

# مدل پروبیت و کاربرد آن در داده های پزشکی

سعیده پور احمد<sup>۲</sup>تقی حیدری<sup>۲،۴</sup>سید محمد تقی آیت الله<sup>۱</sup>محمد علی وکیلی<sup>۲،۳</sup>

## چکیده

یکی از مدل های خطی تعمیم یافته که امروزه به طور وسیعی در مطالعات پزشکی مورد استفاده قرار می گیرد، مدل پروبیت می باشد. به کمک این مدل می توان یک مدل رگرسیونی بین متغیر پاسخ دو حالتی با متغیرهای پیش بینی کننده کمی یا کیفی را ارائه نمود. مهمترین کاربرد مدل پروبیت مربوط به عیارگیری زیستی، احتمالات بقا و بررسی فاکتورهای خطر یک بیماری است.

در این مقاله، ابتداء به معرفی مدل پروبیت، نحوه برآورد مناسب بودن مدل و چگونگی برآوردها پرداخته شد و پس از مقایسه آن با مدل لجستیک و مدل لگاریتم دو گانه مکمل، کاربرد مدل پروبیت و لجستیک در سه مثال از داده های پزشکی مورد بررسی قرار گرفت.

**واژه های کلیدی:** پروبیت، لوجیت، داده های پزشکی، مدل خطی تعمیم یافته، مدل لگاریتمی دو گانه مکمل، نرم افزارهای آماری، عیارگیری زیستی

پروبیت<sup>۷</sup> هستند که این مقاله با هدف آشنایی با مدل پروبیت و مقایسه آن با مدل لجستیک با استفاده از مجموعه ای از داده های پزشکی نگارش یافته است.

## ۱. مقدمه

اغلب برای آنالیز داده های عیارگیری زیستی (*Bioassay*) مثل نسبت حشرات کشته شده با غلظت های مختلف یک حشره کش، محاسبه احتمالات بقا در مطالعات مختلف و بررسی فاکتورهای خطر یک بیماری، با مواردی مواجه می شویم که متغیر پاسخ یک متغیر دو حالتی است<sup>[۳]</sup> و [۶]، لذا استفاده از مدل های خطی ساده (مثل رگرسیون معمولی یا آنالیز واریانس) مقدور نمی باشد و باید به دنبال یافتن مدل هایی باشیم که شرط پیوستگی و نرمال بودن متغیر پاسخ را نداشته باشد، به عبارت دیگر در این شرایط بین متغیر وابسته و متغیر(متغیرهای) مستقل یک رابطه خطی وجود ندارد و تنها از طریق تبدیلات مختلف می توان این رابطه را خطی نمود، که این موضوع در قالب مدل های خطی تعمیم یافته بیان می گردد. از بین مدل های خطی تعمیم یافته<sup>۸</sup>، مهم ترین مدل های برای متغیرهای وابسته دو حالتی، مدل های لجستیک<sup>۹</sup> و

۲. مدل پروبیت

در بسیاری از مطالعات مواردی پیش می آید که متغیر پاسخ در مقیاس دو حالتی اندازه گیری می شود، مثلاً پاسخ ها ممکن است مواردی مثل مردن یا زنده ماندن، داشتن یا نداشتن یک بیماری، موفقیت یا شکست یک روش درمانی و نظری آنها باشند. در این حالات اگر متغیر پاسخ را  $Z$  بنامیم و موفقیت را با عدد یک و شکست را با عدد صفر نشان دهیم، خواهیم داشت:

$$Z = \begin{cases} 1 & \text{موفقیت} \\ 0 & \text{شکست} \end{cases}$$

Logistic  
Probit -۷

۱- استاد

۲- دانشجوی دکتری گروه آمار زیستی و اپیدمیولوژی دانشکده

بهداشت، دانشگاه علوم پزشکی شیاز

۳- مری، دانشگاه علوم پزشکی گرگان

۴- مری، دانشگاه علوم پزشکی لرستان

Generalized Linear Model -۵

وقتی که تابع  $f(s)$  یک تابع چگالی احتمال است و داریم:

$$f(s) \geq 0 \quad 2) \int_{-\infty}^{+\infty} f(s) ds = 1 \quad (1)$$

باید این چگالی احتمال را که به توزیع تحمل<sup>۱</sup> معروف است به گونه ای بیابیم که هیچ محدودیتی بر دامنه مقادیر متغیرهای مستقل حاکم نشود.

در سال ۱۹۳۰، چستر بلیس<sup>۲</sup> تابع احتمال نرمال را برای  $f(s)$  پیشنهاد کرد به گونه ای که اگر  $f(s)$  تابع چگالی احتمال نرمال باشد، آنگاه:

$$\pi = g^{-1}(\beta_0 + \beta_1 x) = \int_{-\infty}^{\beta_0 + \beta_1 x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

که در آن  $\Phi$  تابع احتمال تجمعی نرمال استاندارد است. بنابراین

$$\Phi^{-1}(\pi) \beta_0 + \beta_1 x$$

وقتی که  $\beta_0 = \frac{-\mu}{\sigma}$  و  $\beta_1 = \frac{1}{\sigma}$  تابع ارتباط  $g$  هم

معکوس تابع احتمال تجمعی نرمال استاندارد ( $\Phi^{-1}$ ) است، به این مدل اصطلاحاً مدل پروبیت (Probit) گفته می شود که برگرفته از دو کلمه *Probability Unit* است.<sup>[۳]</sup>

