

مدل پروبیت و کاربرد آن در داده های پزشکی

سید محمد تقی آیت اللهی^۱
محمد علی وکیلی^{۲،۳}

سعیده پور احمد^۲
تقی حیدری^{۲،۴}

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 برازش مناسب بودن مدل و چگونگی برآورد پارامترها 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

۱. مقدمه

پروبیت^۷ 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.

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 [۳] 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^۵، مهم ترین مدل ها برای متغیرهای وابسته دو حالتی، مدل های لجستیک^۶ 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 نظیر 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 -۶
Probit -۷

۱- استاد

۲- دانشجوی دکتری گروه آمار زیستی و اپیدمیولوژی دانشکده

بهداشت، دانشگاه علوم پزشکی شیراز

۳- مربی، دانشگاه علوم پزشکی گرگان

۴- مربی، دانشگاه علوم پزشکی لرستان

۵- Generalized Linear Model

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

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

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

در سال ۱۹۳۰، چستر بلیس^۲ 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 Φ تابع احتمال تجمعی نرمال استاندارد است. بنابراین

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

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

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

مدل پروبیت در زمینه های مختلف از جمله علوم اجتماعی، زیست شناسی، ژنتیک، سم شناسی و پزشکی کاربردهای فراوان دارد. به خصوص در تحقیقات مربوط به تعیین دوز دارو یا کشندگی 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 π 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 $P(z_1 = 1) = \pi_1$ ، احتمال توأم آنها برابر است با:

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

که عضوی از خانواده نمایی است [۴].

بنابراین اگر علاقه مند به مدل سازی احتمال موفقیت بر اساس متغیرهای مستقل مورد نظر باشیم می توانیم از مدل های 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 [۵].

برای مدل سازی احتمالات π_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، β بردار پارامترها و g تابع ارتباط بین X_i و π_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 π_i یک احتمال است و همواره در بازه $[0, 1]$ قرار می گیرد، مقدار برازش شده $X_i^T \beta$ ممکن است با تغییر مقادیر X_i در خارج از این بازه قرار گیرد. برای اطمینان از اینکه مقدار برازش شده برای π_i در فاصله $[0, 1]$ محدود می شود، باید از توابع ارتباطی استفاده نمود که برد تابع معکوس آنها در محدوده $[0, 1]$ قرار گیرد، یکی از توابع ارتباط، تابع توزیع احتمال تجمعی متغیرهای پیوسته است. بر اساس این تابع ارتباط می توان یک مدل برای π_i بدست آورد.

در حالتی که یک متغیر مستقل داشته باشیم، داریم:

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

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

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

بگیریم، بسط سری تیلور^۵ $\Psi(p)$ حول نقطه $p = \pi$ (احتمال موفقیت) برابر است با:

$$\Psi(p) = \Psi\left(\frac{y}{n}\right) = \Psi(\pi) + \left(\frac{y}{n} - \pi\right) \Psi'(\pi) + o\left(\frac{1}{n^2}\right) \quad (1)$$

از اولین تقریب در معادله (۱) نتیجه می‌شود:

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

چون

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

همچنین

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

$$\downarrow$$

$$[\Psi'(\pi)]^2 \frac{\pi(1-\pi)}{n} \quad \text{var}\left(\frac{y}{n}\right) = \text{var}(p)$$

بنابراین آماره حداقل مربعات وزنی برابر است با

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

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 Ψ 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 ($X = \mu$) دوز کشنده میانه^۱ یا $LD_{(50)}$ گفته می‌شود که برای نشان دادن قدرت کشندگی مواد سمی از جمله حشره کشها، شاخص مناسبی به حساب می‌آید [۴].

۳. روش‌های برآورد پارامترهای مدل

پارامترهای مدل پروبیت را با استفاده از دو روش ماکزیمم درستنمایی (MLE) و حداقل مربعات وزنی، که روش‌های معمول برآورد پارامترها در مدل‌های رگرسیونی با پاسخ دو حالتی هستند، برآورد می‌کنیم [۴و۵].

