Subject: $y = XB + \varepsilon \rightarrow E(y|X) = XB$ $P = XB$

Year. Month. Date. ( )

$y|x \sim ber(p) \rightarrow p^{y}(1-p)$ $y\ln p + (1-y)\ln(1-p)$

$y|x \sim N(XB, \sigma^2 I)$ $e^{y \ln \frac{p}{1-p} + \ln(1-p)}$

جلسه دهم مدل‌های خطی تعمیم یافته فضل چهارم

$g(E(y|x)) = XB$

هدف از ارائه معادلات و طرح روش‌های تحلیل پیوند بررسی اثر متغیرهای تبیینی بر روی متغیرهای پاسخ. هدف از این فصل استفاده از مدل آماری مناسب برای تحلیل داده‌ها و ارتباط بین آن‌ها است.

مدل‌های خطی تعمیم یافته:

مدل‌های خطی تعمیم یافته که از این به بعد آن را با GLM (Generalized linear models) نشان می‌دهیم به وسیله سه مؤلفه زیر مشخص می‌شوند:

Y ۱) مؤلفه تصادفی Random component

X ۲) مؤلفه سیستماتیک systematic component

g ۳) تابع ربط (پیوند) link function

نکته: خانواده GLM شامل دامنه گسترده‌ای از مدل‌ها مانند رگرسیون معمولی و مدل‌های ANOVA برای متغیرهای پاسخ پیوسته نیز هستند در این درس GLM برای داده‌های رسته‌ای بررسی می‌شود.

مؤلفه تصادفی: این مؤلفه شامل مشاهدات مستقل $y_1, \dots, y_n$ از متغیر تصادفی Y با توزیعی از یک خانواده نمایی طبیعی به صورت زیر است پس منظور از مؤلفه تصادفی GLM، تعیین متغیر پاسخ Y و انتخاب توزیع آماری برای $Y_1, \dots, Y_n$ است.

$$f(y_i, \theta_i) = a(\theta_i) b(y_i) \exp\{y_i Q(\theta_i)\}.$$

که $Q(\theta)$ را پارامتر طبیعی می‌نامند. بسیاری از توزیع‌ها مثل دوجمله‌ای، پواسن، چند جمله‌ای، نرمال، نمایی و ... مثال‌هایی از این خانواده می‌باشند. با توجه به شکل توزیع مناسب انتخاب می‌شود برای مثال برای حالتی که Y دو حالتی است و حجم نمونه ثابت است فرض می‌شود توزیع دوجمله‌ای است.

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \rightarrow e^{-\lambda + x\log\lambda - \log x!} \Rightarrow e^{-\lambda} \times e^{-\log x!} \times e^{x\log\lambda}$$

$$f(x) = \pi^{x}(1-\pi)^{1-x} \rightarrow e^{x\log\pi + (1-x)\log(1-\pi)} = e^{x\ln\frac{\pi}{1-\pi} + \log(1-\pi)}$$

Subject:
Year. Month. Date. ( )

مؤلفه سیستماتیک : مؤلفه سیستماتیک GLM متغیرهای تبیینی را مشخص می‌کند و شامل یک بردار η است که ترکیبی خطی از مجموعه متغیرهای تبیینی می‌باشد یعنی

$$\eta = \underset{\sim}{x}'\beta = \beta_0+\beta_1x_1+\dots+\beta_kx_k$$

$\beta' = (\beta_0, \dots, \beta_k)$ $\quad x' = (x_1, \dots, x_k)$

پس $\underset{\sim}{x}$ یک بردار از متغیرهای تبیینی و β یک بردار از پارامترهای مدل است.
این ترکیب خطی از متغیرها پیشگوی خطی نامیده می‌شود.

نکته متغیرهای تبیینی در GLM می‌توانند از نوع پیوسته، رسته‌ای یا از هر دو نوع باشند.

تابع ربط : ارتباط بین مؤلفه سیستماتیک و مقادیر مورد انتظار مؤلفه‌ی تصادفی را توصیف می‌کند؛ فرض کنید مقدار مورد انتظار Y، $\mu = E(Y)$ باشد به طوری که

$i=1,\dots,n \qquad \mu_i = E(Y_i)$

نرمال

y_1	x_{11}	x_{12}	$\dots$	x_{1k}	μ_1
$\vdots$					$\vdots$
y_n	x_{n1}	x_{n2}	$\dots$	x_{nk}	μ_n

$\underset{\sim}{x}_1$, $\underset{\sim}{x}_n$

$Y_i|x_i \sim N(\beta_0+\beta_1x, \sigma^2)$

$E(Y_i|x_i) = \beta_0+\beta_1x_i$

$$g(\mu_i) = \underset{\sim}{x}_i\beta = \sum_{j=0}^{k} x_{ij}\beta_j$$

که به طور کلی به صورت $g(\mu)=\beta_0+\beta_1x_1+\dots+\beta_kx_k$ می‌توان نوشت $\pi(x)$
با این حالت که مقدار μ با تغییر سطوح متغیر تبیینی تغییر می‌کند.

به g تابع ربط یا پیوندی گویند که تابعی مشتق‌پذیر یکنوا است. ساده‌ترین تابع ربط $g(\mu_i)=\mu_i$ است که به تابع ربط همانی معروف است.

نکته تئوری مدل لگ خطی و لوجیت :
مدلهای دیگری نیز وجود دارد شامل $g(\mu)=\log(\mu)$ به GLM ها با ربط لگاریتمی

P4PCO

استفاده کند مدل لگ خطی گویند (برای مقادیر غیرمنفی μ)

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

یا اگر $g(\mu) = \log(\mu/(1-\mu))$ باشد تابع ربط لوجیت نامیده می‌شود و مدل

$= \log\mu - \log(1-\mu)$

GLM ای که از ربط لوجیت استفاده می‌کند مدل لوجیت گویند. (ابزار مواقعی که μ بین صفر و یک است)

مثلاً در برنولی μ، احتمالات

نکته: اگر $g(\mu_i) = Q(\theta_i)$ یعنی اگر تابع ربط پارامتر طبیعی توزیع‌ها باشد

$f(y) = \pi^y (1-\pi)^{1-y}$ $\quad e^{y\log\pi + (1-y)\log(1-\pi)}$ ربط کانونی نامیده می‌شود

$\mu = E(y) = \pi$ $\quad e^{y\log\frac{\pi}{1-\pi}} \times e^{\log(1-\pi)}$

Q

انواع مختلف مدلهای خطی تعمیم یافته: انواع GLM در جدول زیر آورده شده است.

مؤلفه تصادفی ۴	تابع ربط ۳	مؤلفه سیستماتیک ۲ (تبیینی)	مدل ۱
نرمال	همانی	پیوسته	رگرسیون
نرمال	همانی	رسته‌ای	تحلیل واریانس
نرمال	همانی	پیوسته ـ رسته‌ای	تحلیل کوواریانس
برنولی	لوجیت	پیوسته ـ رسته‌ای	رگرسیون لوجستیک (داده‌ دودویی)
پواسن	لگاریتم	پیوسته ـ رسته‌ای	لگ خطی (داده شمارشی)
چندجمله‌ای	لوجیت تعمیم یافته	پیوسته ـ رسته‌ای	پاسخ اسمی با چند سطح

در رگرسیون یکی از فرض‌های اساسی نرمال بودن و ثابت بودن واریانس متغیر پاسخ است که در صورتی که داده‌ها نرمال نباشد از تبدیلاتی استفاده و آنها را نرمال می‌کنند اما در عمل یافتن تبدیلی که هم داده‌ها را نرمال کند و هم واریانس ثابت باشد تقریباً ناممکن است و به همین دلیل اکثر مدلهای خطی ساده مثل رگرسیون پس از انجام تبدیل به خوبی برازش نمی‌شوند اما در مدلهای خطی تعمیم یافته (GLMها) انتخاب تابع ربط از انتخاب

$Y \sim ber(\pi)$ $EY = \pi$ $V(Y) = \pi(1-\pi)$

یکی از ایرادات اصلی مدل خطی در مورد متغیر پاسخ دو حالتی ناهمسانی واریانس

$E(Y|X) = \pi = XB$ $\rightarrow V(Y) = XB(1-XB)$

امکان ندارد ثابت باشد چون وابسته به X است و از فردی به فرد دیگر تغییر می‌کند

Subject:
Year.   Month.   Date.   ( )

متغیر تصادفی جدا است و همچنین در برازش GLM با توجه به متغیر تصادفی انتخاب شده از روش های ماکسیمم درستنمایی استفاده می‌شود پس GLM ها تحت تأثیر نرمال بودن قرار نمی‌گیرند.

چون OLS در GLM معتبر نیست ~~چون ثابت بودن واریانس و نرمال بودن~~

در اکثر نرم افزارهای آماری مدل‌های خطی تعمیم یافته به طور کامل قابل برازش هستند.

~~واریانس~~

GLM ها برای پاسخ دودویی:

بسیاری از متغیرهای پاسخ رسته‌ای (Y) شامل دو رسته هستند که آنها را با صفر و یک نشان می‌دهیم. پاسخ دودویی در واقع یک متغیر تصادفی برنولی است به طوری که

$$P(Y=1) = \pi \quad , \quad P(Y=0) = 1-\pi$$

که می‌دانیم این توزیع دارای میانگین $E(Y)=\pi$ و $var(Y) = \pi(1-\pi)$ است

* برای n مشاهده مستقل با پاسخ دودویی با پارامتر π، تعداد موفقیت‌ها دارای توزیع دوجمله‌ای با پارامترهای n و π است. برای متغیرهای تصادفی دودویی با توزیع برنولی:

$$f(y_i ; \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}$$

$$= \underbrace{(1-\pi_i)}_{a(\pi_i)} \exp\left\{y_i \underbrace{\log\left(\frac{\pi_i}{1-\pi_i}\right)}_{Q(\pi_i)}\right\} \qquad y_i = 0,1$$

که نشان می‌دهد f عضو خانواده نمایی طبیعی با پارامتر طبیعی لوجیت π_i است که یک رابط کانونی است.

