به نام خداوند بخشنده مهربان

روش‌های آماری و اقتصادسنجی

در تحلیل و مدل‌سازی داده‌های حمل و نقلی

محمد مهدی بشارتی

besharati@iut.ac.ir

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

- مقدمه
- پروبیت دوگانه
- لوجیت دوگانه (لوجستیک)
- معیارهای نیکویی برازش
- ریسک و شانس
- لوجیت چندگانه
- لوجیت شرطی

❑ انواع خاصی از متغیرهای گسسته؛

- مقادیر ۰، ۱، ۲، ... داشته باشد.
- مقادیر ۰ و ۱ (دو انتخابی)
- مقادیر محدود (مثلا ۴ گروه: ۰، ۱، ۲، ۳)

- **متغیر وابسته ممکن است در اختیار کردن مقادیر خود با محدودیت روبرو باشد.**
- **مثال؛**
  - **عوامل موثر بر اشتغال افراد (متغیر وابسته؟)،**
  - **عوامل موثر بر فوتی بودن یک تصادف (متغیر وابسته؟)،**
  - **عوامل موثر بر انتخاب حمل‌ونقل شخصی (متغیر وابسته؟)،**
  - **و ...**

- به طور کلی، متغیر وابسته می‌تواند یک متغیر کیفی باشد که نتیجه تصمیم‌گیری‌های افراد (یا یک رویداد مشخص) را نشان دهد.
- یک انتخاب یا وقوع یک رویداد را به مجموعه‌ای از متغیرهای توصیفی (توضیحی) مرتبط می‌کنیم (رگرسیون).
- پایه اصلی این موارد در چارچوب کلی مدل‌های احتمال مورد بررسی قرار می‌گیرد.

- پایه اصلی این موارد در چارچوب کلی مدل‌های احتمال مورد بررسی قرار می‌گیرد.
- مثلاً وقوع تصادف فوتی را به عنوان متغیر Y در نظر می‌گیریم: (وقوع تصادف فوتی: Y=1)
- احتمال وقوع حادثه موردنظر (احتمال فوتی بودن تصادف):

$$P(\text{فوتی بودن}) = P(Y=1)$$

- احتمال فوتی بودن تصادف می‌تواند تحت تأثیر عوامل مختلفی باشد (سن، جنسیت، سرعت برخورد، نوع خودرو و ...)
- پس می‌توان احتمال بالا را تابعی از مجموعه‌ای از متغیرهای توصیفی دانست.
- در بسیاری از موضوعات، متغیر وابسته (Y) بیانگر دو حالت است (مدل‌های دو انتخابی)؛
  - Y=0 عدم وقوع/انتخاب موضوع موردنظر
  - Y=1 وقوع/انتخاب موضوع موردنظر

❑ مدل‌های دو انتخابی

- **برای مدل‌های دو انتخابی می‌توان مدل احتمال را به صورت زیر تعریف کرد؛**

$$P(Y=1|x_i)=F(x_i,\beta)$$

$$P(Y=0|x_i)=1-F(x_i,\beta)$$

بردار متغیرهای موثر بر احتمال وقوع Y : $x_i'=[1\ \ x_{2i}\ \ \ldots x_{Ki}]$

سوال: چه رابطه‌ای بین متغیرهای توصیفی و متغیر وابسته وجود دارد؟

پاسخ: بستگی به شکل تابع $F(x_i,\beta)$ دارد (ممکن است خطی یا غیرخطی باشد).

در چارچوب رگرسیون، مدل Y_i را به صورت روبرو می‌نویسیم: $Y_i=E(Y|x_i)+u_i=F(x_i,\beta)+u_i$

در واقع، تابع $F(x_i,\beta)$ همان امیدریاضی شرطی Y است.

اگر برای $F(x_i,\beta)$ یک معادله خطی تعریف کنیم، آنگاه معادله روبرو یک مدل رگرسیون خطی چندمتغیره خواهد بود.

به طور کلی برای $F(x_i,\beta)$ توابع مختلفی معرفی می‌شود که هرکدام با یک نام مخصوص شناخته می‌شود.

❑ به طور کلی برای $F(x_i,\beta)$ توابع مختلفی معرفی می‌شود که هرکدام با یک نام مخصوص شناخته می‌شود؛

- **مدل احتمال خطی**
- **مدل پروبیت**
- **مدل لوجیت**

❑ مدل احتمال خطی ((Linear Probability Model (LPM)؛

- **در مدل احتمال خطی برای** $F(x_i,\beta)$ یک معادله خطی تعریف می‌شود:

$$F(x_i,\beta) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

- **امید ریاضی شرطی Y:**

$$E(Y|x_i) = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

- **مدل رگرسیون Y:**

$$Y_i = E(Y|x_i) + u_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

- **این یک مدل احتمال خطی (LPM) است. مدلی ساده شبیه مدل رگرسیون معمولی و دارای مشکلات زیر؛**

۱- ناهمسانی واریانس دارد (واریانس u_i وابسته به مقدار متغیرهای توصیفی است).