مدل پروبیت در زمینه های مختلف از جمله علوم اجتماعی، زیست شناسی، ژنتیک، سم شناسی و پژوهشی کاربردهای فراوان دارد. به خصوص در تحقیقات مربوط به تعیین دوز دارو یا کشنیدگی یک سم که در آنها متغیر پاسخ بصورت موفقیت و شکست تعریف می شود، در این حالات تعبیرهای طبیعی جالبی از این مدل می توان داشت، مثلاً در مطالعه درصد حیوانات آزمایشگاهی که بوسیله سطوح مختلف دوز یک ماده سمی کشته می شوند اگر متغیر مستقل  $X$  را سطوح مختلف دوز ماده سمی در نظر بگیریم و متغیر پاسخ درصد مرگ حیوانات از این دوزها باشد میتوانیم دوزی را بیابیم که بطور میانگین باعث از بین رفتن ۵۰ درصد حیوانات می شود:

$$0.5 = \Phi\left(\frac{x-\mu}{\sigma}\right) \Rightarrow \frac{x-\mu}{\sigma} \Rightarrow x = \mu$$

که اگر احتمال موفقیت را با  $\pi$  نشان دهیم داریم:

$$P(Z = z) = \begin{cases} \pi & z = 1 \\ 1 - \pi & z = 0 \end{cases}$$

برای یک نمونه  $n$  تایی از متغیرهای تصادفی  $z_1, z_2, \dots, z_n$  مستقل از هم با  $P(z_i = 1) = \pi$ ، احتمال توأم آنها برابر است با:

$$\prod_{i=1}^n \pi^{z_i} (1 - \pi)^{1-z_i} = \exp\left[\sum_{i=1}^n z_i \ln\left(\frac{\pi}{1 - \pi}\right) + \sum_{i=1}^n \ln(1 - \pi)\right]$$

که عضوی از خانواده نمایی است.<sup>[۴]</sup>

بنابراین اگر علاقه مند به مدل سازی احتمال موفقیت بر اساس متغیرهای مستقل مورد نظر باشیم می توانیم از مدلهای خطی تعمیم یافته استفاده کنیم زیرا همانطور که می دانیم از عناصر و فرضیات یک الگوی خطی تعمیم یافته، علاوه بر استقلال متغیرهای پاسخ تبعیت تابع احتمال متغیر پاسخ از فرم تابع احتمال خانواده نمایی است.<sup>[۵]</sup>

برای مدل سازی احتمالات  $\pi_i$  با پیروی از الگوی مدل های خطی تعمیم یافته باید تابعی مثل  $g$  بیابیم که

$$g(\pi_i) = X_i^T \beta$$

که در آن  $X_i$  بردار متغیرهای مستقل،  $\beta$  بردار پارامترها و تابع ارتباط بین  $i$  و  $\pi_i$  خواهد بود، که نمی تواند یک تابع همانی باشد، زیرا اگر

$$g(\pi_i) = \pi_i = X_i^T \beta$$

آنگاه با توجه به اینکه  $\pi_i$  یک احتمال است و همواره در بازه [۰، ۱] قرار می گیرد، مقدار برازش شده  $X_i^T \beta$  ممکن است با تغییر مقادیر  $i$  در خارج از این بازه قرار گیرد. برای اطمینان از اینکه مقدار برازش شده برای  $i$  در فاصله [۰، ۱] محدود می شود، باید از توابع ارتباطی استفاده نمود که برد تابع معکوس آنها در محدوده [۰، ۱] قرار گیرد، یکی از توابع ارتباط، تابع توزیع احتمال تجمعی متغیرهای پیوسته است. بر اساس این تابع ارتباط می توان یک مدل برای  $\pi_i$  بدست آورد.

در حالی که یک متغیر مستقل داشته باشیم، داریم:

$$\pi = g^{-1}(\beta_0 + \beta_1 x) = \int_{-\infty}^x f(s) ds$$

ارجحیت دارد. برای استفاده از روش فوق به ترتیب زیر عمل می کنیم:

$$\text{اگر } p = \frac{y}{n} \text{ سهم موفقیت باشد و } \Psi \text{ را تابعی از } p \text{ در نظر}$$

بگیریم، بسط سری تیلور<sup>۵</sup>  $\Psi(p) \approx \Psi(\pi) + (\frac{y}{n} - \pi) \Psi'(\pi)$  حول نقطه  $p = \pi$  (احتمال موفقیت) برابر است با:

$$\Psi(p) = \Psi\left(\frac{y}{n}\right) = \Psi(\pi) + \left(\frac{y}{n} - \pi\right) \Psi'(\pi) + o\left(\frac{1}{n}\right) \quad (1)$$

از اولین تقریب در معادله (1) نتیجه می شود:

$$E(\Psi(p)) = \Psi(\pi)$$

چون

$$E(p) = E\left(\frac{y}{n}\right) = \pi$$

همچنین

$$\text{var}(\Psi(p)) = E[\psi(p) - \psi(\pi)]^2 = [\Psi'(\pi)]^2 E\left(\frac{y}{n} - \pi\right)^2$$

$$[\Psi'(\pi)]^2 \frac{\pi(1-\pi)}{n}$$

$$\text{var}\left(\frac{y}{n}\right) = \text{var}(p)$$

بنابراین آماره حداقل مربعات وزنی برابر است با:

$$\chi^2 = \sum_{i=1}^N \left[ \frac{\left[ \Psi\left(\frac{y_i}{n_i}\right) - \Psi(\pi_i) \right]^2}{[\Psi'(\pi_i)] \pi_i (1-\pi_i)/n_i} \right]$$

در واقع مقدار خطای برآورد عبارت است از:

$$\Psi\left(\frac{y_i}{n_i}\right) - \Psi(\pi_i)$$

که با واریانس آن وزن داده شده است. بعضی توابع معمول  $\Psi$  در جدول زیر خلاصه شده است.(جدول ۱)

اگر  $\Psi(\pi_i) = \pi_i$  و  $\pi_i = X^T \beta$  مقدار کای اسکور اصلاح شده برابر است با :

$$\chi^2_{\text{mod}} = \sum_{i=1}^N \frac{(p_i - \pi_i)^2}{\pi_i (1-\pi_i)/n_i}$$

به این سطح ( $\mu = x$ ) دوز کشنده میانه<sup>۶</sup> یا  $LD_{(50)}$  گفته می شود که برای نشان دادن قدرت کشنده‌گی مواد سمی از جمله حشره کشها، شاخص مناسبی به حساب می آید.<sup>۷</sup>