لگاریتم بخت برای پاسخ ۱ یعنی:

$$g(\pi_i) = Q(\pi_i) = logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x'B$$

$$\rightarrow \pi_i = \frac{e^{(x'B)}}{1+e^{(x'B)}} \quad \text{موفقیت} \quad Y_i \sim ber(\pi_i)$$

بنابراین GLM برای پاسخ دودویی رابطه غیرخطی بین x_i و π_i را بیان می‌کند.

نکته: چون مقدار π می‌تواند با توجه به مقدار مشخص x از X تغییر کند از نماد $\pi(x)$ به جای π استفاده می‌شود. پس GLM برای پاسخ دودویی:

$$g(E(Y)) \leftarrow \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x'B = B_0 + B_1x_1 + \cdots + B_K x_K$$

رگرسیون لوژستیک:

در این نوع GLM، تابع ربط همان پارامتر طبیعی توزیع برنولی لوجیت $\pi(x)$ است یعنی:

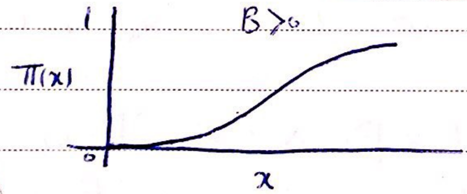
$$logit(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \underline{x}'\underline{\beta}$$

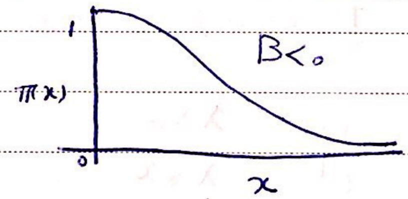
$$\Longrightarrow \pi(x) = \frac{\exp(\underline{x}'\underline{\beta})}{1+\exp(\underline{x}'\underline{\beta})}$$ رگرسیون لوژستیک چندگانه

که به این تابع، رگرسیون لوژستیک می گویند.

پس مؤلفه تصادفی در این مدل برنولی (شکست و موفقیت) است تابع ربط آن نیز لوجیت است. مدلهای رگرسیونی لوژستیک مدلهای لوجیت نیز نامیده می شوند.

پارامتر β_1 نرخ افزایش یا کاهش منحنی را تعیین می کند (در حالتی که فقط یک متغیر تبیینی داریم)





فرض کنید فقط یک متغیر تبیینی داشته باشیم در این صورت

$$\pi(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

نمودار پراکنش

که برای این مدل نسبت موفقیت:

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0+\beta_1 x) = e^{\beta_0}\, e^{\beta_1 x} \rightarrow \text{ص ۵۴}$$

نکته: برای یافتن یک شکل مناسب برای مؤلفه ی سیستماتیک در یک مدل رگرسیون لوژستیک رسم نمودار پراکنش لوجیت های نمونه ای در برابر x مفید است. $\beta_0+\beta_1 x$

تعداد کل در این رسته

$$logit\, y_i = \log\frac{y_i}{n_i-y_i} \longrightarrow$$ لوجیت نمونه ای

تعداد موفقیت در این رسته

x	n	تعداد ۱ ها
1	n_1	y_1
2	n_2	y_2

$i=1 \quad y_1=2 \quad n_1$

$$\frac{y_i/n_i}{1-y_i/n_i}$$

و چون به ازای $y_i = n_i$ و $y_i = 0$ وجود ندارد از لوجیت تجربی استفاده می‌کنیم

لوجیت تجربی $= \log\left(\frac{y_i + 1/2}{n_i - y_i + 1/2}\right)$

نکته ۱: اگر x پیوسته باشد باید قبل محاسبه لوجیت‌های تجربی داده‌ها رده‌بندی شوند.

نکته ۲: مدل‌های ~~غیرخطی~~ دیگر شکل $\pi(x)=F(x'B)$ وجود دارند که $F(x'B)$ در بازه $[0,1]$ تغییر می‌کند. تابع‌های توزیع تجمعی مثال‌هایی از $F(\cdot)$ هستند که دارای این خاصیت می‌باشند. به عبارتی در این موارد GLM کی با تابع ربط F^{-1} داریم. * $\cancel{F^{-1}(x'B)=\pi(x)}$

مدل پروبیت: اگر در مدل $\pi(x)=F(x'B)$ به جای F از Φ (تابع توزیع تجمعی نرمال استاندارد) استفاده شود مدل پروبیت حاصل می‌شود پس در این مدل پروبیت

$$\pi(x) = \Phi(x'B) \xrightarrow[\text{متغیر تبیینی}]{\text{برای یک}} \Phi^{-1}(\pi(x)) = B_0 + B_1 x$$

نکته

$F_X(x) = P(X \le x) \quad -\infty < x < \infty$ تابع توزیع تجمعی — می‌دانیم $0 \le F_X(x) \le 1$

$$P(Y=y) = \begin{cases} 0 & X > x \\ 1 & X \le x \end{cases} \qquad \pi(x) = P(X \le x) = F(x)$$

در این مدل F تابع توزیع تجمعی است و به ازای تمام توزیع‌ها برقرار است. *

نکته: در اغلب مثال‌های کاربردی منحنی‌های برازش لوجستیک و پروبیت بسیار شبیه هستند به طوری که برآوردهای شیب در مدل‌های رگرسیونی لوجستیک تقریباً در حدود ۱٫۶ تا ۲ برابر شیب در مدل‌های پروبیت است و علت این دو مدل ~~غیرخطی~~ در حالتی که سرعت میل کردن $\pi(x)$ به صفر و یک برابر باشد مناسب است. در حالت‌هایی که $\pi(x)$ سرعتی متفاوت در نزدیک شدن به صفر و یک دارد نامناسب می‌باشند.

پروبیت ← $\Phi^{-1}(\pi(x)) = B_0 + B_1 x$

لوجستیک ← $\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = B_0 + B_1 x$ ← لینک

مدل خطی معمولی ← $\pi(x) = B_0 + B_1 x$

Subject:
Year. Month. Date. ( )

$E(Y|X) = B_0 + B_1 X \longleftarrow Y \sim N(B_0 + B_1 x, \sigma^2)$

$Y = \begin{cases} 1 & \pi \\ 0 & 1-\pi \end{cases} \longrightarrow E(Y|X) = \pi$

ص ۱۰۵ $\pi(x) = B_0 + B_1 x$

مثال: بررسی ارتباط بین خرناس کشیدن و بیماری قلبی. این مطالعه بر روی 2484 نفر انجام شده است. $Y=1$ (بیماری قلبی)

برای یافتن مدل به x امتیاز (0, 2, 4, 5) بدهید.

X (خرناس کشیدن)	بله	خیر
0 هرگز	24	1355
2 گاه‌گاهی	35	603
4 تقریباً هر شب	21	192
5 هر شب	30	224

به عنوان تمرین برآورد کنید — جزوه

$E(Y|X) = \pi(x)$

سه مدل مد نظر است:

① مدل احتمال خطی $\pi_1(x) = B_0 + BX$

(چون احتمال موفقیت به طور خطی با تغییر x تغییر می‌کند)

مدل احتمال ناخطی:

② $logit(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ یا $\pi_2(x) = \frac{\exp(B_0 + B_1 x)}{1 + \exp(B_0 + B_1 x)}$

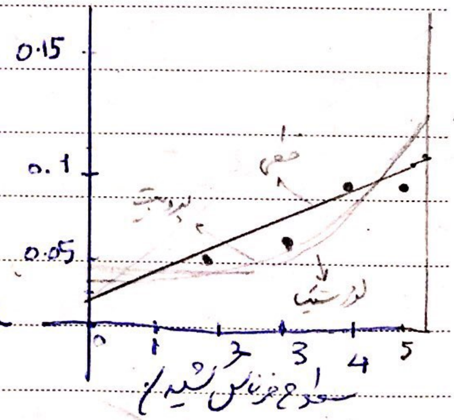
③ $probit(\pi(x))$ $\pi_3(x) = \Phi(B_0 + B_1 x)$

پارامترها با نرم افزار برآورد و در روش ML محاسبه می‌شوند (برآوردگر ML در GLM اغلب به صورت جواب بسته قابل محاسبه نیستند و از روش‌های عددی استفاده می‌شود)

$\hat{\pi}_1 = 0.0172 + 0.0198x$

$logit(\hat{\pi}_2) = -3.87 + 0.4x$

$probit(\hat{\pi}_3) = -2.061 + 0.188x$



نکته: اگر امتیاز انتخاب شده سطوح خرناس کشیدن با نمره‌های مختلف به غیر از {0, 2, 4, 5} در نظر گرفته شود مقادیر برازش شده در چهار رده خرناس کشیدن تغییر خواهد کرد.

Complementry-log-log — مدل لگ - لگ مکمل

در مدل لگ - لگ مکمل تابع رابط به صورت $\log(-\log(1-\pi(x)))$ است یعنی مدل به صورت

$$\log\{-\log(1-\pi(x))\} = x'B \Rightarrow -\log(1-\pi(x)) = e^{x'B}$$

$$\rightarrow \log(1-\pi(x)) = -e^{x'B}$$
$$\rightarrow (1-\pi(x)) = \exp\{-\exp(x'B)\}$$

است یا به طور معادل

$$\pi(x) = 1-\exp\{-\exp(x'B)\}$$

اگر فقط یک متغیر تبیینی داشته باشیم

$$\log\{-\log(1-\pi(x))\} = B_0 + B_1 x \rightarrow \pi(x) = 1-\exp\{-\exp(B_0+B_1 x)\}$$

اگر $B_1 < 0$ مدل برای $\pi(x)$ نزولی و به ازای $B_1 > 0$ مدل صعودی است.