۱- روش ماکزیمم درستنمایی (MLE): این روش حتی

برای حالتی که $n_i = 1$ یا $y_i = 0$ می‌باشد نیز امکان پذیر است (برخلاف بعضی روش‌ها مثل روش حداقل مربعات). اگر U بردار نمره‌ها (Score vector) و I ماتریس اطلاع فیشر باشد:

$$U_{\%} = \left[\frac{\partial l}{\partial \beta_i} \right] \quad i = 1, 2, \dots, n$$

و

$$I = \left[\frac{\partial^2 l}{\partial \beta_i \partial \beta_j} \right] \quad i, j = 1, 2, \dots, n$$

آن‌گاه با تشکیل معادله زیر و حل آن به روش تکراری می‌توان برآوردهای ماکزیمم 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 ام 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 می‌کنیم که اصطلاحاً همگرایی^۳ حاصل شود. به عبارت دیگر دو تقریب متوالی از برآوردها کاملاً به هم نزدیک شوند در این صورت تقریب نهایی را می‌توان به عنوان برآورد پارامترهای مدل در نظر گرفت.

۲- روش حداقل مربعات وزنی^۴: برای پرهیز از تکرار در

محاسبات، روش حداقل مربعات وزنی بر روش ماکزیمم درستنمایی

۱- Median lethal dose

۲- Maximum Likelihood Estimation

۳- Converge

۴- Weighted Least Square Method

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

$$D = \sum_{i=1}^n o \ln \frac{o}{e}$$

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

۲- روش مینیمم کردن مجموع مربعات وزنی: به جای

استفاده از روش درست‌نمایی می توان از روش کمینه سازی^۳ مجموع مربعات وزنی برای برازش 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^۴ می باشد

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

که در آن o مقادیر مشاهده شده و e مقادیر برازش 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 ، χ^2 ، χ^2_{mod} دارای توزیع χ^2_{N-p} می باشند. شواهد نشان می دهد که χ^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 β 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 $\Psi(\pi_i) = \log \text{it}(\pi_i)$ ، داریم:

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

پس:

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

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

در این مدل نیز برای برآورد پارامترها نیاز به روش تکراری نمی باشد و π_i ها در دامنه صفر و یک قرار می گیرند. کاکس این روش را تبدیل لجستیک کاربردی می نامد و پیشنهاد می کند از $E(z_i - x_i^T \beta)$ برای کاهش اریبی $z_i = \ln \frac{y_i + o/5}{n_i - y_i + o/5}$ استفاده شود. مینیمم مقداری که از (۲) بدست می آید را آماره کای اسکور لجستیک مینیمم می نامیم.

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

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

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.

۴. سنجش برازش^۱ مدل پروبیت

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

۱- استفاده از آماره نسبت لگاریتم درست‌نمایی^۲:

آماره به صورت زیر تعریف می گردد:

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

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

راهنمایی و متوسطه مدارس دولتی، نمونه مردمی و غیر انتفاعی شیراز در سال تحصیلی ۷۷-۱۳۷۶ انجام شده است ([۷]). در این مطالعه با روش نمونه گیری چند مرحله ای 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 می دهد. لازم به ذکر است که متغیر مستقل سن (x) به صورت میانگین سن در هر گروه سنی در نظر گرفته شده و متغیر وابسته (P) (احتمال قاعدگی) از طریق فراوانی نسبی تعداد افراد قاعده شده در هر گروه سنی محاسبه شده است (جدول شماره ۳).