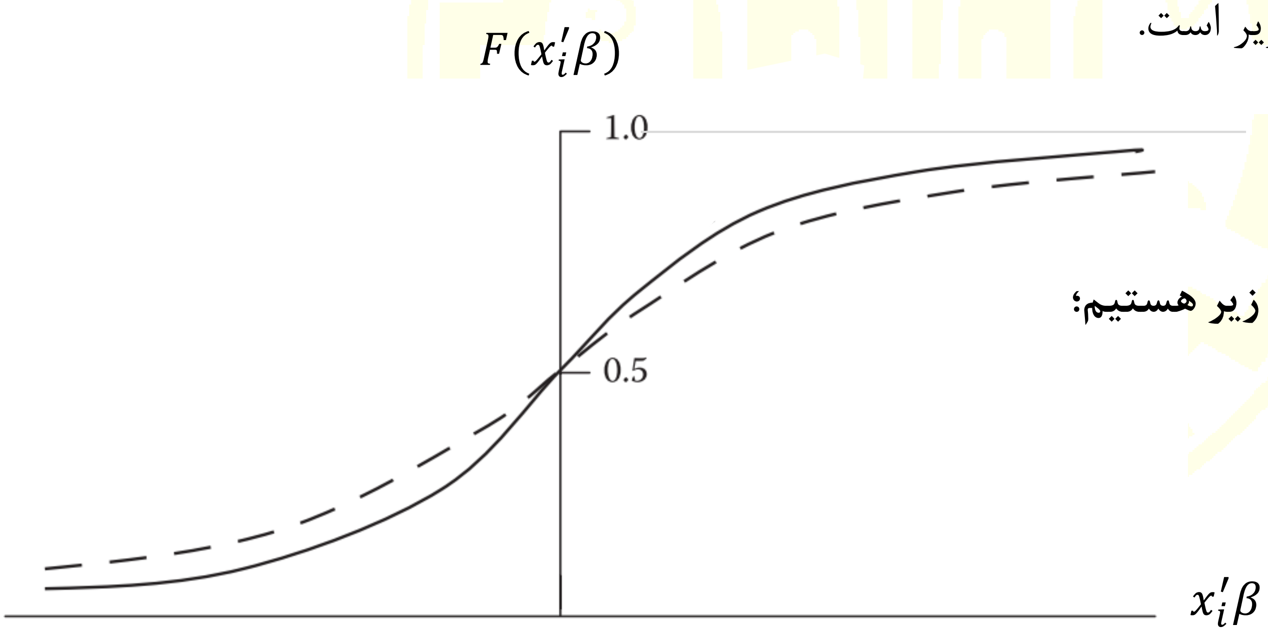
۲- چون $F(x_i,\beta)$ بیانگر احتمال است، مقادیر آن همواره باید بین صفر و یک قرار بگیرد. در مسائل واقعی ممکن است $x_i'\beta$ خارج از محدوده $[0,1]$ قرار بگیرد.

❑ نظریه مطلوبیت تصادفی

- یکی از مشکلات مدل احتمال خطی (LPM) آن است که مقدار $F(x_i,\beta)$ که بیانگر یک احتمال است، ممکن است گاهی اوقات مقادیر بیشتر از ۱ و یا کمتر از صفر را اختیار کند.

- برای حل این مشکل، از مدل‌های دیگری استفاده می‌کنیم که **قیدی را روی تابع احتمال** $F(x_i,\beta)$ اِعمال می‌کنند تا مقدار احتمال محدوده به بازه $[0,1]$ گردد.

- نمودار چنین تابعی به صورت زیر است.



- **پس برای حل این مشکل، نیازمند موارد زیر هستیم؛**
  - یک معادله رگرسیون،
  - یک تابع احتمال

❑ نظریه مطلوبیت تصادفی

○ **برای حل این مشکل، نیازمند موارد زیر هستیم؛**

- یک معادله رگرسیون،
- یک تابع احتمال

بر مبنای **نظریه مطلوبیت تصادفی** می‌توان این دو موضوع را مرتبط کرده و مشکل مذکور را حل کرد.

مطلوبیت تصادفی از نظریه انتخاب استفاده می‌کند. نظریه انتخاب در حالت دو انتخابی:

فرض کنید هر فرد ۲ انتخاب دارد. او گزینه‌های خود را به گونه‌ای انتخاب می‌کند که مطلوبیتش (U_i) حداکثر شود.

برای فرد i-ام

U_{1i} : مطلوبیت حاصل از انتخاب گزینه ۱

U_{0i} : مطلوبیت حاصل از انتخاب گزینه ۰

اگر $U_{1i} \geq U_{0i}$ و یا $U_{1i} - U_{0i} \geq 0$ باشد، آنگاه گزینه ۱ انتخاب می‌شود.

در ادامه، متغیر وابسته Y_i^* را به صورت روبرو تعریف می‌کنیم:

$$Y_i^* = U_{1i} - U_{0i}$$

اگر $Y_i^* \geq 0$ باشد، فرد i-ام گزینه <u>۱</u> و اگر $Y_i^* < 0$ باشد، فرد i-ام گزینه <u>0</u> را انتخاب می‌کند.

❑ نظریه مطلوبیت تصادفی

می‌توان ادعا کرد که مطلوبیت، ترجیحات فرد را منعکس می‌کند، بنابراین، مقداری که Y_i^* اختیار می‌کند، می‌تواند تحت تأثیر مجموعه‌ای از عوامل و متغیرهای دیگر باشد.

یعنی اینکه کدام گزینه برای فرد i مطلوب‌تر است، به عوامل مختلفی بستگی دارد:

$$Y_i^* = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i \qquad ; X_{1i} = 1$$

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i \quad ; \ \mathbf{x}_i' = [1 \quad X_{2i} \quad \cdots \quad X_{ki}] \quad ; \quad \boldsymbol{\beta}' = [\beta_1 \quad \beta_2 \quad \cdots \quad \beta_k]$$

برای کاربردهای عملی باید مدل بالا را برآورد کنیم.

چالش: Y_i^* یک متغیر غیرقابل مشاهده است.

برای حل این چالش، متغیر قابل مشاهده Y_i را معرفی می‌کنیم (که عبارتست از انتخاب مشاهده شده برای فرد i-ام).