### ۳. روش های برآورد پارامترهای مدل

پارامترهای مدل پربویت را با استفاده از دو روش ماکریم درستنمایی (MLE) و حداقل مربعات وزنی، که روش های معمول برآورد پارامترها در مدل های رگرسیونی با پاسخ دو حالتی هستند، برآورد می کنیم.<sup>۴ و ۵</sup>

۱-روش ماکریم درستنمایی<sup>۲</sup> (MLE): این روش حتی برای حالتی که  $y_i = 0$  یا  $n_i = 1$  باشد نیز امکان پذیر است (برخلاف بعضی روش ها مثل روش حداقل مربعات). اگر  $U$  بردار نمره ها (Score vector) و  $I$  ماتریس اطلاع فیشر باشد :

$$U = \begin{bmatrix} \frac{\partial l}{\partial \beta_i} \end{bmatrix} \quad i = 1, 2, \dots, n$$

$$I = \begin{bmatrix} \frac{\partial^2 l}{\partial \beta_i \partial \beta_j} \end{bmatrix} \quad i, j = 1, 2, \dots, n$$

آن گاه با تشکیل معادله زیر و حل آن به روش تکراری می توان برآوردهای ماکریم درستنمایی را برای پارامترهای مدل بدست آورد:

$$I^{(m-1)} b^m = I^{(m-1)} b^{(m-1)} + U^{(m-1)}$$

که در آن  $m$  نشان دهنده تقریب  $m$  و  $b$  بردار برآوردهاست.

ابتدا با جایگذاری مقادیر اولیه برای بردار پارامترها ( $b^0$ ) در مدل، تقریب بعدی برآورد پارامترها ( $b^1$ ) را محاسبه کرده و این روش را تا جایی تکرار می کنیم که اصطلاحاً همگرایی<sup>۳</sup> حاصل شود. به عبارت دیگر دو تقریب متوالی از برآوردها کاملاً به هم نزدیک شوند در این صورت تقریب نهایی را می توان به عنوان برآورد پارامترهای مدل در نظر گرفت.

۲-روش حداقل مربعات وزنی<sup>۴</sup>: برای پرهیز از تکرار در محاسبات، روش حداقل مربعات وزنی بر روش ماکریم درستنمایی

Median lethal dose -۱

Maximum Likelihood Estimation -۲

Converge -۳

Weighted Least Square Method -۴

درستنمائی برای مدل پروبیت می باشد. آماره  $D$  از توزیع کای اسکور با درجه آزادی  $N-p$  پیروی می کند ( $D: \chi^2_{N-p}$ ). مناسب بودن مدل با مقایسه آماره  $D: \chi^2_{N-p}, \alpha$  با  $\beta$  بررسی می شود، که در آن  $p$  تعداد پارامترهای برآورد شده  $\beta$  است. می توان  $D$  را با فرمول زیر نیز محاسبه نمود :

$$D = 2 \sum o \ln \frac{o}{e}$$

وقتی که  $o$  مقادیر مشاهده شده  $y_i$  و  $(n_i - y_i)$  و  $e_i$  مقادیر مورد انتظار از مدل برآش شده اند (جدول ۲).

### ۲- روش مینیمم کردن مجموع مربعات وزنی: به جای

استفاده از روش درستنمائی می توان از روش کمینه سازی<sup>۳</sup> مجموع مربعات وزنی برای برآش مدل استفاده کرد. داریم:

$$S_w = \sum_{i=1}^N \frac{(y_i - n_i \pi_i)^2}{n_i \pi_i (1 - \pi_i)}$$

که این روش معادل کمینه سازی آماره کای اسکور پیرسن<sup>۴</sup> می باشد

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

که در آن  $o$  مقادیر مشاهده شده و  $e$  مقادیر برآش می باشد (جدول ۲).

### ۳- استفاده از آماره کای اسکور اصلاح شده

$$\chi^2_{\text{mod}} = \sum_{i=1}^N \frac{(y_i - n_i \pi_i)^2}{n_i \pi_i (1 - \pi_i)}$$

اگر مدل صحیح باشد هر سه آماره  $D$ ,  $\chi^2$ ,  $\chi^2_{\text{mod}}$  داری توزیع  $\chi^2_{N-p}$  می باشند. شواهد نشان می دهد که  $\chi^2$  بهتر از  $D$  است، زیرا  $D$  تحت تأثیر فراوانی های کوچک قرار می گیرد. هر سه آماره وقتی مقادیر مورد انتظار کوچک باشند، ضعیف عمل می کنند. (جدول ۲)

## ۵. روش های جایگزین مدل پروبیت

در مواردی که متغیر پاسخ دو حالتی است، علاوه بر مدل پروبیت می توان از مدل های جایگزین زیر هم استفاده کرد:

برای برآورد  $\beta$  در این حالت نیاز به روش تکراری نیست. اگر چه ممکن است مقدار  $\hat{\pi}_i = X^T \beta$  همواره بین صفر و یک نباشد.

اگر  $\Psi(\pi_i) = \log it(\pi_i)$  داریم:

$$\frac{\exp(X^T \beta)}{1 + \exp(X^T \beta)}$$

پس:

$$\chi^2_{\text{mod}} = \sum_{i=1}^N (z_i - x_i^T \beta)^2 \frac{y_i (n_i - y_i)}{n_i} \quad (2)$$

$$\text{وقتی که } z_i = \log it(p_i) = \ln \frac{y_i}{n_i - y_i}$$

در این مدل نیز برای برآورد پارامترها نیاز به روش تکراری نمی باشد و  $\pi_i$  ها در دامنه صفر و یک قرار می گیرند. کاکس این روش را تبدیل لجستیک کاربردی می نامد و پیشنهاد می کند از  $E(z_i - x_i^T \beta)$   $z_i = \ln \frac{y_i + 0/5}{n_i - y_i + 0/5}$  استفاده شود. مینیمم مقداری که از (2) بدست می آید را آماره کای اسکور لجستیک مینیمم می نامیم.