تفسیر مدل: فرض کنید x دو مقدار x_1 و x_2 را می‌گیرد:

$$\log\{-\log(1-\pi(x_2))\} - \log\{-\log(1-\pi(x_1))\} = B_1(x_2-x_1)$$

$$\Rightarrow \log\frac{\log(1-\pi(x_2))}{\log(1-\pi(x_1))} = B_1(x_2-x_1)$$

$$\Rightarrow \frac{\log(1-\pi(x_2))}{\log(1-\pi(x_1))} = \exp\{B_1(x_2-x_1)\}$$

$$\log(1-\pi(x_2)) = \{\log(1-\pi(x_1))\}\exp\{B_1(x_2-x_1)\}$$

$$\Rightarrow (1-\pi(x_2)) = (1-\pi(x_1))^{\exp\{B_1(x_2-x_1)\}}$$

پس احتمال شکست در x_2 برابر است با احتمال شکست در x_1 به توان e^{B_1} به ازای هر یک واحد اختلاف در فاصله بین x_1 و x_2.

نکته: در مدل لگ - لگ مکمل سرعت رسیدن $\pi(x)$ به 1 از سرعت رسیدن $\pi(x)$ به صفر بیشتر است.

نکته: در مدل لگ - لگ مکمل یعنی $\log\{-\log(1-\pi(x))\} = x'B$

$$\Rightarrow \pi(x) = 1-\exp\{-\exp(x'B)\} \Rightarrow \pi(x) = 1-\exp\{-\exp(B_0+B_1x)\}$$

اگر $B_1<0$ مدل برای $\pi(x)$ نزولی و به ازای $B_1>0$ مدل صعودی است.

نکته ۲: اگر $\pi(x)=F(x'B)$ در نظر بگیریم و F را تابع توزیع گامبل در نظر بگیریم مدل لگ - لگ مکمل بدست می آید

توزیع گامبل $G(x)=\exp\{-\exp(-\frac{x-a}{b})\}$ $b>0$ $-\infty<a<\infty$

$E(X)=a+0.577b$ $\sqrt{v(x)}=\frac{\pi b}{\sqrt{6}}$

مدل لگ - لگ به صورت زیر تعریف می شود

$$\log\{-\log\pi(x)\} = x'B \Rightarrow \pi(x)=\exp\{-\exp(x'B)\}$$

$$\Rightarrow e^{x'B} = -\log\pi(x)$$

و اگر یک متغیر پیشبین داشته باشیم

$$\pi(x) = \exp\{-\exp(B_0+B_1x)\} \Rightarrow \log(-\log(\pi(x)) = B_0+B_1x$$

یک GLM با تابع ربط لگ - لگ

در مدل لگ لگ بر عکس لگ لگ مکمل سرعت رسیدن $\pi(x)$ به صفر بیشتر از سرعت رسیدن به یک است و هنگامی که $B_1>0$ باشد منحنی $\pi(x)$ نزولی و وقتی $B_1<0$ منحنی صعودی خواهد بود.

مدل لگ - لگ را می توان به صورت $\pi(x)=F(x'B)$ تعریف کرد که F حالت خاصی از توزیع گامبل با $a=0$ و $b=1$ است.

استنباط و بررسی مدل:

همان طور که گفتیم در اغلب GLM ها با پاسخ های رسته ای، محاسبه برآوردهای ML با رابطه های بسیار پیچیده هستش و از روشهای عددی استفاده می شود. در این قسمت دو روش درستنمایی و شبه درستنمایی به طور خلاصه بیان می شود که اغلب حل معادلات با نرم افزار امکان پذیر است.

روش درستنمایی:

$$Y \qquad \begin{array}{l} y_1 \quad x_1 \to Y_1 \sim ber(\pi_1) \\ \vdots \\ y_n \quad x_n \to Y_n \sim ber(\pi_n) \end{array}$$

Subject : Date ....

فرض کنید مدل لوجیت مدنظر باشد پس داریم

برای n مشاهده و یک متغیر تبیینی:

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \cancel{x'\beta} = B_0 + B_1 x_i$$

$$\underset{(B_0, B_1)}{L(B; y, x)} = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1-\pi(x_i)]^{1-y_i}$$

$$= \prod_{i=1}^{n} \left[(1-\pi(x_i)) \left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i} \right]$$

به صورت نمایی می‌نویسیم

$$= \prod_{i=1}^{n} \exp\left\{ \log(1-\pi(x_i)) + y_i \log \frac{\pi(x_i)}{1-\pi(x_i)} \right\} \quad (1)$$

از طرفی می‌دانیم

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = B_0 + B_1 x_i$$

$$\Rightarrow \exp\{B_0 + B_1 x_i\} = \frac{\pi(x_i)}{1-\pi(x_i)} \Rightarrow$$

پس داریم:

$$\pi(x_i) = \exp\{B_0 + B_1 x_i\} - \pi(x_i) \exp\{B_0 + B_1 x_i\}$$

$$\Rightarrow \pi(x_i) = \frac{\exp\{B_0 + B_1 x_i\}}{1 + \exp\{B_0 + B_1 x_i\}} \Rightarrow 1 - \pi(x_i) = (1 + \exp(B_0 + B_1 x_i))^{-1} \quad (2)$$

در نهایت از معادلات (1) و (2):

$$L(B; y, x) = \exp\left\{ \sum_{i=1}^{n} \left[\log(1-\pi(x_i)) + y_i \log \frac{\pi(x_i)}{1-\pi(x_i)} \right] \right\}$$

$$\stackrel{(2)}{=} \exp\left\{ \sum_{i=1}^{n} \left[-\log(1 + \exp(B_0 + B_1 x_i)) + y_i (B_0 + B_1 x_i) \right] \right\}$$

و اگر از معادله فوق لگاریتم بگیریم:

$$\ell(B) = \sum_{i=1}^{n} \underbrace{[y_i (B_0 + B_1 x_i)]}_{B_0 y_i + B_1 x_i y_i} - \sum_{i=1}^{n} [\log(1 + \exp(B_0 + B_1 x_i))]$$

$$\partial B_0 = y_i \qquad \partial B_0 \to \frac{\exp(B_0 + B_1 x_i)}{1 + \exp(B_0 + B_1 x_i)} = \pi(x_i)$$

$$\partial B_1 = x_i y_i \qquad \partial B_1 \to \frac{x_i \exp(B_0 + B_1 x_i)}{1 + \exp(B_0 + B_1 x_i)} = x_i \pi(x_i)$$

کافیست نسبت به B_0 و B_1 مشتق بگیریم:

معادلات نرمال

$$\begin{cases} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(x_i) = \sum_{i=1}^{n} (y_i - \pi(x_i)) = 0 \\ \sum_{i=1}^{n} (x_i y_i - x_i \pi(x_i)) = \sum_{i=1}^{n} x_i (y_i - \pi(x_i)) = 0 \end{cases}$$

MICRO

Subject : Date ....

مثال: فرض کنید مدل خطی تعمیم یافته $logit(\pi(x)) = \beta_0 + \beta_1 x$ با پاسخ دودویی مد نظر است و جدول زیر بدست آمده است. برآوردگر β_0 و β_1 را بدست آورید و $H_0: \beta_1 = 0$ را آزمون کنید.

	x	y = 0	y = 1	n
x_1	0	۵۰	۵۰	$n_1 \leftarrow 100$
x_2	1	٦۰	۴۰	$n_2 \leftarrow 100$

$n = 200 \leftarrow$

$$\pi(x_1) = \frac{\exp\{\beta_0 + \beta_1 x_1\}}{1 + \exp\{\beta_0 + \beta_1 x_1\}} \xRightarrow{x_1 = 0} \pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$\pi(x_2) = \pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$$

از معادلات نرمال

$$\begin{cases} \sum_{i=1}^{200} [y_i - \pi(x_i)] = 0 & (1) \\ \sum_{i=1}^{200} x_i [y_i - \pi(x_i)] = 0 & (2) \end{cases}$$

از معادله اول:

$$\underbrace{\sum_{i=1}^{100} y_i - \sum_{i=1}^{100} \pi(0)}_{\text{در سطح اول } x=0} + \underbrace{\sum_{i=1}^{100} y_i - \sum_{i=1}^{100} \pi(1)}_{\text{در سطح دوم } x=1} = 0$$

$$50 - 100\pi(0) + 40 - 100\pi(1) = 0 \Rightarrow \pi(0) + \pi(1) = \frac{90}{100} \quad (3)$$

از معادله دوم

$$\underbrace{\sum_{i=1}^{100} x_i y_i}_{\substack{0 \\ x=0}} - \sum_{i=1}^{100} x_i \pi(0) + \underbrace{\sum_{i=1}^{100} x_i y_i}_{40} - \underbrace{\sum_{i=1}^{100} x_i \pi(1)}_{100}$$

$$= +40 - 100\pi(1) = 0$$

$$\Rightarrow \hat{\pi}(1) = 0/4$$

در (۳) $\hat{\pi}(0) = 0/5$

$$\hat{\pi}(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} = 0/5 \Rightarrow \hat{\beta}_0 = 0$$

$$\hat{\pi}(1) = \frac{e^{0 + \beta_1}}{1 + e^{0 + \beta_1}} = 0/4 \Rightarrow \frac{1}{1 + e^{-\beta_1}} = 0/4 \Rightarrow e^{-\beta_1} = 3/2$$

$$\Rightarrow \hat{\beta}_1 = \log(2/3)$$

Subject :   Date ....