نظریه مطلوبیت تصادفی

مثلا Y_i انتخاب مترو برای جابجایی است. اگر $Y_i = 1$ باشد، فرد i-ام مترو را برای جابجایی انتخاب کرده است (و بالعکس).

رابطه میان Y_i^* و Y_i به صورت زیر است؛

۱- اگر $Y_i = 1$ باشد، بدین معناست که $Y_i^* \geq 0$ است. (گزینه ۱ برای فرد i-ام مطلوب‌تر بوده است).

۲- اگر $Y_i = 0$ باشد، بدین معناست که $Y_i^* < 0$ است. (گزینه 0 برای فرد i-ام مطلوب‌تر بوده است).

- **بنابراین، یک مدل آماری داریم که دو جزء دارد؛**

۱- یک معادله رگرسیون برای متغیر غیرقابل مشاهده Y_i^*

۲- یک معادله برای مرتبط ساختن متغیر غیرقابل مشاهده Y_i^* و متغیر قابل مشاهده Y_i (Link function)

✓ **برای مرتبط کردن Y_i^* و Y_i از مفهوم احتمال استفاده می‌کنیم.**

❑ نظریه مطلوبیت تصادفی

برای مرتبط کردن Y_i^* و Y_i از مفهوم احتمال استفاده می‌کنیم.

$$P(\mathrm{Y}_i = 1) = P(Y_i^* \geq 0) = P\left(\mathbf{x}_i'\boldsymbol{\beta} + u_i \geq 0\right) = P(u_i \geq -\mathbf{x}_i'\boldsymbol{\beta})$$

پس نهایتاً احتمال انتخاب مترو برای جابجایی

معادل است با اینکه مترو مطلوبیت بیشتری داشته باشد $(Y_i^* \geq 0)$

معادل است با احتمال اینکه متغیر تصادفی u_i بزرگتر از $-\mathbf{x}_i'\boldsymbol{\beta}$ باشد.

لازم است برای u_i یک تابع احتمال تعریف کنیم و خصوصیات آن را بررسی کنیم.

دو تابع رایج مورد استفاده برای تابع احتمال u_i؛

۱- تابع توزیع نرمال (مدل **پروبیت**)

۲- تابع توزیع لوجستیک (مدل **لوجیت**)

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

مقدمه

پروبیت دوگانه

لوجیت دوگانه (لوجستیک)

معیارهای نیکویی برازش

ریسک و شانس

لوجیت چندگانه

لوجیت شرطی

1. ساختن تابع توزیع خطای مدل

فرض کنید u_i دارای توزیع نرمال باشد. برای هر متغیری مانند Z که تابع چگالی احتمال آن نرمال استاندارد باشد، می‌توان تابع توزیع یا تابع احتمال تجمعی را به صورت زیر معرفی کرد؛

$$P(Z \leq z) = \int_{-\infty}^{z} \varphi(z)dz = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz = \Phi(z)$$

$\varphi(z)$ تابع چگالی و $\Phi(z)$ تابع توزیع Z است.

اگر u_i دارای توزیع نرمال باشد؛

$$P(Y_i = 1) = P(Y_i^* \geq 0) = P\left(u_i \geq -\mathbf{x}_i'\boldsymbol{\beta}\right) = 1 - P\left(u_i < -\mathbf{x}_i'\boldsymbol{\beta}\right) = \mathbf{1} - \Phi(-\mathbf{x}_i'\boldsymbol{\beta})$$
$$= \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

از سوی دیگر،

$$P(Y_i = 0) = P(Y_i^* < 0) = 1 - P(Y_i = 1) = 1 - \Phi(\mathbf{x}_i'\boldsymbol{\beta}) = \Phi(-\mathbf{x}_i'\boldsymbol{\beta})$$

2. برآورد پارامترهای مدل

برای برآورد پارامترهای مدل از روش MLE استفاده می‌کنیم.

می‌دانیم متغیر تصادفی Y_i دارای توزیع برنولی است.

Y_i	0	1
p_i	$1-\Phi_i$	Φ_i

که در آن، $\Phi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$

این توزیع را به صورت زیر نیز می‌توان نشان داد؛

$$p_i = P(Y_i|\mathbf{x_i}) = \Phi_i^{Y_i}(1-\Phi_i)^{1-Y_i} \qquad , Y_i = 0\,,1$$

تابع درستنمایی؛

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i|\mathbf{x}_i) = \prod_{i=1}^{n} \Phi_i^{Y_i}(1-\Phi_i)^{1-Y_i}$$

2. برآورد پارامترهای مدل

تابع درستنمایی؛

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i|\mathbf{x}_{\boldsymbol{i}}) = \prod_{i=1}^{n} \Phi_i^{Y_i}(1-\Phi_\mathrm{i})^{1-Y_i}$$

لگاریتم تابع درستنمایی را محاسبه کرده و نسبت به β مشتق می‌گیریم.

بدین ترتیب، برآوردگر $\hat{\beta}_{ML}$ به دست می‌آید.

3. تفسیر نتایج مدل پروبیت

در هر مدل رگرسیون، β بیانگر اثرات نهایی متغیرهای توصیفی بر متغیر وابسته است.

در اینجا، متغیر وابسته، بیانگر مطلوبیت است (که غیرقابل مشاهده است).