مدل  $\Psi(\pi_i) = \arcsin \sqrt{\pi_i}$  با هر انتخابی برای  $\pi_i$  دارای خاصیت پایایی واریانس می باشد زیرا

$$\text{var}(\Psi(p_i)) = (\Psi'(\pi_i))^2 \pi_i (1 - \pi_i) / n = (4n_i)^{-1}$$

بنابراین وزن ها به پارامترها یا پاسخ ها بستگی ندارد و در نتیجه محاسباتی که از این تبدیل استفاده می کند بسیار ساده بوده و با ماشین حساب معمولی هم قابل محاسبه می باشد.

## ۶. سنجش برآش<sup>۱</sup> مدل پروبیت

پس از برآش کردن مدل پروبیت می توان مناسب بودن مدل برآش شده به داده ها را با یکی از روش های زیر بررسی کرد:

### ۱- استفاده از آماره نسبت لگاریتم درستنمائی<sup>۲</sup>: این

آماره به صورت زیر تعریف می گردد:

$$D = 2 [l(\hat{\pi}_{\max}; y) - l(\hat{\pi}; y)]$$

وقتی که  $\hat{\pi}$ ,  $\hat{\pi}_{\max} = \frac{y_i}{n_i}$  هم بردار برآوردهای ماقزیم

راهنمایی و متوسطه مدارس دولتی، نمونه مردمی و غیر انتحاعی شیراز در سال تحصیلی ۱۳۷۶-۷۷ انجام شده است<sup>(۷)</sup>.

در این مطالعه با روش نمونه گیری چند مرحله‌ای یک نمونه دو درصدی از کل دانش آموزان مورد نظر به حجم ۱۸۶۶ نفر مورد بررسی قرار گرفته و تعداد و درصد افراد قاعده شده در هر گروه سنی محاسبه شده اند، جدول زیر اطلاعات مربوطه را نشان می‌دهد.

لازم به ذکر است که متغیر مستقل سن (۲) به صورت میانگین سن در هر گروه سنی در نظر گرفته شده و متغیر وابسته ( $P$ ) (احتمال قاعده‌گی) از طریق فراوانی نسبی تعداد افراد قاعده شده در هر گروه سنی محاسبه شده است (جدول شماره ۳).

پس از برازش مدل پربویت به وسیله نرم افزار SPSS و محاسبه ضرایب مدل احتمال قاعده‌گی در هر گروه سنی با استفاده ازتابع ارتباط نرمال  $(\hat{\alpha} + \hat{\beta}\chi^* = \Psi)$  و مقادیر برازش شده مدل از فرمول  $e_i = n_i p_i$  محاسبه و سپس سنجش برازش مدل با آماره  $\chi^2$  بررسی شده است. در جدول ۴، مقادیر برازش شده و در جدول ۵، هر دو روش برآورد ضرایب مدل (MLE و WLS) در دو مدل لجستیک و پربویت ارائه و مقایسه شده است. (جدول ۳ و ۴)

### مثال ۲: در یک مطالعه مورد شاهدی در شهرستان جهرم،

ضمن جوری‌سازی سن و جنس، عوامل مؤثر بر پیرگوشی در ۱۵۲ فرد بیمار و ۱۵۲ فرد بیمار مورد بررسی قرار گرفتند. با استفاده از مدل پربویت رابطه بین متغیر پاسخ (ابتلا یا عدم ابتلا به بیماری) و متغیرهای مستقل (آلودگی صوتی محیط کار، سابقه خانوادگی، بیماری دیابت، فشارخون و چربی خون، سرگیجه و وزوز گوش) مورد بررسی قرار گرفت<sup>[۸]</sup>. تمام متغیرهای مستقل دو حالتی بودند و مشخص گردید که متغیرهای آلودگی صوتی محیط کار، سابقه خانوادگی و وزوز گوش بر پیرگوشی مؤثر هستند، با بکارگیری مدل لجستیک نیز همین نتایج تأیید گردید، که نتایج حاصله در جدول ۶ تا ۸ ارائه گردیده است. جدول شماره ۶ مقادیر مشاهده شده و برازش شده را در هر دو مدل نشان می‌دهد. همانطور که مشاهده می‌گردد تفاوت قابل توجهی بین مقادیر برازش شده در هر دو مدل وجود ندارد. در جدول شماره ۷ برآورد MLE پارامترها در مدل‌های برازش شده به همراه خطای استاندارد آنها آورده شده است. در جدول شماره ۸ مقادیر  $D^2$  برای سنجش برازش مدل‌ها آمده

۱- مدل لجستیک: این مدل بسیار به مدل پربویت نزدیک است به گونه‌ای که گاهی اوقات با مدل پربویت اشتباه گرفته می‌شود. در این مدل، که به جای تابع ارتباط توزیع تجمعی نرمال از تابع ارتباط لوجیت استفاده می‌شود، داریم:

$$g(\pi) = \ln \frac{\pi}{1-\pi} = X^T \beta \Rightarrow \pi = \frac{\exp(X^T \beta)}{1+\exp(X^T \beta)}$$

$$\frac{\pi}{1-\pi} = odds \quad , \quad \ln \frac{\pi}{1-\pi} = \ln(odds) = \log it$$

نمودار دو تابع لوجیت و پربویت S شکل، برد آنها بین صفر و یک، حول  $\pi = 0.5$  متقاضی، در بازه  $0.2-0.8$  تقریباً خطی و اختلاف دو مدل در نقاط ابتدایی و انتهایی مقادیر  $x$  می‌باشد (شکل شماره ۱). مدل لجستیک بیشتر برای مطالعات مشاهده‌ای مناسب است در حالی که مدل پربویت برای مطالعات تجربی به کار می‌رود. در مطالعاتی که علاقه مند به محاسبه نسبت شانس<sup>۱</sup> برای متغیرهای مستقل باشیم از مدل لجستیک استفاده می‌کنیم در حالی که روش پربویت برای تعیین مقادیر کلیدی و مؤثر متغیر مستقل که  $\hat{\pi}$  دارای مفهوم خاصی برای ما باشد مثل دوز کشته میانه یعنی دوزی که موجب مرگ و میر حداقل ۵۰ درصد حیوانات آزمایشگاهی شود

(LD<sub>۵۰</sub>). [۲].