برای بررسی فرض $H_0: B_1=0$

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = B_0 + B_1 x$$

تحت موفقیت رسوال

$$\theta = \frac{\pi(0)/1-\pi(0)}{\pi(1)/1-\pi(1)} = \frac{e^{B_0}/1+e^{B_0}}{e^{B_0}/1+e^{B_0}} = 1 \quad : H_0 \text{ تحت}$$

تحت موفقیت در سوال دوم

پس $H_0: B_1=0$ معادل است با $H_0: \theta=1$ و می‌توان از آزمون X^2 یا G^2 استفاده کنید

روش شبه درستنمایی :

شبه درستنمایی، اولین بار توسط ودربرن (1974) و سپس مک‌کولاه (1983) ارائه شده است. روش شبه درستنمایی روشی است برای برآوردکردن پارامترهای رگرسیونی است که تفاوت آن با درستنمایی در این است که در درستنمایی شکل توزیع باید مشخص باشد ولی در شبه درستنمایی فقط روابط بین میانگین و متغیرهای کمکی و رابطه بین میانگین و واریانس باید معلوم باشد (به عبارت دیگر در شبه درستنمایی فقط فرض‌های کمکی درباره توزیع متغیر پاسخ مورد نیاز است و یعنی مثلاً بدانیم مدل لوجیت است. (تابع ربط مشخص باشد و نه دقیقاً توزیع)

$$l(\beta) = \sum_{i=1}^{n} \underbrace{\left[y_i \overbrace{\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)}^{\theta_i} + \overbrace{\log(1-\pi(x_i))}^{-\log(1+\exp(\theta_i))} \right]}_{l_i}$$

در درستنمایی دیدیم لوجیت

$\rightarrow l_i$

$\theta_i = \text{logit}(\pi(x_i))$ فرض شد $= \log \frac{\pi(x_i)}{1-\pi(x_i)}$

$$\log \frac{\pi(x_i)}{1-\pi(x_i)} = \theta_i \Rightarrow \exp(\theta_i) = \frac{\pi(x_i)}{1-\pi(x_i)} \Rightarrow 1+\exp(\theta_i) = \frac{1}{1-\pi(x_i)}$$

$$\pi(x_i) = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} \qquad \Rightarrow \log(1-\pi(x_i)) = -\log(1+\exp(\theta_i)) \quad (\star)$$

برای ساده‌نویسی $\pi(x_i) = \pi_i$ و از قاعده زنجیره‌ای مشتق‌گیری داریم تراناده

$$\frac{\partial l_i}{\partial \beta} = \left(\frac{\partial \pi_i}{\partial \beta}\right)' \cdot \frac{\partial \theta_i}{\partial \pi_i} \cdot \frac{\partial l_i}{\partial \theta_i} \quad (\star\star)$$

بردار

$$\frac{\partial \theta_i}{\partial \pi_i} = \frac{\partial \log \pi_i/1-\pi_i}{\partial \pi_i} = \frac{1/(1-\pi_i)^2}{\pi_i/1-\pi_i} = \frac{1}{\pi_i(1-\pi_i)} = \nu_i^{-1} \Rightarrow var(y_i) = \nu_i$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left[\theta_i y_i - \underbrace{\log(1+\exp(\theta_i))}_{\star}\right] = y_i - \underbrace{\frac{e^{\theta_i}}{1+e^{\theta_i}}}_{\pi_i}$$

$$(\star\star) \quad \frac{\partial l_i}{\partial \beta} = \left(\frac{\partial \pi_i}{\partial \beta}\right)' \nu_i^{-1} \{y_i - \pi_i\}$$

و در نهایت جواب‌های رگرسیون از حل معادلات :

$$S(\beta)=\sum_{i=1}^{n}\left(\frac{\partial \pi_i}{\partial \beta}\right)' \nu_i^{-1}\{y_i-\pi_i\}=0 \quad \circledast$$

به دست می‌آید که فقط به $E(y_i)=\pi_i$ و $Var(y_i)=\nu_i$ بستگی دارد
$=\pi_i(1-\pi_i)$

و در این ادعا که از $\circledast$ $S(\beta)$ برای هر نوع تابع ربط می‌توان استفاده کرد و فقط نیاز به فرض‌هایی درباره تابع ربط و واریانس داریم و فرض توزیع احتمالی لازم نیست.

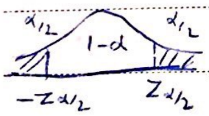
نکته: مکولا (1983) نشان داد جواب معادله شبه درستنمایی ($\hat{\beta}$) با تابع ربط لوجیت دارای توزیع مجانبی نرمال با میانگین β و واریانس زیر است.

$$V=\left\{\sum_{i=1}^{n}\left(\frac{\partial \pi_i}{\partial \beta}\right)'\nu_i^{-1}\left(\frac{\partial \pi_i}{\partial \beta}\right)\right\}^{-1}\times\left\{\sum_{i=1}^{n}\left(\frac{\partial \pi_i}{\partial \beta}\right)'\nu_i^{-1}\{y_i-\pi_i(\beta)\}^{2}\nu_i^{-1}\left(\frac{\partial \pi_i}{\partial \beta}\right)\right\}^{-1}\cdot\left\{\sum_{i=1}^{n}\left(\frac{\partial \pi_i}{\partial \beta}\right)'\nu_i^{-1}\left(\frac{\partial \pi_i}{\partial \beta}\right)\right\}^{-1}$$

فاصله اطمینان : $\beta:\ \hat{\beta}\pm Z_{\alpha/2}\underbrace{\sqrt{V}}_{ASE}$ $\qquad Z=\frac{\hat{\beta}-\beta}{ASE}\sim N(0,1)$

آزمون والد، نسبت درستنمایی و امتیاز:

معمولاً ۳ روش برای انجام آزمون $H_0:\beta=0$ در مدل GLM وجود دارد که ساده‌ترین آن استفاده از آماره زیر است:



$$Z=\frac{\hat{\beta}}{ASE=\sqrt{V}}$$

برای حجم نمونه بزرگ

که تحت فرض صفر دارای توزیع نرمال استاندارد است و می‌توان از جدول نرمال استاندارد استفاده کرد و در سطح α آزمون را براساس فرض مقابل که یکطرفه باشد یا دوطرفه انجام داد.

$|Z|>Z_{\alpha/2}\to RH_0$

به طور معادل می‌توان از Z^2 استفاده کرد که دارای توزیع χ^2_1 است

به آماره Z^2، آماره والد گویند.

دومین روش آزمون از نسبت دو ماکسیمم تابع درستنمایی استفاده می‌شود. در آماره آزمون نسبت درستنمایی استفاده می‌شود.

۱) ماکسیمم بر روی مقادیر ممکن پارامتر با شرط برقراری فرض صفر.

۲) ماکسیمم بر روی مجموعه‌ای بزرگتر بدون در نظر گرفتن فرض صفر یا فرض مقابل

پس فرض کنید L_1 مقدار ماکسیمم تابع درستنمایی برای مدل کامل و L_0 مقدار ماکسیمم درستنمایی تحت فرض صفر باشد. همواره $L_1 \geq L_0$ است. آماره آزمون نسبت درستنمایی برابر است با

$$-2\log(L_0/L_1) = -2[\log L_0 - \log L_1]$$

$$= -2(l_0 - l_1)$$

که می‌دانیم دارای توزیع χ^2_1 است.

سومین آزمون که به آماره امتیاز کارا مشهور است و دارای توزیع کای‌دو است در اغلب نرم‌افزارها محاسبه می‌شود در این درس مطرح نمی‌شود.

نکته: هر سه آماره برای حجم نمونه بزرگ دارای توزیع χ^2 هستند و رفتار یکسانی دارند اما برای نمونه‌های کوچک استفاده از آزمون نسبت درستنمایی قابل اعتمادتر است.

باقیمانده‌ها در مدل‌های لوجیت:

$n_i\pi_i = E(y_i)$

فرض کنید y_i معرف تعداد موفقیت‌ها در iامین موقعیت در n_i آزمایش باشد

داده برنولی

پس می‌توان باقیمانده‌ها را در یک GLM با مؤلفه تصادفی دوجمله‌ای به صورت زیر تعریف کرد که باقیمانده پیرسن در موقعیت iام می‌گویند.

$$r_{P_i} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

Subject:

Year. Month. Date. ( )

که $\sum_i r_{p_i}^2 = \chi^2$ را آماره پیرسن می‌نامند. اگر n_i ها بزرگ باشد

r_{p_i} ها دارای توزیع تقریبی نرمال هستند و اگر مدل برازش شده مناسب باشد با حجم نمونه بزرگ دارای امید ریاضی صفر و واریانس کوچکتر از یک هستند.

(در مقابل متغیر کراندار)

معمولاً برای بررسی یک مدل مناسب مانده‌ی خام یعنی $e_i = y_i - n\hat{\pi}_i$ رسم می‌شوند انتظار داریم روندی نداشته باشد. ولی چون این مانده‌ها استاندارد نشده نیستند برای تشخیص نقاط دورافتاده خیلی نمی‌توان از آنها استفاده کرد. مانده‌ی استاندارد شده را معمولاً از روی مانده‌ی پیرسن و به صورت زیر محاسبه می‌کنند

$$r^*_{p_i} = \frac{r_{p_i}}{\sqrt{1-H_{ii}}}$$

که در آن H_{ii} درایه‌ی روی قطر ماتریس کلاه هستند که به صورت زیر تعریف می‌شود

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}$$

که در آن W ماتریس قطری با عناصر $\frac{\left(\frac{\partial \pi_i}{\partial \eta_i}\right)^2}{var(y_i)}$ است و η_i پیشگو

برای شکل دودویی $\eta_i = \ln \frac{\pi_i}{1-\pi_i} = \sum x'\beta$ است.

پس برای تشخیص نقاط دورافتاده مقادیر $r^*_{p_i}$ محاسبه و قدر مطلق بزرگتر از ۳ نقاط دورافتاده را مشخص می‌کند.