احتمال برآورد شده در مدل پروبیت (بر مبنای $\hat{\beta}$) به صورت زیر است؛

$$p_i = P(Y_i = 1|x_i) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}) = \Phi(\hat{\beta}_۱ + \hat{\beta}_۲ X_{۲i} + \ldots + \hat{\beta}_k X_{ki})$$

مثال؛ فرض کنید یک مدل تک متغیره داریم (و پارامترها به صورت ۰.۸ و ۰.۰۱ برآورد شده است)؛

Y_i : انتخاب حمل‌ونقل شخصی برای جابجایی ($Y_i = 1$ به معنی انتخاب خودرو شخصی و بالعکس)

X_i : سطح درآمد فرد i-ام

$$p_i = P(Y_i = 1|x_i) = \Phi(-0.8 + 0.01\, X_i)$$

3. تفسیر نتایج مدل پروبیت

مثال؛ فرض کنید یک مدل تک متغیره داریم (و پارامترها به صورت ۰.۸- و ۰.۰۱ برآورد شده است)؛

Y_i : انتخاب حمل‌ونقل شخصی برای جابجایی ($Y_i = 1$ به معنی انتخاب خودرو شخصی و بالعکس)

X_i : سطح درآمد فرد i-ام

$$p_i = P(Y_i = 1|x_i) = \Phi(-0.8 + 0.01\,X_i)$$

احتمال انتخاب خودرو شخصی به ازای سطوح مختلف درآمدی؛

سطح درآمدی X_i=40

$$P(Y_i = 1|x_i = 40) = \Phi\big(-0.8 + 0.01\,(40)\big) = 0.345$$

سطح درآمدی X_i=80

$$P(Y_i = 1|x_i = 80) = \Phi\big(-0.8 + 0.01\,(80)\big) = 0.5$$

سطح درآمدی X_i=150

$$P(Y_i = 1|x_i = 150) = \Phi\big(-0.8 + 0.01\,(150)\big) = 0.758$$

سطح درآمدی X_i=200

$$P(Y_i = 1|x_i = 200) = \Phi\big(-0.8 + 0.01\,(200)\big) = 0.885$$

پس با افزایش سطوح درآمدی، احتمال انتخاب خودرو شخصی افزایش می‌یابد.

3. تفسیر نتایج مدل پروبیت

برای بررسی اثرات نهایی، اثر تغییر در X_i ها را بر $\mathbf{Y}$ اندازه‌گیری می‌کنیم.

در این مدل‌ها که Y_i^* یک متغیر کیفی و غیرقابل مشاهده است، $\boldsymbol{\beta}$ اثر تغییرات X_i ها را بر Y_i اندازه‌گیری می‌کند.

اگر $\boldsymbol{\beta}$ مثبت باشد، مطلوبیت انتخاب گزینه ۱ همراه با افزایش X_i، افزایش می‌یابد.

اما سوال اینست: در واکنش به تغییر X_i، احتمال انتخاب گزینه ۱ چقدر افزایش می‌یابد؟

اثر تغییر X_i بر احتمال انتخاب گزینه ۱ به صورت روبرو محاسبه می‌شود؛

$$\frac{d\,P(Y_i=1)}{dX_i}=\varphi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$$

مثلاً اثر تغییر در X_{ki} بر $P(Y_i=1)$ به صورت روبرو محاسبه می‌شود؛

$$\frac{d\,P(Y_i=1)}{dX_{ki}}=\varphi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}=\varphi(\hat{\beta}_۱+\hat{\beta}_۲X_{۲i}+\ldots+\hat{\beta}_kX_{ki})\hat{\beta}_k$$

در مثال قبلی، اثر تغییر در درآمد بر احتمال انتخاب خودرو شخصی برای جابجایی:

$$\frac{d\,P(Y_i=1)}{dX_i}=\varphi(-0.8+0.01\,X_i)\times 0.01$$

3. تفسیر نتایج مدل پروبیت

در مثال قبلی، اثر تغییر در درآمد بر احتمال انتخاب خودرو شخصی برای جابجایی:

$$\frac{d\,P(\mathrm{Y}_i=1)}{d\mathrm{X}_i}=\varphi(-0.8+0.01\,\mathrm{X_i})\times 0.01$$

این بدان معناست که به ازای مقادیر مختلف X، اثر نهایی X بر Y نیز تغییر می‌کند.

می‌توانیم اثر نهایی X بر Y را براساس مقدار متوسط X برآورد و گزارش کنیم.

$$\frac{d\,P(\mathrm{Y}_i=1)}{d\mathrm{X}_{ki}}=\varphi(\hat{\beta}_1+\hat{\beta}_2\bar{\mathrm{X}}_2+\ldots+\hat{\beta}_k\bar{\mathrm{X}}_k)\hat{\beta}_k$$

مثال قبلی، اگر مقدار **متوسط درآمد برابر با ۱۵۰ باشد،**

اثر تغییر در درآمد بر احتمال انتخاب خودرو شخصی برای جابجایی **به طور متوسط:**

$$\frac{d\,P(\mathrm{Y}_i=1)}{d\mathrm{X}_i}=\varphi\left(-0.8+0.01\bar{\mathrm{X}}_{\mathrm{i}}\right)\times 0.01=\varphi(0.7)\times 0.01=0.3123\times 0.01=0.031$$

پس، اگر درآمد افراد به طور متوسط افزایش یابد، احتمال انتخاب خودرو شخصی ۳.۱٪ افزایش می‌یابد.

3. تفسیر نتایج مدل پروبیت

اگر درآمد فردی کمتر از متوسط باشد (مثلا ۴۰ باشد)، احتمال انتخاب خودرو شخصی توسط او با افزایش درآمد، چقدر است؟

$$\frac{d\,P(\mathrm{Y}_i = 1)}{d\mathrm{X}_i} = \varphi(-0.8 + 0.01 \times 40) \times 0.01 = 0.03683$$

حدود ۳.۷٪ افزایش می‌یابد.

پس با افزایش درآمد برای این فرد احتمال انتخاب خودرو شخصی بیشتر از متوسط افزایش می‌یابد (حدود ۳.۷٪).