۲- مدل لگاریتم دوگانه مکمل<sup>۲</sup>: تابع ارتباط دیگری که می‌تواند جایگزین مدل پربویت شود تابع لگاریتم دوگانه مکمل می‌باشد که بصورت زیر تعریف می‌شود:

$$g(\pi) = \ln[-\ln(1-\pi)] = X^T \beta \Rightarrow \pi = 1 - \exp[-\exp(1-\pi)]$$

این مدل نیز وقتی که  $\pi$  به  $0.5$  نزدیک باشد شبیه مدل پربویت و لجستیک است، اما هر چه  $\pi$  به صفر یا یک نزدیک تر می‌شود از دو مدل دیگر فاصله می‌گیرد و برای مواردی استفاده می‌شود که پراکنندگی بین داده‌ها زیاد باشد یا مقادیر افراطی در داده‌ها موجود باشد.

## ۶. کاربرد مدل پربویت در داده‌های پزشکی و مقایسه آن با مدل لجستیک

مثال ۱: در یک بررسی به منظور برآورد احتمال قاعده‌گی در گروههای سنی مختلف، مطالعه‌ای بر روی دانش آموزان دختر مقاطع

متغیرهای توضیحی را در قسمت Model وارد کرده و در صورتی که فاکتور باشند در قسمت Factor آنها را تعریف می کنیم، بطور پیش فرض Minitab رگرسیون لجستیک را برآش می کند، برای برآش مدل پروبیت گزینه Option را باز نموده و گزینه Probit را انتخاب می کنیم (در ضمن می توانیم مدل لگاریتم دوگانه مکمل را به همین طریق انتخاب کنیم) و سپس به منوی اصلی باز می گردیم. اگر بخواهیم اطلاعات بیشتری درباره برآش مدل بدست آوریم مثل مقادیر باقیمانده ها، نمودارهای برآش مدل و ... گزینه های دیگر این منو را انتخاب می کنیم.

**روش WLS:** از منوی Calc گزینه Calculator را انتخاب نموده و مقادیر احتمال در هر گروه را بدست می آوریم، سپس از همان منو گزینه Probability Distribution را انتخاب نموده Inverse گزینه Normal را انتخاب می کنیم، گزینه Input cumulative probability را انتخاب نموده در قسمت optional strong ستون خروجی را مشخص می کنیم. از منوی Stat گزینه Regression را انتخاب نموده، متغیر فوق را بعنوان پاسخ و متغیرهای توضیحی (که باید پیوسته باشند) را بسته به موضوع انتخاب می کنیم و در گزینه عکس تعداد موارد را بعنوان وزن تعریف می کنیم و سپس به منوی اصلی بازگشته و آنالیز را به اتمام می رسانیم.

است که مناسب بودن برآش مدل در هر دو مدل تأیید می شود.  
(جدول ۶۷ و ۸)

**مثال ۳:** در شهرستان گرگان، مطالعه ای با هدف تعیین فرسودگی شغلی و عوامل موثر بر آن در جامعه پرستاران شاغل در مرکز آموزشی درمانی پنج آذر انجام گردید. از ۱۹۲ فرد مورد مطالعه، ۴۷ نفر (۲۴/۵ درصد) با فرسودگی شغلی بودند [۱]، با استفاده از نرم افزار آماری Minitab ، رابطه فرسودگی شغلی با جنس، وضعیت تأهل، سن، ساعات کار در هفته و سابقه کار با مدل پروبیت و لجستیک مورد بررسی قرار گرفت. جهت بررسی مناسب بودن مدل از آماره های  $D^2$ ،  $\chi^2$  بهره گرفتیم که هر دو مقدار آماره های پیش گفت در هر دو مدل، بیانگری برآش مناسب داده ها می باشد (جدول شماره ۹).

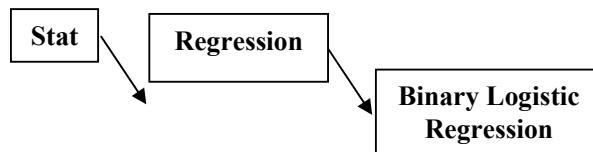
در بررسی رابطه فرسودگی شغلی با متغیرهای مستقل با استفاده از مدل پروبیت و با اثرات اصلی، یافته های بیانگر آن است که فرسودگی شغلی با وضعیت تأهل با اختلاف معنی دار می باشد، ولی نسبت به سایر متغیرها مستقل باشد، بدین ترتیب که فرسودگی شغلی در متأهلین نسبت به مجردین بیشتر می باشد، که این نتایج با یافته های حاصله از مدل لجستیک نیز همخوانی دارد (جدول شماره ۱۰).  
(جدول شماره ۹).

## ۷. طریقه استفاده از نرم افزارهای آماری

اکثر نرم افزارهای آماری قابلیت انجام آنالیز پروبیت و روشهای جایگزین را دارند. در این قسمت به اختصار طرز استفاده از نرم افزارهای SPSS [۹] و Minitab [۱۰] برای انجام محاسبات فوق بیان می شود:

### نرم افزار Minitab

**روش MLE:** از منوی stat گزینه Regression و سپس گزینه Binary Logistic Regression را انتخاب می کنیم.



فرمت داده ها باید به صورت جدول فراوانی باشد. در گزینه Response متغیر پاسخ (فراوانی پیروزی های مشاهده شده)، در گزینه Total Observed مقدار کل شکست و پیروزی، در گزینه Factor گزینه Total Observed بررسی (باید دامنه تغییرات فاکتور را نیز مشخص نمود) و در گزینه Covariate متغیرهای توضیحی مدل را وارد می کنیم.