پس از برازش مدل پروبیت به وسیله نرم افزار SPSS و محاسبه ضرائب مدل احتمال قاعدگی در هر گروه سنی با استفاده از تابع ارتباط نرمال $P^* = \Psi(\hat{\alpha} + \hat{\beta}x)$ و مقادیر برازش شده مدل از فرمول $e_i = n_i p_i$ محاسبه و سپس سنجش برازش مدل با آماره χ^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 لوجستیک و پروبیت ارائه و مقایسه شده است. (جدول ۳ و ۴ و ۵)

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

ضمن جوری سازی سن و جنس، عوامل مؤثر بر پیرگوشی در ۱۵۲ فرد بیمار و ۱۵۲ فرد بیمار مورد بررسی قرار گرفتند. با استفاده از مدل پروبیت 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 [۸]، 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 لوجستیک 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 می دهد. همانطور که مشاهده می گردد تفاوت قابل توجهی بین مقادیر برازش شده در هر دو مدل وجود ندارد. در جدول شماره ۷ برآورد MLE پارامترها در مدل های برازش شده به همراه خطای استاندارد آنها آورده شده است. در جدول شماره ۸ مقادیر D ، χ^2 برای سنجش برازش مدل ها آمده

۱- مدل لوجستیک: این مدل بسیار به مدل پروبیت

نزدیک است به گونه ای که گاهی اوقات با مدل پروبیت اشتباه گرفته می شود. در این مدل، که به جای تابع ارتباط توزیع تجمعی نرمال از تابع ارتباط لوجیت استفاده می شود، داریم:

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

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

نمودار دو تابع لوجیت و پروبیت S شکل، برد آنها بین صفر و یک، حول $\pi = 0.5$ متقارن، در بازه $0.18 - 0.82$ تقریباً خطی و اختلاف دو مدل در نقاط ابتدایی و انتهایی مقادیر x می باشد (شکل شماره ۱). مدل لوجستیک بیشتر برای مطالعات مشاهده ای مناسب است در حالی که مدل پروبیت برای مطالعات تجربی به کار می رود. در مطالعاتی که علاقه مند به محاسبه نسبت شانس^۱ برای متغیرهای مستقل باشیم از مدل لوجستیک استفاده می کنیم در حالی که روش پروبیت برای تعیین مقادیر کلیدی و مؤثر متغیر مستقل که $\hat{\pi}_i$ دارای مفهوم خاصی برای ما باشد مثل دوز کشنده میانه یعنی دوزی که موجب مرگ و میر حداقل ۵۰ درصد حیوانات آزمایشگاهی شود (LD_{50}) ([۲]).

۲- مدل لگاریتم دوگانه مکمل^۲: تابع ارتباط دیگری که

می تواند جایگزین مدل پروبیت شود تابع لگاریتم دوگانه مکمل می باشد که بصورت زیر تعریف می شود:

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

این مدل نیز وقتی که π به 0.5 نزدیک باشد شبیه مدل پروبیت و لوجستیک است، اما هر چه π به صفر یا یک نزدیک تر می شود از دو مدل دیگر فاصله می گیرد و برای مواردی استفاده می شود که پراکندگی بین داده ها زیاد باشد یا مقادیر افراطی در داده ها موجود باشد.

۶. کاربرد مدل پروبیت در داده های پزشکی و مقایسه آن با مدل لوجستیک

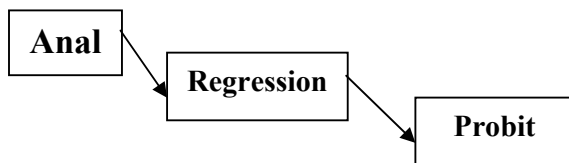
مثال ۱: در یک بررسی به منظور برآورد احتمال قاعدگی در گروههای سنی مختلف، مطالعه ای بر روی دانش آموزان دختر مقاطع

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 برازش می کند، برای برازش 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 برازش مدل بدست آوریم مثل مقادیر باقیمانده ها، نمودارهای برازش 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 گزینه Normal را انتخاب می کنیم، گزینه Inverse cumulative probability را انتخاب نموده در قسمت optional strong ستون احتمالات و 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 option عکس تعداد موارد را بعنوان وزن تعریف می کنیم و سپس به منوی اصلی بازگشته و آنالیز را به اتمام می رسانیم.