سوال: بیشترین تاثیر درآمد بر احتمال انتخاب خودرو شخصی به ازای چه شرایطی رخ می‌دهد؟

فصل یازدهم

مدل رگرسیون داده‌های گسسته

Discrete data Regression Model

- مقدمه
- پروبیت دوگانه
- **لوجیت دوگانه (لوجستیک)**
- معیارهای نیکویی برازش
- ریسک و شانس
- لوجیت چندگانه
- لوجیت شرطی

1. ساختن تابع توزیع خطای مدل

به جای تابع چگالی نرمال، توابع چگالی دیگری نیز می‌توان استفاده کرد.

یکی از رایج‌ترین توابع، تابع لوجستیک است.

برای هر متغیر تصادفی Z تابع توزیع لاجستیک به صورت زیر است،

$$P(Z \leq z) = \frac{1}{1+e^{-Z}} = \frac{e^{Z}}{1+e^{Z}}$$

احتمال آنکه $Y_i = 1$ باشد، برابر است با؛

$$P(Y_i = 1|\boldsymbol{x}_i') = \boldsymbol{G}(\mathbf{x}_i'\boldsymbol{\beta}) = \frac{1}{1+e^{-\mathbf{x}_i'\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

احتمال آنکه $Y_i = 0$ باشد، برابر است با؛

$$P(Y_i = 0|\boldsymbol{x}_i') = 1 - P(Y_i = 1|\boldsymbol{x}_i') = \frac{1}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

2. برآورد پارامترهای مدل

در واقع شکل مدل رگرسیون لوجستیک به صورت روبروست

$$Y_i=\text{logit}(p_i)=\text{Ln}\left(\frac{p_i}{1-p_i}\right)=\mathbf{x}_i'\boldsymbol{\beta}=\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki}$$

برای برآورد پارامترهای مدل از روش MLE استفاده می‌کنیم.

می‌دانیم متغیر تصادفی Y_i دارای توزیع برنولی است.

تابع درستنمایی؛

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}G_i^{Y_i}(1-G_i)^{1-Y_i}$$

برای به دست آوردن برآوردگرهای MLE،

لگاریتم تابع درستنمایی را به دست آورده، از آن مشتق گرفته و برابر با صفر قرار می‌دهیم.

3. تفسیر نتایج مدل لوجیت

احتمال برآورد شده در مدل لوجیت؛

$$p_i = P(Y_i = 1|x_i) = G(\mathbf{x}_i'\boldsymbol{\beta}) = G(\hat{\beta}_۱ + \hat{\beta}_۲ X_{۲i} + \ldots + \hat{\beta}_k X_{ki})$$

- به عبارت دیگر، احتمال اینکه $Y_i = 1$ باشد، بستگی به پارامترهای برآوردشده و مقدار متغیرهای توصیفی دارد.

3. تفسیر نتایج مدل لوجیت

مثال؛ فرض کنید یک مدل تک متغیره داریم (و پارامترها به صورت ۰.۸– و ۰.۰۱ برآورد شده است)؛

Y_i : انتخاب حمل‌ونقل شخصی برای جابجایی

X_i : سطح درآمد فرد i–ام

$$P(Y_i = 1|x_i) = G(-0.8 + 0.01\,X_i)$$

سطح درآمدی X_i=40

$$P(Y_i = 1|x_i = 40) = G\big(-0.8 + 0.01\,(40)\big) = \frac{1}{1 + e^{-(-0.4)}} = 0.401$$

سطح درآمدی X_i=80

$$P(Y_i = 1|x_i = 80) = G\big(-0.8 + 0.01\,(80)\big) = \frac{1}{1 + e^{-(0)}} = 0.5$$

سطح درآمدی X_i=150

$$P(Y_i = 1|x_i = 150) = G\big(-0.8 + 0.01\,(150)\big) = \frac{1}{1+e^{-(0.7)}} = 0.668$$

سطح درآمدی X_i=200

$$P(Y_i = 1|x_i = 200) = G\big(-0.8 + 0.01\,(200)\big) = \frac{1}{1 + e^{-(1.2)}} = 0.7685$$

- بنابراین، با افزایش درآمد، احتمال انتخاب حمل‌ونقل شخصی بیشتر می‌شود.

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

- مقدمه
- پروبیت دوگانه
- لوجیت دوگانه (لوجستیک)
- **معیارهای نیکویی برازش**
- ریسک و شانس
- لوجیت چندگانه
- لوجیت شرطی

❖ نسبت درستنمایی (Likelihood ratio)

برای برآورد پارامترهای مدل‌های پروبیت و لوجیت از روش **MLE** استفاده می‌شود.

بنابراین، با استفاده از نسبت درستنمایی می‌توان هر محدودیتی را مورد آزمون قرار داد.

نسبت درستنمایی برای مقایسه دو رگرسیون غیرمقید و مقید استفاده می‌شود؛

$$LR = 2\,[\ln L_{UR} - \ln L_0]$$

L_{UR} : مقدار تابع درستنمایی غیرمقید (مدل ساخته شده)

L_0 : مقدار تابع درستنمایی مدل ساده (فقط شامل عرض از مبدأ)

نسبت درستنمایی از توزیع χ^2_m پیروی می‌کند که **m** تعداد محدودیت‌ها را نشان می‌دهد.

❖ **شاخص نسبت درستنمایی (Likelihood ratio index) یا ρ^2 مک فادن**

$$\rho_M^2 = LRI = 1 - \frac{\ln L_{UR}}{\ln L_0}$$

L_{UR} : مقدار تابع درستنمایی غیرمقید (مدل ساخته شده)

L_0 : مقدار تابع درستنمایی مدل ساده (فقط شامل عرض از مبدأ)

LRI بین صفر و یک است.