**روش WLS:** از منوی Transform گزینه Compute را انتخاب نموده و مقادیر احتمال را در هر گروه بدست می آوریم، سپس دوباره به همان گزینه برگشته و Function IDF.Normal و

در منوی ظاهر شده متغیر وابسته را در فیلد Response (اگر داده ها به صورت جدول فراوانی وارد شده باشند ستون مربوط به فراوانی داده ها را در فیلد Frequency وارد می کنیم) و

توزیع تجمعی نرمال (مدل پروبیت) و تابع توزیع تجمعی لجستیک (مدل لجستیک) برای تحلیل چنین داده هایی بوده است. در مدل لجستیک امکان محاسبه مقادیر نسبت خطر در مطالعاتی که یک گروه از اعضای نمونه با عامل خاصی مواجه داده شده اند و هدف بررسی اثر عامل خطر در نحوه پاسخگویی اعضای نمونه است (متغیر پاسخ) وجود دارد، که این مقادیر در تحقیقات پزشکی حائز اهمیت می باشد. با این وجود به کارگیری مدل لجستیک و پروبیت در سه مثال ارائه شده نشان داد که دو مدل در تعیین تاثیر متغیرهای مستقل با متغیر وابسته و میزان معنی داری آنها نتایج مشابهی ارائه نمودند که این نتایج با الگوی مقایسه ای نمودار یک همخوانی دارد. می دانیم که تفاوت عمدی مدل لجستیک و پروبیت در بررسی رابطه یک متغیر مستقل با متغیر وابسته دو حالتی در نقاط ابتدایی و انتهایی مقادیر متغیر مستقل می باشد (نمودار یک) و مهم ترین تفاوت کاربردی آنها نیز در مناسب بودن مدل پروبیت برای مطالعات تجربی و مدل لوจیت برای مطالعات مشاهده ای می باشد، که به دلیل محدودیت مثالهای ارائه شده به خوبی نتوانستند تمایز جدی دو روش را نمایش دهند. ولی می توان گفت در اغلب موارد، به کارگیری هر کدام از دو مدل لجستیک یا پروبیت در داده های با متغیر پاسخ دو حالتی مناسب بوده و اختلاف قابل توجهی مطرح نمی باشد.

علاوه بر آن همانطور که قبلاً گفته شده مناسب بودن برآش مدل لگاریتم دوگانه مکمل بر داده های ابتدائی و انتهائی بیانگر وجه تمایز این مدل با مدلهای پروبیت و لجستیک است، و این مدل به عنوان جایگزینی برای مدل پروبیت در مواردی استفاده می شود که داده های پرت در مشاهدات وجود داشته باشد.

ستون احتمالات را بعنوان  $p$  و اعداد  $\pi_i$  را که میانگین و واریانس توزیع نرمال استاندارد می باشد را تعریف می کنیم. پس از انجام این تبدیلات از منوی Analyze گزینه Regression را انتخاب نموده، متغیر فوق را بعنوان پاسخ و متغیرهای توضیحی (که باید پیوسته باشند) را بسته به موضوع انتخاب می کنیم و در گزینه WLS عکس تعداد موارد را بعنوان وزن تعریف می کنیم و سپس به منوی اصلی بازگشته و آنالیز را به اتمام می رسانیم.

**تذکر :** نرم افزارهای SPSS و Minitab ، به طور متداول، قابلیت محاسبه برآورد پارامترها به روش WLS در مدل های لجستیک و لوجیت را ندارند و به همین دلیل تنها در مثال اول که شامل متغیر مستقل پیوسته است، با استفاده از روش رگرسیون معمولی و با متغیر پاسخ تبدیل یافته، برآورد پارامترها به روش حداقل مربعات وزنی محاسبه شده است ولی در مثال های ۲ و ۳، با توجه به حضور متغیرهای مستقل ناپیوسته، امکان محاسبه برآورد پارامترها به روش WLS نبوده است.

## ۸. نتیجه گیری

کاربرد بر جسته مدل پروبیت مربوط به مطالعات عیارگیری زیستی و بیشتر به منظور تعیین دوزی از دارو است که موجب از بین رفتن حداقل ۵۰ درصد نمونه ها (حشرات) در بررسی اثر حشره کشها می باشد.

در این مقاله، ما کاربرد مدل پروبیت را به عنوان یکی از سه مدل مورد استفاده در داده هایی با متغیر پاسخ دو حالتی با ذکر سه مثال در مجموعه تحقیقات پزشکی مورد بررسی قرار داده ایم. همانطور که از نتایج قابل مشاهده است این مدل بسیار شبیه به مدل لجستیک بوده و هر دو مدل این نوع داده ها را به خوبی برآش می کنند، ولی بدیهی است که دو نگرش مختلف از نحوه پراکندگی داده ها بر اساس احتمالات مربوطه (نمودار ۱) ایده اصلی استفاده از

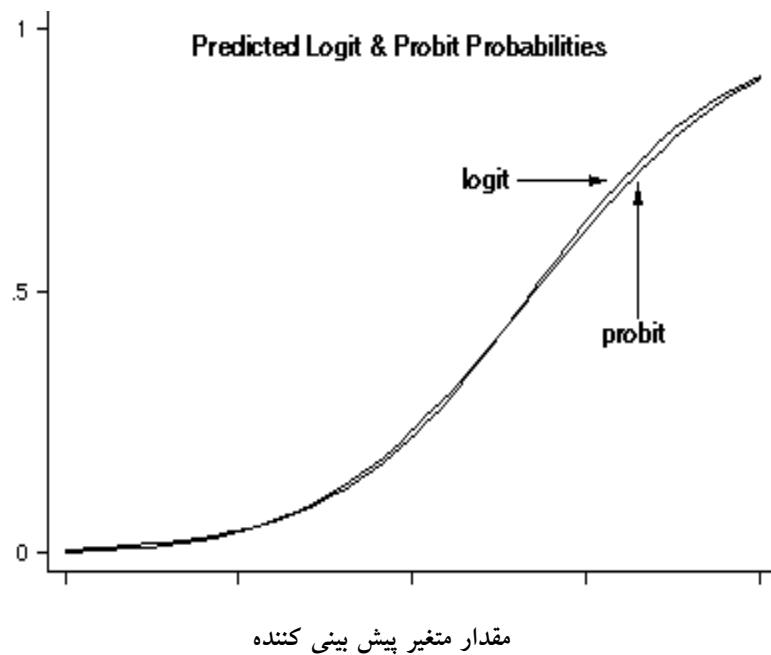
جدول ۱. مدل های حداقل مربعات وزنی معمول برای داده های دو حالتی