نرم افزار SPSS

روش MLE: از منوی Analyze گزینه 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 فاکتور مورد 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

است که مناسب بودن برازش مدل در هر دو مدل تأیید می شود. (جدول ۷ 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 ، χ^2 بهره گرفتیم که هر دو مقدار آماره های پیش گفت در هر دو مدل، بیانگری برازش مناسب داده ها می باشد (جدول شماره ۹).

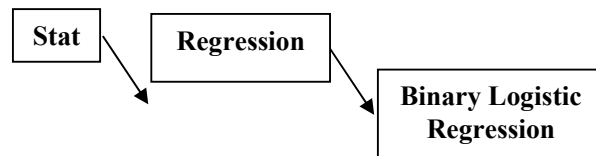
در بررسی رابطه فرسودگی شغلی با متغیرهای مستقل با استفاده از مدل پروبیت و با اثرات اصلی، یافته ها بیانگر آن است که فرسودگی شغلی با وضعیت تأهل با اختلاف معنی دار می باشد، ولی نسبت به سایر متغیرها مستقل باشد، بدین ترتیب که فرسودگی شغلی در متأهلین نسبت به مجردین بیشتر می باشد، که این نتایج با یافته های حاصله 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.



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 عمده 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 لوجیت 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 برآزش مدل لگاریتم دوگانه مکمل بر داده های ابتدایی و انتهایی 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 0 و 1 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 نا پیوسته، امکان محاسبه برآورد پارامترها به روش WLS نبوده است.

۸. نتیجه گیری

کاربرد برجسته 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 برآزش می کنند، ولی بدیهی است که دو نگرش مختلف از نحوه پراکندگی داده ها بر اساس احتمالات مربوطه (نمودار ۱) ایده اصلی استفاده از

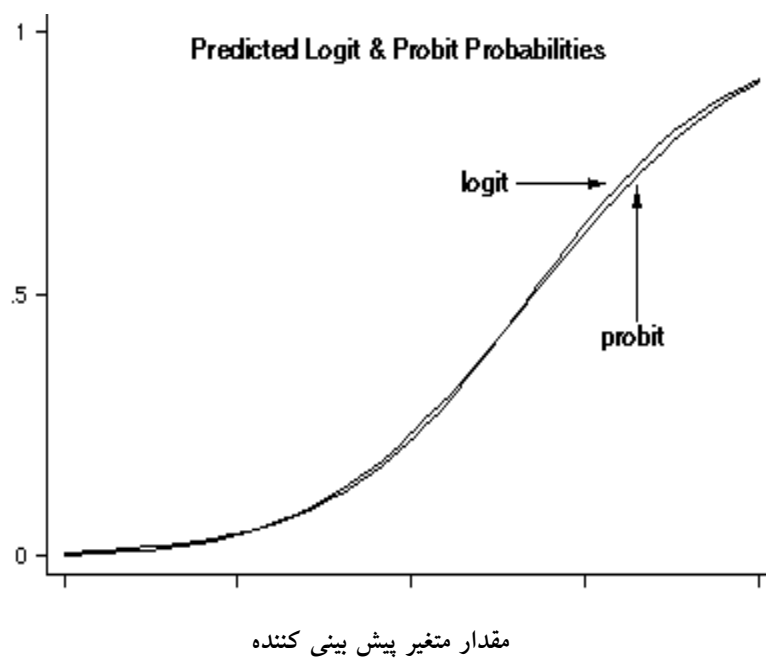
جدول ۱. مدل های حداقل مربعات وزنی معمول برای داده های دو حالتی