اگر $L_{UR}= L_0$ یعنی قدرت توضیح مدل ساخته شده با یک مدل صفر، یکسان است. $LRI = 0$

$LRI = 1$ به معنای برازش کامل است. (یعنی به ازای $Y_i = 1$ همواره $G(x_i'\beta)=1$ و به ازای $Y_i = 0$ همواره $G(x_i'\beta)=0$ باشد (بدین ترتیب لگاریتم تابع درستنمایی برابر با 0 می‌شود و $LRI = 1$)- این یک **حالت خاص** است.

✓ **بیشترین مقدار LRI زمانی به دست می‌آید که تمام ضرایب متغیرهای توصیفی درون مدل، معنادار باشند.**

❖ **درصد پیش‌بینی صحیح (Percent correctly predicted)**

به ازای هر i احتمال تخمینی را برای وضعیت $Y_i = 1$ حساب می‌کنیم.

این احتمال برابر است با $\Phi\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)$ یا $G\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)$. در حالت کلی به صورت $F\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)$ نمایش می‌دهیم.

برای مقادیر $F\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right) > 0.5$ پیش‌بینی می‌شود که $Y_i = 1$ است.

برای مقادیر $F\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right) \leq 0.5$ پیش‌بینی می‌شود که $Y_i = 0$ است.

درصد مواردی که Y_i پیش‌بینی شده با Y_i مشاهده شده مطابقت دارد= «درصد پیش‌بینی صحیح»

❖ **درصد پیش‌بینی صحیح (Percent correctly predicted)**

مثال؛ در یک نمونه ۲۰۰ تایی، نتایج به دست آمده از مدل لوجیت به صورت جدول زیر است. مطلوبست مجاسبه درصد پیش‌بینی صحیح؟

		پیش‌بینی شده		
		$Y_i = 0$	$Y_i = 1$	جمع
مشاهده شده	$Y_i = 0$	۵۰	۳۰	۸۰
	$Y_i = 1$	۲۰	۱۰۰	۱۲۰
	جمع	۷۰	۱۳۰	۲۰۰

❖ شاخص انحراف (Deviance)

این شاخص معیاری برای مقایسه و انتخاب از میان مدل‌های کاندیدا می‌باشد.

در این روش دو مدل را با هم مقایسه می‌کنیم؛

۱- مدل کامل: مدلی که تعداد پارامترهای آن برابر با تعداد مشاهدات است.

در این حالت، مدل به صورت کامل برازش می‌شود.

مثلا اگر یک مدل رگرسیون با یک متغیر توصیفی داشته باشیم، و فقط ۲ مشاهده داشته باشیم، آنگاه هیچ خطایی وجود ندارد، و پارامترهای مدل به صورت دقیق به دست می‌آید.

در حالت کلی، در مدل کامل، **n** پارامتر و **n** مشاهده داریم.

۲- مدل موردنظر: مدلی ساخته شده که تعداد پارامترهای آن از تعداد مشاهدات کمتر است.

در حالت کلی، به تعداد **p** پارامتر و **n** مشاهده داریم ($p<n$).

$$E(Y_i|x_i) = \beta_۱ + \beta_۲ X_{۲i} + \dots + \beta_p X_{pi}$$

❖ **شاخص انحراف (Deviance)**

برای مقایسه این دو مدل از تابع درستنمایی استفاده می‌کنیم.

$$\boldsymbol{D = 2[\ln L(\mu_i) - \ln L(\beta)]}$$

$L(\mu_i)$ بیشترین مقدار برای تابع درستنمایی است.

کوچک‌تر بودن مقدار D بدین معناست که مدل رگرسیون موردنظر، برازش بهتری بر روی داده‌ها داشته است. زیرا انحراف آن از مدل کامل، کمتر است.

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

مقدمه

پروبیت دوگانه

لوجیت دوگانه (لوجستیک)

معیارهای نیکویی برازش

ریسک و شانس

لوجیت چندگانه

لوجیت شرطی

❖ نسبت ریسک

- **ریسک از مبانی مدل‌سازی داده‌های شمارشی است.**
- **ریسک عبارت است از مواجهه با احتمال یک پیامد.**
- **مثلا ریسک تصادف: احتمال وقوع تصادف**
- **ریسک** به معنای تجربه یک پیامد معین توسط یک فرد در شرایط مشخص است.
- **ریسک** بیانگر احتمال آن است که فرد واقعا آن پیامد را تجربه کند.
- **از طرف دیگر، این احتمال، بیانگر نسبت افرادی است که این پیامد را تجربه می‌کنند.**

❖ نسبت ریسک

- **عامل ریسک: شرط یا شرایطی که ریسک تحت آن اندازه‌گیری می‌شود.**
- **ریسک: معیاری از رابطه میان پیامد و عامل (عوامل) ریسک خاص.**
- **ریسک نسبی (نسبت ریسک):** نسبت احتمال وقوع پیامد برای یک عامل ریسک معین در مقایسه با احتمال وقوع پیامد برای آن‌هایی که آن عامل ریسک را ندارند.
- پس ریسک نسبی، نسبت دو نسبت است.
- وقتی که با داده‌های شمارشی کار می‌کنیم، به آن «**نسبت نرخ وقوع** (**Incident Rate Ratio (IRR)**)» می‌گویند.

