرکدام از دو روش تحلیل عاملی و تحلیل مؤلفه‌های اصلی به دست آورده، بر مبنای آن‌ها پارامترهای مدل‌های لجستیک مجدداً محاسبه شدند. کارایی مدل‌ها، با استفاده از خطای استاندارد پارامترها مقایسه گردید.

**نتایج:** مدل رگرسیون لجستیک براساس متغیرهای اولیه حاکی از مقادیر غیرعادی نسبت شانس برای سن در اولین زایمان زنده (۴۵۳۵۰۳) و  $OR=67960$  و  $CI95\%=[10184, 67960]$  و سن در اولین حاملگی ( $OR=0/00029$ ) بود. درحالی که پارامترهای مدل‌های لجستیک حاصل از متغیرهای پنهان به دست آمده از هر دو روش تحلیل عاملی و تحلیل مؤلفه‌های اصلی، از نظر آماری معنی دار ( $p<0/003$ ) و خطای استاندارد همه‌ی آن‌ها کوچک‌تر از خطای استاندارد مربوط به رگرسیون لجستیک معمولی بود. فاکتورها و مؤلفه‌های اصلی تولید شده توسط دو روش حداقل ۸۵ درصد کل واریانس را تبیین کردند.

**نتیجه‌گیری:** تحقیق نشان داد انحراف استاندارد پارامترهای برآورد شده در رگرسیون لجستیک براساس متغیرهای پنهان از رگرسیون لجستیک براساس مشاهدات اولیه کوچکتر بوده و در نتیجه این‌گونه مدل‌بندی در تحلیل برخی عوامل خطر سرطان پستان که هم خطی دارند کارآتر است.

**واژگان کلیدی:** هم خطی چندگانه، متغیر پنهان، تحلیل عاملی، تحلیل مؤلفه‌های اصلی، رگرسیون لجستیک، سرطان پستان.

## مقدمه

تعداد متغیرهای توضیحی افزایش می‌یابد، مدل‌سازی مشکل شده و کارایی آن نیز کاهش می‌یابد؛ به خصوص اگر برخی از متغیرها علی‌رغم فرض استقلال در مدل‌سازی، با یکدیگر هم‌بستگی قوی

رگرسیون لجستیک یکی از کاربردی‌ترین مدل‌های خطی تعمیم‌یافته است که برای تحلیل رابطه‌ی یک یا چند متغیر توضیحی بر متغیر پاسخ رسته‌ای به کار می‌رود (۱). زمانی که