ص ۴۹ تحقیق

r_{p_i}

مقایسه مدل‌ها :

برای مقایسه دو مدل می‌توان از آماره نسبت درستنمایی استفاده کرد که از اختلاف کیش‌ها بدست می‌آید. فرض کنید دو مدل M_0 و M_1 می‌خواهند مقایسه شوند

$$\begin{cases} H_0 : M_0 \\ H_1 : M_1 \end{cases}$$

ابتدا فرض کنید می‌خواهیم مدل M_1 را بررسی کنیم:

فرض کنید $l(\mu ; y)$ را نشان‌دهنده لگاریتم درستنمایی برحسب $\mu = (\mu_1 , \dots , \mu_n)'$ برای یک بردار داده $y' = (y_1 , \dots , y_n)$ باشد و $l(\hat{\mu}_{M_1} ; y)$ لگاریتم درستنمایی ماکسیمم تحت فرض مناسب بودن مدل M_1 باشد (یعنی یک مدل مناسب M_1 درنظر گرفتیم و $\hat{\mu}$ را برآورد کردیم). و لگاریتم درستنمایی ماکسیمم برای مدلی که تعداد پارامترها برابر تعداد مشاهدات است (مدل اشباع شده) برابر $l(y ; y)$ باشد که با جایگذاری $\hat{\mu}_i = y_i$ در $l(\mu ; y)$ بدست آمده است در این صورت کیش مدل M_1 به صورت زیر تعریف می‌شود:

$$D(M_1) = 2\left[\underbrace{l(y ; y)} - \underbrace{l(\hat{\mu}_{M_1} ; y)}\right] \quad \text{Devince}$$

$$= -2\ln\Lambda = ?$$

$y_i \sim bin(n_i , \pi_i) \quad i = 1 , \dots , \overset{\text{رسته}}{m}$ (p تعداد پارامتر) $\qquad f(y_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$

$$l(\mu ; y) = k + \sum\left[y_i \ln \pi_i + (n_i - y_i)\ln(1-\pi_i)\right]$$

تحت مدل M : $\hat{\mu}_i = n_i \hat{\pi}_i \Rightarrow \hat{\pi}_i = y_i / n_i$

تحت مدل اشباع شده $\hat{\mu}_i = n_i \underbrace{\hat{\pi}_i}_{y_i/n_i} = n_i \frac{y_i}{n_i} = y_i$

۲۸-۱

Subject:

Year. Month. Date. ( )

$$2[l(y;y) - l(\hat{\mu}_{M_1}, y)] = 2\Big[2\sum_{i=1}^{m} y_i \ln y_i - 2\sum_{i=1}^{m} y_i \ln \hat{y}_i$$

$$+ 2\sum_{i=1}^{m} (n_i - y_i)\ln(n_i - y_i) - 2\sum_{i=1}^{m} (n_i - y_i)\ln(n_i - \hat{y}_i)\Big]$$

$$\sim \chi^2_{m-p}$$ (m رسته با p پارامتر موجود در M_1)

(وقتی M_1 برقرار است یعنی تحت M_1
مدل دارای توزیع مجانبی χ^2 است)

$$l(y;y) = \sum\Big[y_i \ln \frac{y_i}{n_i} + (n_i - y_i)\ln\Big(1 - \frac{y_i}{n_i}\Big)\Big]$$

$$= \sum\Big[y_i \ln y_i - y_i \ln n_i + (n_i - y_i)\ln(n_i - y_i)$$

$$- (n_i - y_i)\ln n_i\Big] = \sum\Big[y_i \ln y_i - (n_i - y_i)\ln(n_i - y_i)$$

$$- n_i \cancel{\ln} n_i\Big]$$

$$l(\hat{\mu}_{M_1};y) = \sum\Big[y_i \ln \frac{\hat{y}_i}{n_i} + (n_i - y_i)\ln\Big(1 - \frac{\hat{y}_i}{n_i}\Big)\Big]$$

$$= \sum\Big[y_i \ln \hat{y}_i - y_i \ln n_i + (n_i - y_i)\ln(n_i - \hat{y}_i)$$

$$- (n_i - y_i)\ln n_i\Big]$$

$$= \sum\Big(y_i \ln \hat{y}_i - (n_i - y_i)\ln(n_i - \hat{y}_i) - n_i \cancel{\ln n_i}\Big]$$

برای مقایسه دو مدل M_0 و M_1 با پیکش زیر را محاسبه کنید (به طوری که M_1 پارامتر بیشتری نسبت به M_0 دارد)

$$2\left[l(\hat{\mu}_{M_1};y) - l(\hat{\mu}_{M_0};y)\right]$$

$$= 2\left[l(y;y) - l(\hat{\mu}_{M_0};y)\right] - 2\left[l(y;y) - l(\hat{\mu}_{M_1};y)\right]$$

$$= D(M_0) - D(M_1)$$

نوشت که اختلاف کیش‌ها در مدل M_1 و M_0 را نشان می‌دهد، اگر این اختلاف کوچک باشد به این معناست که دو مدل خوب هستند. تحت مدل M_0 این اختلاف برای نمونه‌های بزرگ دارای توزیع کای دو با درجه آزادی اختلاف بین پارامتر دو مدل M_0 و M_1 است.

اگر اختلاف بزرگ باشد مدلی بهتر است که کیش کوچکتری دارد.

مثال: بررسی اثر دوزهای مختلف (دی‌سولفید کربن بر روی سوسک‌ها)

لوجیت	پروبیت	لگ لگ مکمل (مقادیر برازش شده)	$Y=0$ کشته نشده	تعداد سوسک‌های کشته شده $Y=1$	تعداد سوسک‌ها n_i	لگاریتم دوز X
3.5	3.4 (55.6)	$5.7 \to n_1\hat{\pi}_1$	53	6	59	1.691
9.8	10.7 (49.3)	11.3	47	13	60	1.724
22.4	23.4 (38.6)	20.9	44	18	62	1.755
33.9	33.8 (22.2)	30.3	28	28	56	1.784
50	44.6 (13.4)	47.7	11	52	63	1.811
53.3	53.4 (5.6)	54.2	6	53	59	1.837
59.2	59.7 (2.3)	61.1	1	61	62	1.861
58.8	59.2 (2.8)	59.9	0	60	60	1.884

(در مثال موش کمترین صفحه ۷۱ جزوه)

Subject
Date

سه مدل لگ لگ مکمل، لوجیت، پروبیت را به داده‌ها برازش کنید و مقایسه کنید در مدل لوجیت در سطوح مختلف x نسبت بخت‌ها را بدست آورید

minitab: Stat / Regression / Binary logistic Regression

option → logit / Probit / cloglog

{ number of event y_i
number of trial n_i
Model x_i

لوجیت : $\hat{\beta}_0 = -60.74$ $\hat{\beta}_1 = 34.2859$

پروبیت : $\hat{\beta}_0 = -34.96$ $\hat{\beta}_1 = 19.74$

لگ لگ مکمل : $\hat{\beta}_0 = -39.52$ $\hat{\beta}_1 = 22.01$

$$\text{logit}\,\pi = \log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -60.74 + 34.2859x$$

$$\Rightarrow \hat{\pi}(x) = \frac{\exp(-60.74+34.2859x)}{1+\exp(-60.74+34.2859x)}$$

عبارت محاسبه $\hat{\pi}(x)$ برای هر سطح x → $\hat{\mu}_i = \hat{y}_i = n_i\hat{\pi}_i$

$\hat{y}=0$	$\hat{y}=1$	$y=0$	$y=1$	n_i		
55.5	$59 \times \hat{\pi}(1) = 3.5$	53	6	59	1.691	در سطح
$59-3.5$						

probit : $\Phi^{-1}(\hat{\pi}(x)) = -34.96 + 19.74x \rightarrow \hat{\pi}(x) = \Phi(-34.96 + 19.74x)$

cloglog : $\log(-\log(1-\hat{\pi}(x))) = -39.52 + 22.01x$

$$\Rightarrow \hat{\pi}(x) = 1 - \exp(-\exp(-39.52 + 22.01x))$$

بعد $\hat{y}_i$ ها بدست آمد می‌توانید χ^2 را بدست آورید

x سطح m

$$\chi^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

دسته‌بندی برای دو ستون کشته شده و نشده باید

($\hat{y}_i = ?$ و $\hat{y}_i = ?$)
نشده — کشته شده

جمع ستون بشود

PAPCO

Subject

Date

$\chi^2_{logit} = 9.9 \qquad \chi^2_{Probit} = 9.4 \qquad \chi^2_{cloglog} = ۳,۳۵$

ظاهراً مدل لگ لگ مکمل مقادیری که برازش کرده نزدیک‌تر به مقدار واقعی است اکنون

می‌توانید Devince (کشیدگی) را برای هر سه مدل بدست آورید

$$\underset{\text{Cloglog}}{D(M)} = ۲[۶\log_e ۶ + \dots + 60 \log_e 60] - ۲[۶ \times \log_e 5{,}7 + \dots$$

$$+ 60 \log_e 59.9] + ۲[(59-6)\log_e(59-6) + \dots]$$

$$-2[(59-6)\ln(59-5.7) + \dots]$$

محاسبه χ^2 و Devince:

	y_i کشته شده	~~کشته~~ $n_i - y_i$ کشته نشده	$\hat{y}_i$	$n_i - \hat{y}_i$
			cloglog بدست آورد	
۵۹	۶	۵۳	۵,۷	۵۳,۳
۶۰	۱۳	۴۷	۱۱,۳	۴۸,۷
۶۲	۱۸	۴۴	۲۰,۹	۴۱,۱
۵۶	۲۸	۲۸	۳۰,۳	۲۵,۷
۶۳	۵۲	۱۱	۴۷,۷	۱۵,۳
۵۹	۵۳	۶	۵۴,۲	۴,۸
۶۲	۶۱	۱	۶۱,۱	۰,۹
۶۰	۶۰	۰	۵۹,۹	۰,۱

$$\chi^2_{cloglog} = \frac{(۶ - ۵,۷)^۲}{۵,۷} + \dots + \frac{(۶۰ - ۵۹,۹)^۲}{۵۹,۹} + \frac{(۵۳ - ۵۳,۳)^۲}{۵۳,۳} + \dots$$

$$+ \frac{(۰ - ۰,۱)^۲}{۰,۱} \simeq ۳,۵$$

"χ^2 و کشش نزدیک به هم درمی‌آید"

$$\text{Deviance} = ۲y_i \sum \ln y_i - ۲\sum y_i \ln \hat{y}_i + ۲\sum (n_i - y_i)\ln(n_i - y_i)$$

$$- ۲\sum (n_i - y_i)\ln(n_i - \hat{y}_i) = ۲[۶\ln ۶ + \dots + ۶۰\ln ۶۰] - ۲[۶\ln ۵,۷$$

$$+ \dots + ۶۰\ln ۵۹,۹] + ۲[(۵۹-۶)\ln(۵۹-۶) + \dots + (۶۰-۶۰)\ln(۶۰-۶۰)]$$

$$-۲[(۵۹-۶)\ln(۵۹-۵,۷) + \dots + (۶۰-۶۰)\ln(۶۰-۵۹,۹)] \simeq ۳,۵$$

PAPCO

مدل‌های خطی تعمیم یافته برای مدل داده‌های شمارشی (رگرسیون پواسن)

تاکنون در مورد مطالعات مقطعی با پاسخ‌های دو دویی صحبت کردیم و مدل‌های مختلف جهت برازش این داده‌ها مثل لوجیت و پروبیت، لگ لگ مکمل را بیان کردیم. اما در عمل گاهی اوقات متغیرهای پاسخ گسسته و به صورت شمارشی هستند. مثلاً تعداد اتومبیل‌های سرقت شده در سال ۲۰۰ در تعدادی از شهرهای ایران و .... در این بخش GLM های داده‌های شمارشی بررسی می‌شوند. پس هدف مطالعات مقطعی با پاسخ‌های شمارشی است. در GLM ها با پاسخ‌های شمارشی فرض بر این است که مولفه تصادفی دارای توزیع پواسن است چون شمارش‌ها متغیرهایی هستند که می‌توانند هر مقدار صحیح غیرمنفی را بپذیرند پس توزیع پواسن برای مولفه تصادفی مناسب است.