❖ نسبت ریسک

○ مثال؛

		(X)		
جمع	۱	۰		
A+B	B	A	۰	(Y)
C+D	D	C	۱	
n	B+D	A+C	جمع	

ریسک $Y=1$ به ازای $X=1$ برابر است با؛

$$\frac{D}{B+D}$$

ریسک $Y=1$ به ازای $X=0$ برابر است با؛

$$\frac{C}{A+C}$$

نسبت ریسک (ریسک نسبی) برای $Y=1$ به ازای $X=1$ در مقایسه با $X=0$ برابر است با؛

$$\text{IRR}=\frac{\frac{D}{B+D}}{\frac{C}{A+C}}=\frac{DA+CD}{CB+CD}$$

❖ نسبت ریسک

○ **مثال** ریسک فوت/مصدومیت در تصادف؛

		سن (X)		
		خردسال: ۰	بزرگسال: ۱	جمع
وضعیت مصدومیت (Y)	بدون مصدومیت: ۰	۴۵	۷۵۵	۸۰۰
	مصدوم/فوت شده: ۱	۵۵	۴۴۵	۵۰۰
	جمع	۱۰۰	۱۲۰۰	۱۳۰۰

ریسک فوت/مصدومیت برای بزرگسالان:

$$\frac{445}{1200} = 0.371$$

ریسک فوت/مصدومیت برای خردسالان

$$\frac{55}{100} = 0.55$$

نسبت ریسک (**ریسک نسبی**) فوت/مصدومیت بزرگسالان در مقایسه با خردسالان؛

$$\frac{0.371}{0.55} = 0.6742$$

		(X)		
		۰	۱	جمع
(Y)	۰	A	B	A+B
	۱	C	D	C+D
	جمع	A+C	B+D	n

❖ تفسیر: احتمال فوت/مصدومیت برای بزرگسالان نسبت به خردسالان ۳۲.۶٪ کمتر است.

❖ فاصله اطمینان برای ریسک

محاسبه فاصله اطمینان نیازمند انحراف معیار می‌باشد.

		خردسال: ۰	بزرگسال: ۱	جمع
		سن (X)		
وضعیت مصدومیت (Y)	بدون مصدومیت: ۰	۴۵	۷۵۵	۸۰۰
	مصدوم/فوت شده: ۱	۵۵	۴۴۵	۵۰۰
	جمع	۱۰۰	۱۲۰۰	۱۳۰۰

در جدول ۲*۲ لگاریتم انحراف معیار نسبت ریسک را می‌توان به صورت زیر حساب کرد؛

$$\ln(SE(IRR)) = ln\sqrt{\frac{1}{D}+\frac{1}{B+D}+\frac{1}{C}+\frac{1}{A+C}} = ln\sqrt{\frac{1}{445}+\frac{1}{1200}+\frac{1}{55}+\frac{1}{100}} = 0.1768$$

قبلا نسبت ریسک را برابر با 0.6742 به دست آوردیم؛ لگاریتم نسبت ریسک: $\ln(0.6742) = -0.3942$

فاصله اطمینان با فرض توزیع نرمال: $RR \pm Z\alpha_{/2}\, SE\ (IRR)$

فاصله اطمینان ۹۵درصدی برای لگاریتم نسبت ریسک: $-0.3942 \pm 1.96 * 0.1768 \rightarrow (-0.741\ , -0.048)$

فاصله اطمینان ۹۵درصدی برای نسبت ریسک: $(e^{-0.741}\ , e^{-0.048}) = (0.477\ , 0.954)$

		خردسال: ۰	بزرگسال: ۱	جمع
		سن (X)		
وضعیت مصدومیت (Y)	بدون مصدومیت: ۰	۴۵	۷۵۵	۸۰۰
	مصدوم/فوت شده: 1	۵۵	۴۴۵	۵۰۰
	جمع	۱۰۰	۱۲۰۰	۱۳۰۰

❖ تفاضل ریسک

تفاضل ریسک بیانگر کاهش مطلق ریسک است.

این تفاضل را به عنوان «**معیار کاهش ریسک**» درنظر می‌گیریم.

ریسک فوت/مصدومیت برای بزرگسالان:

$$\frac{445}{1200} = 0.371$$

ریسک فوت/مصدومیت برای خردسالان

$$\frac{55}{100} = 0.55$$

تفاضل ریسک:

$$0.371 - 0.55 = -0.179$$

❖ **تفسیر: نرخ فوت/مصدومیت بزرگسالان حدود ۱۸٪ کمتر از خردسالان است.**

❖ نسبت ریسک

○ **مثال** ریسک فوت/مصدومیت در تصادف؛

		سن (X)			
		خردسال: ۰	میانسال: ۱	بزرگسال: ۲	جمع
وضعیت مصدومیت (Y)	بدون مصدومیت: 0	۱۲۰	۱۶۰	۵۲۰	۸۰۰
	مصدوم/فوت شده: 1	۲۰۰	۱۲۰	۱۸۰	۵۰۰
	جمع	۳۲۰	۲۸۰	۷۰۰	۱۳۰۰

اکنون یک **متغیر سه سطحی** را درنظر بگیرید.