$\Psi(\pi_i)$	$\chi^2$
$\pi_i$	$\sum_{i=1}^N \frac{(p_i - \pi_i)^2}{\pi_i(1 - \pi_i)/n_i}$
$\log it(\pi_i)$	$\sum_{i=1}^N [(\log it(p_i)) - \log it(\pi_i(1 - \pi_i)/n_i)]$
$\arcsin \sqrt{\pi_i}$	$\sum_{i=1}^N n_i [\arcsin \sqrt{p_i} - \arcsin \sqrt{\pi_i}]$

جدول ۲. فراوانی  $N$  جامعه دارای توزیع دوجمله‌ای

زیر گروهها				
	۱	۲	...	$N$
موفقیت	$y_1$	$y_2$		$y_N$
شکست	$n_1 - y_1$	$n_2 - y_2$		$n_N - y_N$
مجموع	$n_1$	$n_2$		$n_N$

نمودار ۱. مقایسه منحنی پروبیت و لجستیک



جدول ۳. مقادیر احتمال مشاهده شده و برآش شده قاعده‌گی به وسیله مدل پروبیت در گروههای سنی مختلف

گروههای سنی	X	تعداد کل افراد	تعداد افراد قاعده شده	P	$P^*$
۱۱-۱۱/۵	۱۱/۲۵	۶۰	۵	•/•۸۳۳	•/•۸۶۲
۱۱/۵-۱۲	۱۱/۷۵	۱۵۷	۲۰	•/۱۲۷۴	•/۱۶۷۴
۱۲-۱۲/۵	۱۲/۲۵	۱۳۱	۳۸	•/۲۹۰۱	•/۲۸۶۲
۱۲/۵-۱۳	۱۲/۷۵	۲۵۳	۱۱۲	•/۴۴۲۸	•/۴۳۴۵
۱۳-۱۳/۵	۱۳/۲۵	۱۳۵	۸۶	•/۸۳۷۰	•/۵۹۲۹
۱۳/۵-۱۴	۱۳/۷۵	۲۵۳	۱۸۷	•/۷۳۹۱	•/۷۳۷۲
۱۴-۱۴/۵	۱۴/۲۵	۲۰۳	۱۷۹	•/۸۸۱۸	•/۸۴۹۶
۱۴/۵-۱۵	۱۴/۷۵	۱۴۸	۱۳۱	•/۸۸۵۱	•/۹۲۴۳
۱۵-۱۵/۵	۱۵/۲۵	۱۷۳	۱۶۸	•/۹۷۱۱	•/۹۶۶۷
۱۵/۵-۱۶	۱۵/۷۵	۳۰	۲۹	•/۹۶۶۷	•/۹۸۷۳
۱۶-۱۶/۵	۱۶/۲۵	۱۷۲	۱۷۰	•/۹۸۸۴	•/۹۹۵۸
۱۶/۵ و بیشتر	۱۶/۷۵	۱۴۷	۱۴۷	۱	•/۹۹۸۸

جدول ۴. مقایسه مقادیر برآش شده مدل پروبیت و لجستیک

گروههای سنی	مقادیر مشاهده شده	مقادیر برآش شده مدل پروبیت	مقادیر برآش شده مدل لجستیک
۱۱-۱۱/۵	۵	۵/۱۷۴	۵/۰۸۰
۱۱/۵-۱۲	۲۰	۲۶/۲۸۳	۲۴/۶۸۶
۱۲-۱۲/۵	۳۸	۳۷/۴۸۷	۳۵/۸۱۷
۱۲/۵-۱۳	۱۱۲	۱۰۹/۹۴۰	۱۰۹/۱۶۵
۱۳-۱۳/۵	۸۶	۸۰/۰۴۱	۸۱/۶۵۶
۱۳/۵-۱۴	۱۸۷	۱۸۶/۵۲	۱۹۱/۱۰۳
۱۴-۱۴/۵	۱۷۹	۱۷۲/۴۶۶	۱۷۴/۹۱۱
۱۴/۵-۱۵	۱۳۱	۱۳۶/۷۹۴	۱۳۷/۰۸۵
۱۵-۱۵/۵	۱۶۸	۱۶۷/۲۳۸	۱۶۶/۴۳۰
۱۵/۵-۱۶	۲۹	۲۹/۶۱۸	۲۹/۴۲۴
۱۶-۱۶/۵	۱۷۰	۱۷۱/۲۷۴	۱۷۰/۳۴۷
۱۶/۵ و بیشتر	۱۴۷	۱۴۶/۸۲۲	۱۴۶/۲۹۶

جدول ۵. برآورد ضرایب و سنجش برآش مدل پروبیت و لجستیک

مدل لجستیک			مدل پروبیت				
P-value	خطای معیار	مقدار	P-value	خطای معیار	مقدار	ضرائب	
•/۰۰۰	•/۸۹۸	-۱۸/۱۶۶	•/۰۰۰	•/۴۷	-۱۰/۳۶۰	MLE	
•/۰۰۰	•/۰۹۳	-۱۸/۳۳۸	•/۰۰۰	•/۰۶	-۱۰/۵۴۳		
•/۰۰۰	•/۰۶۸	۱/۴۵۳	•/۰۰۱	•/۳۶	•/۷۹۹	MLE	
•/۰۰۰	•/۰۰۸	۱/۴۱۲	•/۰۰۰	•/۰۰۵	•/۸۱۰		
•/۶۱۴	df=۱۰	$\chi^2=۸/۱۵۷$	•/۳۲۶	df=۱۰	$\chi^2=۱۱/۴۱$		