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

جدول ۲. فراوانی N جامعه دارای توزیع دوجمله ای

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 برازش شده قاعدگی 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	۰/۰۸۳۳	۰/۰۸۶۲
۱۱/۵-۱۲	۱۱/۷۵	۱۵۷	۲۰	۰/۱۲۷۴	۰/۱۶۷۴
۱۲-۱۲/۵	۱۲/۲۵	۱۳۱	۳۸	۰/۲۹۰۱	۰/۲۸۶۲
۱۲/۵-۱۳	۱۲/۷۵	۲۵۳	۱۱۲	۰/۴۴۲۸	۰/۴۳۴۵
۱۳-۱۳/۵	۱۳/۲۵	۱۳۵	۸۶	۰/۶۳۷۰	۰/۵۹۲۹
۱۳/۵-۱۴	۱۳/۷۵	۲۵۳	۱۸۷	۰/۷۳۹۱	۰/۷۳۷۲
۱۴-۱۴/۵	۱۴/۲۵	۲۰۳	۱۷۹	۰/۸۸۱۸	۰/۸۴۹۶
۱۴/۵-۱۵	۱۴/۷۵	۱۴۸	۱۳۱	۰/۸۸۵۱	۰/۹۲۴۳
۱۵-۱۵/۵	۱۵/۲۵	۱۷۳	۱۶۸	۰/۹۷۱۱	۰/۹۶۶۷
۱۵/۵-۱۶	۱۵/۷۵	۳۰	۲۹	۰/۹۶۶۷	۰/۹۸۷۳
۱۶-۱۶/۵	۱۶/۲۵	۱۷۲	۱۷۰	۰/۹۸۸۴	۰/۹۹۵۸
۱۶/۵ و بیشتر	۱۶/۷۵	۱۴۷	۱۴۷	۱	۰/۹۹۸۸

جدول ۴. مقایسه مقادیر برازش شده مدل پروبیت و لجستیک

گروههای سنی	مقادیر مشاهده شده	مقادیر برازش شده مدل پروبیت	مقادیر برازش 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 برازش مدل پروبیت و لجستیک

مدل لجستیک			مدل پروبیت			ضرایب
P-value	خطای معیار	مقدار	P-value	خطای معیار	مقدار	
۰/۰۰۰	۰/۸۹۸	-۱۸/۱۶۶	۰/۰۰۰	۰/۴۷	-۱۰/۳۶۰	MLE
۰/۰۰۰	۰/۰۹۳	-۱۸/۳۳۸	۰/۰۰۰	۰/۰۶	-۱۰/۵۴۳	WLS
۰/۰۰۰	۰/۰۶۸	۱/۴۵۳	۰/۰۰۱	۰/۳۶	۰/۷۹۹	MLE
۰/۰۰۰	۰/۰۰۸	۱/۴۱۲	۰/۰۰۰	۰/۰۰۵	۰/۸۱۰	WLS
۰/۶۱۴	df=۱۰	$\chi^2=۸/۱۵۷$	۰/۳۲۶	df=۱۰	$\chi^2=۱۱/۴۱$	

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

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

پرانتر مقدار برازش شده مدل پروبیت و براکت مقدار برازش شده مدل لجستیک

جدول ۷. مقادیر برازش 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 برازش شده پروبیت و لجستیک

آماره	درجه آزادی	پروبیت		لجستیک	
		P-Value	مقدار	P-Value	مقدار
χ^2	۴	۰/۵۰۶	۳/۳۲	۰/۶۰۴	۲/۶۷
D	۴	۰/۵۸۱	۲/۸۶	۰/۶۸۱	۲/۳۰

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

آماره	درجه آزادی	پروبیت		لجستیک	
		P-Value	مقدار	P-Value	مقدار
χ^2	۱۶۵	۰/۲۴۹	۱۷۶/۹۰۹	۰/۲۵۳	۱۷۶/۶۸۹
D	۱۶۵	۰/۱۰	۱۸۸/۶۱۳	۰/۱۰	۱۸۸/۶۱۲

جدول ۱۰. مقادیر برازش 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 تاهل
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 مراکز 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 Humman 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.