رگرسیون پواسن:

$$f(y_i;\mu_i)=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

$$\rightarrow e^{-\mu_i}\, e^{-\log y_i!}\, \exp\{y_i \underbrace{\log \mu_i}_{Q(\mu_i)}\}$$

پارامتر طبیعی

ربط کانونی ←

$$\log \mu=\beta_0+\beta_1 x \rightarrow \mu=\exp\{\beta_0+\beta_1 x\}$$

$$=e^{\beta_0}\, e^{\beta_1 x}$$

و بر اساس این مدل هر واحد افزایش در x دارای اثر ضربی e^{β_1} بر μ است. پس میانگین y در $x+1$ برابر میانگین y در x ضرب در e^{β_1} است.

اگر $\beta_1=0$ پس $\mu=e^{\beta_0}$ و میانگین y با تغییرات x تغییر نمی‌کند. اگر $\beta_1>0$ آنگاه $e^{\beta_1}>1$ و میانگین y با افزایش x افزایش می‌یابد. اگر $\beta_1<0$ با افزایش x میانگین کاهش می‌یابد.

نکته ۱: گاهی در عمل از رگرسیون پواسن با رابطه‌های خطی استفاده می‌شود

$$\mu = B_0 + B_1 x$$

نکته ۲: اگر هر پاسخ شمارشی Y دارای شاخص (مانند اندازه جمعیت) t باشد نرخ شمارشی نمونه‌ای آن Y/t است و مقدار مورد انتظار μ/t است و مدل به صورت

$$\ln(\mu/t) = B_0 + B_1 x$$

$$\Rightarrow \ln\mu - \underbrace{\ln t}_{} = B_0 + B_1 x$$

به این مقدار offset می‌گویند

$$\Rightarrow \mu = t\exp\{B_0 + B_1 x\}$$

یعنی میانگین با شاخص t و مقدار ثابتی که متغیر تبیینی وابسته است (x) تناسب دارد به ازای مقدار ثابت x دو برابر کردن t باعث دو برابر شدن تعداد مورد انتظار μ می‌شود

$$\text{نرخ سرقت} = \frac{\text{تعداد سرقت}}{\text{جمعیت}}$$

مثال ص ۱۱۲ اگرستی

cbind(y, n-y)

GLM در splus:

glm(y ~ x, family = binomial (link = logit)
- یا gaussian
- Poisson
- Gamma
- ⋮

(link یا Probit)

, data = data.frame(y, x))

P4PCO

Subject

Date

مدلهای لگ خطی

تا کنون مدلهای منطقی مورد نظر بودند که متغیر پاسخ یکی از متغیرهای مورد مطالعه بود و تاثیر متغیرهای دیگر به عنوان متغیرهای تبیینی بر روی این متغیر مورد بررسی قرار می‌گرفت.

مدلهای لگ خطی حالت خاص از مدلهای GLM هستند که برای تعیین میزان وابستگی شمارش‌ها به سطوح متغیرهای رسته‌ای به کار می‌روند. به عبارت دیگر در مدلهای لگ خطی تاثیر متغیرهای رده بندی شده بر تعداد مورد انتظار (تعداد شمارش‌های) خانه مورد بررسی قرار می‌گیرد که هم اثرات اصلی و هم اثرات متقابل مد نظر می‌باشد. تعداد شمارش‌ها به عنوان یک متغیر پاسخ شمارشی در نظر گرفته می‌شود و تاثیر رده بندی‌ها (شامل تاثیر اصلی هر متغیر و متقابل آنها) بر روی شمارش‌ها مد نظر می‌باشد.

مدل لگ خطی در جدولهای دوطرفه

جدول توافقی $I \times J$ را در نظر بگیرید که n عنصر بر اساس دو متغیر رسته‌ای X و Y به صورت متقاطع دسته بندی شوند. اگر X و Y مستقل باشند می‌دانیم

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad i=1\dots I \qquad j=1\dots J$$

می‌دانیم $\mu_{ij} = n\pi_{ij}$ پس تحت فرض $\mu_{ij} = n\pi_{i+}\pi_{+j}$ (*)

در مدل لگ خطی در واقع از μ_{ij} ها به جای π_{ij} ها استفاده می‌شود. فرض می‌شود شمارش‌های خانه‌ها، n_{ij} ها دارای توزیع پواسن با میانگین‌های μ_{ij} هستند.

با لگاریتم گرفتن از رابطه (*) رابطه ضربی به صورت یک رابطه جمعی می‌شود (همواره برای نوشتن مدل از رابطه جمعی استفاده می‌شود).

$$\Rightarrow \ln \mu_{ij} = \ln n + \ln \pi_{i+} + \ln \pi_{+j}$$

$$\Rightarrow \log(\mu_{ij}) = \underbrace{\lambda}_{\text{میانگین کل}} + \underbrace{\lambda_i^X}_{\text{اثر سطر}} + \underbrace{\lambda_j^Y}_{\text{اثر ستونی}} \qquad (1)$$

Subject

Date

پس $\ln \mu$ به صورت جمعی داشتیم: حجم نمونه، احتمال در سطر i ام و احتمال در ستون j ام

است و برقراری شرط استقلال به صورت رابطه ۱ نیز می‌تواند در نظر گرفته شود.

به مدل ۱ یک مدل لگ خطی استقلال جداول توافقی دوطرفه نامیده می‌شود.

تأثیر رده بندی متغیر X در سطر i ام بر روی لگاریتم شمارش $\rightarrow \lambda_i^X$

(هرچه بزرگتر باشد فراوانی مورد انتظار در این سطر بیشتر می‌شود)

- مدل لگ خطی کلی ۱/ (سلسله مراتبی): برای یافتن تأثیر سطوح مختلف دو متغیر X و Y بر تعداد شمارش‌های خانه‌ها:

$$(2)\qquad \ln \mu_{ij} = \overset{1}{\lambda} + \overset{I}{\lambda_i^X} + \overset{J}{\lambda_j^Y} + \overset{IJ}{\underbrace{\lambda_{ij}^{XY}}_{\text{اثر متقابل سطر i ام X و j ام Y}}} \qquad i = 1 \dots I,\quad j = 1 \dots J$$

← مدل کامل یا مدل سلسله مراتبی

if $\lambda_{ij}^{XY} = 0$ مدل به مدل استقلال تبدیل می‌شود. در مدل ۲ برای اینکه تعداد پارامترها

حداکثر با تعداد خانه‌ها برابر شود باید محدودیت‌هایی بر روی مدل اعمال شود

تعداد پارامترهای مدل ۲ $= \underbrace{1 + I + J + IJ}$; در مدل ۲: تعداد شمارش‌ها: IJ

$\lambda_J^Y = 0$ تأثیر سطح آخر متغیر Y صفر است $\qquad \lambda_{iJ}^{XY} = 0 \quad \forall i$

$\lambda_I^X = 0$ تأثیر سطح آخر متغیر X صفر است $\qquad \lambda_{Ij}^{XY} = 0 \quad \forall j$

تعداد پارامترها: $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$

PAPCO

Scanned with CamScanner

و در این صورت با اعمال محدودیتها، به مدل ۲ که تعداد پارامترها با تعداد خانه‌ها برابر است می‌گویند مدل لگ‌خطی اشباع شده.