جدول ۶. مقایسه مقادیر مشاهده شده و برآذش شده داده های پیرگوشی توسط دو مدل پروبیت و لجستیک

وزوز گوش				
خیر	بلی			آلودگی صوتی محیط کار
(۱۳/۰۴) ۱۳ [۱۲/۹۷]	(۴/۹۷) ۵ [۴/۹۳]	بلی	سابقه خانوادگی	بلی
(۸/۶۳) ۱۹ [۱۷/۹۶]	(۸/۱۳) ۷ [۸/۱۴]	خیر		
(۲۶/۳۳) ۲۷ [۲۶/۶۶]	(۱۰/۵۲) ۱۰ [۱۰/۴۴]	بلی		خیر
(۵۰/۹۷) ۴۹ [۵۰/۴۰]	(۲۰/۰۱) ۲۲ [۲۰/۴۹]	خیر		

پرانتر مقدار برآذش شده مدل پروبیت و برآذش مقدار برآذش شده مدل لجستیک

جدول ۷. مقادیر برآذش شده برآورده MLE پارامترهای مدل

مدل لجستیک			مدل پروبیت			متغیر
P-Value	خطای استاندارد	ضرائب	P-Value	خطای استاندارد	ضرائب	
.۰/۰۰۰	.۰/۶۱	-۴/۲۸	.۰/۰۰۰	.۰/۳۴	-۲/۵۲	ثابت
.۰/۰۰۰	.۰/۳۷	۱/۳۶	.۰/۰۰۰	.۰/۲۲	.۰/۸۰	آلودگی صوتی محیط کار
.۰/۰۰۰	.۰/۳۸	۲/۰۴	.۰/۰۰۰	.۰/۲۲	۱/۲۱	سابقه خانوادگی
.۰/۰۰۰	.۰/۴۰	۱/۷۵	.۰/۰۰۰	.۰/۲۳	۱/۰۳	وزوز گوش

جدول ۸. بررسی مناسب بودن مدل برآذش شده پروبیت و لجستیک

لجدستیک		پروبیت		درجه آزادی	آماره
P-Value	مقدار	P-Value	مقدار		
.۰/۶۰۴	۲/۶۷	.۰/۵۰۶	۳/۳۲	۴	$\chi^2$
.۰/۶۸۱	۲/۳۰	.۰/۵۸۱	۲/۸۶	۴	D

جدول ۹. بررسی مناسب بودن مدل برآذش شده پروبیت و لجستیک

لجدستیک		پروبیت		درجه آزادی	آماره
P-Value	مقدار	P-Value	مقدار		
.۰/۲۵۳	۱۷۶/۶۸۹	.۰/۲۴۹	۱۷۶/۹۰۹	۱۶۵	$\chi^2$
.۰/۱۰	۱۸۸/۶۱۲	.۰/۱۰	۱۸۸/۶۱۳	۱۶۵	D

جدول ۱۰. مقادیر برآورد شده برآورده MLE پارامترهای مدل

مدل لجستیک			مدل پروبیت			متغیر
P-Value	خطای استاندارد	ضرائب	P-Value	خطای استاندارد	ضرائب	
.۰/۶۸۲	۲/۳۲۳	.۰/۹۵۲	.۰/۶۳۷	۱/۳۲۹	.۰/۶۲۷	ثابت
.۰/۲۶۷	.۰/۰۲۷	.۰/۰۳۰	.۰/۳۱۲	.۰/۰۱۶	.۰/۰۱۶	سن
.۰/۱۴۴	.۰/۴۳۶	-.۰/۶۳۷	.۰/۱۳۴	.۰/۲۵۷	-.۰/۳۸۵	جنس
.۰/۰۴۱	.۰/۴۵۴	-.۰/۹۲۶	.۰/۰۴۵	.۰/۲۶۱	-.۰/۵۲۴	وضعیت تأهل
.۰/۳۷	.۰/۰۵۵	-.۰/۰۵۰	.۰/۳۳۹	.۰/۰۳۱	-.۰/۰۳۰	ساعات کار در هفته
.۰/۴۶۳	.۰/۰۷۵	.۰/۰۵۵	.۰/۳۳۱	.۰/۰۴۴	.۰/۰۳۴	سابقه کار

## مراجع

- [۱] مطلبی، هومن، ۱۳۸۳، بررسی میزان شیوع فرسودگی شغلی در کارشناسان شاغل در مرکز آموزشی درمانی شهرستان گرگان. پایان نامه دکترای حرفه ای ، دانشگاه علوم پزشکی گرگان.
- [۲] خردمند نیا، منوچهر، رستگاری، مصطفی، ۱۳۸۲، آشنایی با الگوهای خطی تعمیم یافته ، اندیشه آماری ، سال هشتم ،شماره دوم ،پاییز و زمستان.
- [3]. <http://www.gseis.ucla.edu/courses/ed231c/notes3/probit1.html>
- [4]. Dobson, A,1990, *An Introduction To Generalized Linear Models*, Chapman and Hall, London.
- [5] .McCullagh, P. and Nelder, J. A., 1989, *Generalized Linear Models*, Second Edition, Chapman and Hall, London.
- [6]. Agresti, A., 2002, *An Introduction to Categorical Data Analysis*, John Wiley, New York.
- [7]. Ayatollahi S. M. T., Dowlatabadi E., Ayatollahi S.A.R., 2002, Age at menarche in Iran, *Annals of Human Biology*, 29, 4, 355-362.
- [8]. Iravani K., Heydari S. T., Rosenhall U. L. F., 2005, Effect of Hereditary and Environmental Factors on Presbycusis, in press.
- [9]. SPSS 11.5 Mathsoft Inc. 2002.
- [10]. Minitab 13.2 Minitab Inc. 2000.