بررسی شرط استقلال مدل لگ‌خطی اشباع شده ($I=2$ و $J=2$)

می‌دانیم $\theta = 1$ شرط استقلال است یا $\ln\theta = 0$ پس داریم:

$$\ln\theta = \ln\left(\frac{\overset{n\pi_{11}}{\mu_{11}}\ \overset{n\pi_{22}}{\mu_{22}}}{\underset{n\pi_{12}}{\mu_{12}}\ \underset{n\pi_{21}}{\mu_{21}}}\right)$$

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\frac{n_{11}}{n}\cdot\frac{n_{22}}{n}}{\frac{n_{12}}{n}\cdot\frac{n_{21}}{n}}$$

$$\theta = \frac{\pi_1/1-\pi_1}{\pi_2/1-\pi_2} = \frac{n_{11}\,n_{22}}{n_{12}\,n_{21}}$$

$$= \ln\mu_{11} + \ln\mu_{22} - \ln\mu_{12} - \ln\mu_{21}$$

$$\overset{\text{از مدل ۲}}{=} (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + \dots - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY})$$

$J=2 \quad \lambda_{12}^{XY} = \lambda_{22}^{XY} = 0$

$I=2 \quad \lambda_{21}^{XY} = \lambda_{22}^{XY} = 0$

$$\overset{\text{ساده شد}}{=} \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} \underset{\text{با اعمال محدودیتها}}{=} \lambda_{11}^{XY}$$

$$\Rightarrow \ln\theta = 0 \equiv \lambda_{11}^{XY} = 0$$

در این مدل استقلال معادل است با

تحت پذیرش فرض استقلال داریم:

$$\text{logit}(\pi(x_i)) = \lambda_1^Y = \beta_0 \rightarrow$$ یعنی احتمال موفقیت y به شرط X وابسته نیست

$$\text{logit}(\pi(x_i)) = \log\frac{\pi(x_i)}{1-\pi(x_i)} = \ln\frac{P(Y=1 \mid X=i)}{P(Y=2 \mid X=i)}$$

$$= \ln\frac{P(Y=1, X=i)}{P(Y=2, X=i)} = \ln\frac{\pi_{i1}}{\pi_{i2}} = \ln\frac{\mu_{i1}}{\mu_{i2}}$$

$$= \ln\mu_{i1} - \ln\mu_{i2} \underset{\text{و مدل ۱}}{\overset{\text{تحت استقلال}}{=}} (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y)$$

و محدودیتها $\lambda_2^Y = 0$

$$= \lambda_1^Y$$

نکته: اعمال محدودیتها می‌تواند به صورتهای دیگر نیز در نظر گرفته شود مثلاً

$$\underbrace{\sum_i \lambda_i^{X}}_{I-1} = \underbrace{\sum_j \lambda_j^{Y}}_{J-1} = \underbrace{\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY}}_{(I-1)(J-1)} = 0$$

تعداد پارامتر آزاد $= I-1 + J-1 + (I-1)(J-1)$

مدل لگ خطی برای جدولهای سه طرفه

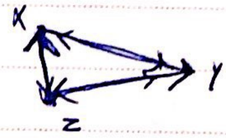
فرض کنید X و Y و Z به ترتیب I و J و K سطح دارند مدل لگ خطی کامل:

$$\begin{array}{ccccccccc} & 1 & I & J & K & IJ & JK & IK & IJK \end{array}$$

$$(۳)\quad \ln \mu_{ijk} = \lambda + \lambda_i^{X} + \lambda_j^{Y} + \lambda_k^{Z} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$$

$i=1,\dots,I \;,\; j=1,\dots,J \;,\; k=1,\dots,K$

و با اعمال محدودیتهای معمول زیر تعداد پارامترهای مدل با تعداد خانه‌ها یعنی IJK برابر می‌شود.



$$\lambda_I^{X} = \lambda_J^{Y} = \lambda_K^{Z} = 0 \qquad \forall j\;\; \lambda_{Ij}^{XY} = \lambda_{jK}^{YZ} = 0$$

$$\forall i\;\; \lambda_{iJ}^{XY} = \lambda_{iK}^{XZ} = 0$$

$$\forall k\;\; \lambda_{Ik}^{XZ} = \lambda_{Jk}^{YZ} = 0$$

$\forall i,j:\; \lambda_{ijK}^{XYZ} = 0$

$\forall i,k:\; \lambda_{iJk}^{XYZ} = 0$

$\forall j,k:\; \lambda_{Ijk}^{XYZ} = 0$

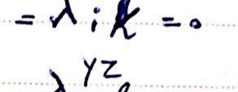
و در این صورت تعداد پارامترها در مدل:

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1) + (I-1)(J-1)(K-1) = IJK$$

و مدل ۳ به یک مدل اشباع شده تبدیل می‌شود برای نمایش مدل ۳ از نماد XYZ استفاده می‌شود

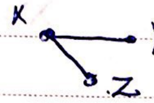
(برای مدل اشباع شده حالتهای مختلف به شرح ذیل بررسی می‌شوند).

۱ـ در مدل ۳ اشباع شده اگر $\forall i,j,k\;\; \lambda_{ijk}^{XYZ} = 0$ آن‌گاه مدل با نماد (XY, XZ, YZ) نمایش داده می‌شود



۲_ اگر $\lambda_{ijk}^{XYZ} = 0$ $\forall i,j,k$ آن گاه
$\forall j,k$ $\lambda_{jk}^{YZ} = 0$

مدل: $\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

با نمادگذاری: (XY, XZ) یا  یا $\underbrace{Z \oplus Y | X}_{\text{Z و Y به شرط X مستقلند}}$ نمایش می‌دهند

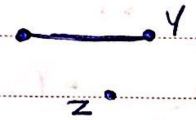
بدیهی است در این حالت تعداد پارامترها:

$\overbrace{IJK}^{\text{مدل اشباع شده}} - (J-1)(K-1) - (I-1)(J-1)(K-1)$

به عبارت دیگر در این مدل: $\pi_{jk|i} = \pi_{j+|i}\,\pi_{+k|i}$

۳_ مدل دیگری تواند به صورت $\lambda_{ijk}^{XYZ} = 0$ ، $\lambda_{jk}^{YZ} = 0$ ، $\lambda_{ik}^{XZ} = 0$ در این صورت

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$$

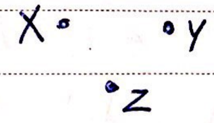
و نمادگذاری: (XY, Z) و  و $Z \oplus Y, X$

نمایش می‌دهند

$$\pi_{ijk} = \pi_{ij+}\,\pi_{++k}$$

۴_ مدل استقلال: علاوه بر شرایط ۳ ، $\lambda_{ij}^{XY} = 0$ آن گاه

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

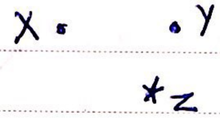
نمادگذاری: (X, Y, Z) ،  ، $X \oplus Y \oplus Z$

$$\pi_{ijk} = \pi_{i++}\,\pi_{+j+}\,\pi_{++k}$$

Subject
Date

۵ـ با علاوه بر موارد ۴ ، $\lambda_k^Z = 0$ آن گاه

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y \qquad i=1\ldots I,\ j=1\ldots J$$

نمودار :  یا (X, Y) یا $(X \oplus Y \oplus Z, Z=u)$

می دانیم $\sum_i \sum_j \sum_k \pi_{ijk} = 1$ و داریم $\pi_{ijk} = c\, \pi_{i++}\, \pi_{+j+}$ بنابراین با جایگذاری

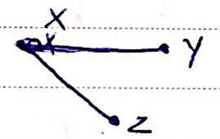
(c اثر ثابت)

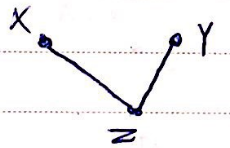
$$\sum_i \sum_j \sum_k c\, \pi_{i++}\, \pi_{+j+} = 1 \Rightarrow \sum_k c = 1$$

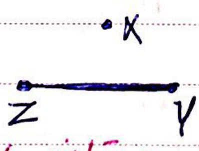
$$\Rightarrow c = 1/K$$

$$\Rightarrow \pi_{ijk} = \frac{\pi_{i++}\, \pi_{+j+}}{K}$$

و همچنین مدل‌های دیگر :

$(XZ, XY) \rightarrow$  $Z \oplus Y \mid X$

$(XZ, YZ) \rightarrow$  $X \oplus Y \mid Z$

$(YZ, X) \rightarrow$  $X \oplus Y, Z$

برای مقایسه مدل‌های مختلف می توانید از X^2 یا G^2 استفاده کنید.

برآورد پارامتر : برای هر کدام از مدل‌ها تحت شرایط قید شده می توان را با روش درستنمایی بدست آورد. در نظر فرض کنید مدل (XZ, YZ) مدل استقلال شرطی مدل :

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

PAPCO

یعنی $\pi_{ij|k} = \pi_{i+|k} \cdot \pi_{+j|k}$

Scanned with CamScanner

Subject
Date

پس

$$L = \prod_i \prod_j \prod_k \frac{e^{-\mu_{ijk}} (\mu_{ijk})^{n_{ijk}}}{n_{ijk}!}$$

$$\xrightarrow{\ln L} l = \sum_i \sum_j \sum_k n_{ijk} \ln \mu_{ijk} - \sum\sum\sum \mu_{ijk}$$

به جای $\ln \mu_{ijk}$ رابطه را جایگذاری می کنید و نسبت به پارامتر مشتق می گیرید

$$\mu_{ijk} = e^{\text{مدل}} = e^{\lambda + \cdots}$$

$$\frac{\partial l}{\partial \lambda_{ik}^{xz}} = n_{i+k} - \mu_{i+k} = 0 \rightarrow \hat{\mu}_{i+k} = n_{i+k}$$

$$\frac{\partial l}{\partial \lambda_{jk}^{yz}} = n_{+jk} - \mu_{+jk} = 0 \Rightarrow \hat{\mu}_{+jk} = n_{+jk}$$

— برآوردگر ML = آماره بسنده

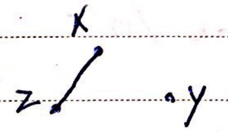
از طرفی داریم:

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{++k}} \Rightarrow \pi_{i+|k}\, \pi_{+j|k} = \pi_{ijk}/\pi_{++k}$$

$$\Rightarrow \frac{\pi_{i+k}}{\pi_{++k}} \cdot \frac{\pi_{+jk}}{\pi_{++k}} = \frac{\pi_{ijk}}{\pi_{++k}}$$

$$\Rightarrow \pi_{ijk} = \frac{\pi_{i+k}\, \pi_{+jk}}{\pi_{++k}} \Rightarrow n\pi_{ijk} = \frac{n\pi_{i+k}\, n\pi_{+jk}}{n\pi_{++k}}$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+k}\, \hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{n_{i+k}\, n_{+jk}}{n_{++k}}$$



(XZ, Y)

تمرین: نشان دهید در مدل $Y \oplus X, Z$

$$\hat{\mu}_{ijk} = \frac{n_{+j+}\, n_{i+k}}{n}$$

می توان نشان داد $\pi_{ijk} = \pi_{+j+}\, \pi_{i+k} \Rightarrow n\pi_{ijk} = n\pi_{+j+}\, n\pi_{i+k}/n$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{+j+}\, \hat{\mu}_{i+k}}{n}$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{n_{+j+}\, n_{i+k}}{n}$$

PAPCO

Subject
Date

- برآورد μ_{ijk} را در مدل $\begin{matrix} X\circ & \\ & *Y \\ Z* & \end{matrix}$ بدست آورید

$$\ln \mu_{ijk} = \lambda + \lambda_i^X$$

یعنی می‌دانیم $\sum\sum\sum \pi_{ijk} = 1$ و داریم π_{i++} پس $\pi_{ijk} = \underset{\text{اثری در Z نداریم}}{\underset{\downarrow}{C}}\ \pi_{i++}$

$$C \sum_i \sum_j \sum_k \pi_{i++} = 1$$

$$\Rightarrow C = \frac{1}{JK}$$

$$\Rightarrow \pi_{ijk} = \frac{\pi_{i++}}{JK}$$

$$\rightarrow n\,\pi_{ijk} = \frac{n\,\pi_{i++}}{JK}$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{i++}}{JK} = \frac{n_{i++}}{JK}$$

- برآورد μ_{ijk} در مدل $\begin{matrix} X\circ & & \circ Y \\ & Z\circ & \end{matrix}$ (X, Y, Z)

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

$$\pi_{ijk} = \pi_{i++}\ \pi_{+j+}\ \pi_{++k}$$

$$n\,\pi_{ijk} = n\,\pi_{i++} \cdot n\,\pi_{+j+} \cdot n\,\pi_{++k} / n^2$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{i++}\ \hat{\mu}_{+j+}\ \hat{\mu}_{++k}}{n^2}$$

پس از تشکیل دستگاه ی تعیین آماره بسنده

$$\hat{\mu}_{i++} = n_{i++} \qquad \hat{\mu}_{+j+} = n_{+j+} \qquad \hat{\mu}_{++k} = n_{++k}$$

PAPCO

Subject
Date

۱ برچ (1963) نشان داد آماره‌ی بسنده‌ی ML است

مدل (XZ, YZ)
$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

$$L = \prod\prod\prod e^{-\mu_{ijk}} (\mu_{ijk})^{n_{ijk}} / n_{ijk}!$$

$$l \propto -\sum\sum\sum \mu_{ijk} + \sum\sum\sum n_{ijk} \ln \mu_{ijk}$$

$$= n\lambda + \sum_i n_{i++} \lambda_i^X + \sum_j n_{+j+} \lambda_j^Y + \sum_k n_{++k} \lambda_k^Z + \sum\sum \underbrace{n_{i+k}} \lambda_{ik}^{XZ} + \sum\sum n_{+jk} \lambda_{jk}^{YZ} - \sum\sum\sum \exp\{\text{مدل}\}$$

$$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda$$

اندیس‌ها جدا شوند و جمع بسته شود

$$\cancel{\frac{\partial l}{\partial \lambda_{ik}^{XZ}} = n_{i+k} - \mu_{i+k} = 0}$$

$\ln \pi_{i+k} =$

~~[illegible]~~ و طبق برچ ضرایب پارامترها که آماره‌ی بسنده هستند برآوردگر ML نیز می‌باشند.

Subject
Date

① جدول زیر را در نظر بگیرید برابر مدل

		Y	
		۱	۲
X	۱	۱۲۶	۱۰۰
	۲	۳۵	۶۱
	۳	۹۰۸	۶۸۸

$\ln \mu_{ij} = \lambda + \lambda_i^X$

برآورد μ_{ij} را بدست آورده و نیکویی برازش؟

$\underbrace{IJ}K = ۶$

$\ln \mu_{ij} = \lambda + \lambda_i^X$ $\qquad$ $۶ - ۱ - ۲ = ۳$

$$\left(\sum\sum \pi_{ij} = ۱ \ , \ \pi_{ij} = C\pi_{i+}\right) \Rightarrow \sum_i \sum_j C\pi_{i+} = ۱ \Rightarrow C = \frac{1}{J}$$

$$n\pi_{ijk} = \frac{n}{J}\pi_{i+} \Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+}}{J} = \frac{n_{i+}}{J}$$

$$\hat{\mu}_{۱۱} = \frac{n_{۱+}}{۲} = \frac{۱۲۶+۱۰۰}{۲} = \frac{۲۲۶}{۲} = ۱۱۳$$

$$\hat{\mu}_{۱۲} = \frac{n_{۱+}}{۲} = \frac{۱۲۶+۱۰۰}{۲} = \frac{۲۲۶}{۲} = ۱۱۳$$

$$\hat{\mu}_{۲۱} = \frac{n_{۲+}}{۲} = \frac{۳۵+۶۱}{۲} = ۴۸ \qquad \hat{\mu}_{۲۲} = ۴۸$$

$$\hat{\mu}_{۳۱} = \frac{n_{۳+}}{۳} = \frac{۹۰۸+۶۸۸}{۲} = ۷۹۸ \qquad \hat{\mu}_{۳۲} = ۷۹۸$$

$$\chi^2 = \frac{(۱۲۶-۱۱۳)^۲}{۱۱۳} + \frac{(۱۰۰-۱۱۳)^۲}{۱۱۳} + \frac{(۳۵-۴۸)^۲}{۴۸} +$$

$$\frac{(۶۱-۴۸)^۲}{۴۸} + \frac{(۹۰۸-۷۹۸)^۲}{۷۹۸} + \frac{(۶۸۸-۷۹۸)^۲}{۷۹۸} = ۴۰٫۳۵۸$$

$$p\text{-value} = P(\chi^۲_۳ > ۴۰٫۱۲۵۸) = ۰/۰۰۰$$

این مدل رد می‌شود $\overset{۶-۶=۰}{\text{منجر به}}$ مدل اشباع شد

PAPCO

Subject
Date

	Y			
	۲	۱	X	Z
	۱۰۰	۱۲۶	۱	۱
	۶۱	۳۵	۲	
	۹۰۸	۶۸۸	۱	۲
	۴۹۷	۸۰۷	۲	

۲) جدول زیر را در نظر بگیرید.

الف) برای مدل

$$\ln \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

تعداد پارامترها، نمودار و برآورد μ_{ijk} را بدست آورید.

ب) با استفاده از داده‌های جدول مدل را ارزش کنید.

X • • Y

• Z

$$\widehat{(IJK)} - \overbrace{(I-1)(J-1)}^{1} - \overbrace{(K-1)(J-1)}^{1}$$
$$- \underbrace{(I-1)(K-1)}_{1} - \underbrace{(I-1)(J-1)(K-1)}_{1}$$

(X^2, G^2)

استقلال $\pi_{ijk} = \pi_{i++}\, \pi_{+j+}\, \pi_{++k}$

$$n\pi_{ijk} = n\pi_{i++}\, n\pi_{+j+}\, n\pi_{++k} / n^2$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{i++}\, \hat{\mu}_{+j+}\, \hat{\mu}_{++k}}{n^2} = \frac{n_{i++}\, n_{+j+}\, n_{++k}}{n^2}$$

$$n_{1++} = 126 + 100 + 688 + 908 = 1822$$
$$n_{+1+} = 126 + 35 + 688 + 807 = 1656$$
$$n_{++1} = 126 + 35 + 100 + 61 = 322$$

$$\hat{\mu}_{111} = \frac{(1822)(1656)(322)}{(3222)^2} = 93.58$$

$\hat{\mu}_{121} = 88.80$ $\hat{\mu}_{211} = 71.91$ $\hat{\mu}_{221} = 68.00$

$\hat{\mu}_{112} = 842.86$ $\hat{\mu}_{122} = 797.05$ $\hat{\mu}_{212} = 647.64$

$\hat{\mu}_{222} = 612.44$

Subject
Date

مدل اشباع شده $IJK = ۸$ — (XZ, Y) $۵ = ۳$

$$\chi^{۲} = \frac{(۱۲۴-۹۳,۵۸)^{۲}}{(۹۳,۵۸)} + \dots + \frac{(۴۹۷-۴۱۲,۴۴)^{۲}}{۴۱۲,۴۴} = \boxed{۱۳۷,۱۵۶}$$

حل ۵: X —— Z, ·Y

مدل: X —— Y, Z·

$\hat{I}JK - (I-1)(J-1) - (J-1)(K-1) - (I-1)(J-1)(K-1) = ۵$ تعداد پارامترها

(XZ, Y) $\quad Y \oplus X,Z$

$$n\pi_{ijk} = \frac{n\pi_{+j+}\, n\pi_{i+k}}{n} \rightarrow \hat{\mu}_{ijk} = \frac{\hat{\mu}_{+j+}\, \hat{\mu}_{i+k}}{n}$$

$$\Rightarrow \hat{\mu}_{ijk} = \frac{n_{+j+}\, n_{i+k}}{n}$$

$$\hat{\mu}_{111} = \frac{n_{+1+}\, n_{1+1}}{n} = \frac{\overbrace{(۱۲۴+۳۵+۶۸۸+\cancel{x}۰۷)}^{۱۹۵۶}\overbrace{(۱۲۴+۱۰۰)}^{۲۲۴}}{۳۲۲۲}$$

$$= ۱۱۹,۱۵۹۴$$

$\hat{\mu}_{121} = ۱۰۹,۸۴۴$ $\quad \hat{\mu}_{211} = ۴۹,۳۴$ $\quad \hat{\mu}_{221} = ۴۹,۴۴$

$\hat{\mu}_{112} = ۸۲۰,۲۹$ $\quad \hat{\mu}_{122} = ۷۷۵,۷۱$ $\quad \hat{\mu}_{212} = ۴۷۰,۲۱$

$\mu_{222} = ۴۶۶,۷۹$

$$\chi^2 = \frac{(۱۲۴-۱۰۹,۸۴۴)^{۲}}{۱۰۹,۸۴۴} + \dots + \frac{(۴۹۷-۴۶۶,۷۹)^{۲}}{۴۶۶,۷۹}$$

$$= \boxed{۱۱۱,۹۳}$$

$$\text{p-value} = P(\chi^2_{۳} > ۱۱۱,۹۳) = ۰$$

مدل اشباع شده بهتر است

مدل ـــ اشباع شده $= ۸ - ۵ = ۳$

PAPCO