- **در اینجا باید یک ستون را به عنوان مبنا انتخاب کنیم.**
- **ستون ۳: بزرگسال را به عنوان مرجع انتخاب می‌کنیم.**

نسبت ریسک فوت/مصدومیت برای گروه ۲ (میانسال)

$$\frac{\frac{120}{280}}{\frac{180}{700}} = 1.6667$$

نسبت ریسک فوت/مصدومیت برای گروه ۱ (خردسال)

$$\frac{\frac{200}{320}}{\frac{180}{700}} = 2.4305$$

❖ نسبت شانس (بخت)

شانس وقوع یک واقعه برابر است با احتمال وقوع آن تقسیم بر احتمال عدم وقوع آن؛

$$\text{شانس وقوع} = \frac{p}{1-p}$$

P **احتمال وقوع است.**

توجه: در مدل رگرسیون لوجستیک در واقع لگاریتم بخت (Odds) را مدل می‌کنیم:

$$Y_i = \text{logit}(p_i) = \text{Ln}\left(\frac{p_i}{1-p_i}\right) = \text{Ln}\,(Odds_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_۱ + \beta_۲ X_{۲i} + \ldots + \beta_k X_{ki}$$

وقتی شانس وقوع را در هر سطحی از عامل ریسک (X) مقایسه می‌کنیم، نسبت بین شانس وقوع $X = 1$ با شانس وقوع $X = 0$ معروف به «**نسبت شانس Odds Ratio (OR)**» است؛

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

❖ وقتی که احتمال $X = 0$ (یعنی p_0) کوچک باشد (مثلا کمتر از ۰.۱)، آنگاه **نسبت شانس** و **نسبت ریسک** خیلی نزدیک به هم خواهند بود.

❖ نسبت شانس (بخت)

		(X)		
		۰	۱	جمع
(Y)	۰	A	B	A+B
	۱	C	D	C+D
	جمع	A+C	B+D	n

شانس وقوع $(Y=1)$ برای $X=1$ برابر است با؛

$$\frac{p_1}{1-p_1}=\frac{\frac{D}{B+D}}{\frac{B}{B+D}}=\frac{D}{B}$$

شانس وقوع $(Y=1)$ برای $X=0$ برابر است با؛

$$\frac{p_0}{1-p_0}=\frac{\frac{C}{A+C}}{\frac{A}{A+C}}=\frac{C}{A}$$

نسبت شانس وقوع $(Y=1)$ برای $X=1$ در مقایسه با $X=0$ برابر است با؛

$$\mathrm{OR}=\frac{\frac{D}{B}}{\frac{C}{A}}=\frac{DA}{CB}$$

❖ نسبت شانس (بخت)

		(X)		
	۰	۱	جمع	
(Y) ۰	A	B	A+B	
۱	C	D	C+D	
جمع	A+C	B+D	n	

نسبت شانس وقوع ($Y = 1$) برای $X = 1$ در مقایسه با $X = 0$

$$OR=\frac{\frac{D}{B}}{\frac{C}{A}}=\frac{DA}{CB}$$

صورت کسر: حاصلضرب (تعداد مواردی که $Y = 1$ و $X = 1$) ❊ (تعداد مواردی که $Y = 0$ و $X = 0$)

مخرج کسر: حاصلضرب (تعداد مواردی که $Y = 1$ و $X = 0$) ❊ (تعداد مواردی که $Y = 0$ و $X = 1$)

نسبت ریسک = $IRR=\dfrac{DA+CD}{CB+CD}$

صورت و مخرج تقسیم بر CB

$$IRR=\frac{DA+CD}{CB+CD}=\frac{OR+\frac{D}{B}}{1+\frac{D}{B}}$$

			سن (X)		
		خردسال: ۰	بزرگسال: ۱	جمع
وضعیت مصدومیت (Y)	بدون مصدومیت: ۰	۴۵	۷۵۵	۸۰۰
	مصدوم/فوت شده: ۱	۵۵	۴۴۵	۵۰۰
	جمع	۱۰۰	۱۲۰۰	۱۳۰۰

❖ نسبت شانس

مثال شانس فوت/مصدومیت در تصادف؛

شانس فوت/مصدومیت برای بزرگسالان:

$$\frac{445}{775} = 0.5894$$

شانس فوت/مصدومیت برای خردسالان

$$\frac{55}{45} = 1.222$$

نسبت ریسک: **ریسک** فوت/مصدومیت بزرگسالان نسبت به **ریسک** فوت/مصدومیت خردسالان؛

$$\frac{0.371}{0.55} = 0.6742$$

نسبت شانس فوت/مصدومیت بزرگسالان در مقایسه با خردسالان؛

$$OR = \frac{0.5894}{1.222} = 0.48224$$

نسبت شانس فوت/مصدومیت خردسالان در مقایسه با بزرگسالان؛

$$OR = \frac{1.222}{0.5894} = \frac{1}{0.48224} = 2.07$$

تفسیر: شانس فوت/مصدومیت خردسالان بیش از ۲ برابرِ شانس فوت/مصدومیت بزرگسالان است.

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

- مقدمه
- پروبیت دوگانه
- لوجیت دوگانه (لوجستیک)
- معیارهای نیکویی برازش
- ریسک و شانس
- **لوجیت چندگانه**
- لوجیت شرطی

مدل‌های لوجیت– لوجیت چندگانه

❖ **شرایطی که پدیده مورد بررسی، بیش از دو حالت دارد. مثلاً؛**

- مسیر انتخاب شده از بین ۳ مسیر آلترناتیو،
- نوع تصادف (صرفاً خسارتی، جرحی خفیف، جرحی شدید، فوتی)
- شیوه حمل‌ونقل انتخابی (تاکسی، اتوبوس، مترو، خودرو شخصی)
- ... ؟

- در این شرایط یکی از سطوح ($Y=0$) را به عنوان مرجع درنظر گرفته، و بقیه سطوح را با آن مقایسه می‌کنیم.
- از مدل لوجیت چندگانه برای بررسی این پدیده‌ها استفاده می‌کنیم.

❖ **مدل لوجیت سه گانه**

○ **فرض کنید p متغیر توصیفی داریم.**

$$\mathbf{x}_i' = [\quad X_{1i} \quad X_{2i} \quad \cdots \quad X_{pi}\,] \quad ; \quad X_{1i} = 1$$

تابع لوجیت برای Y=1 را به صورت زیر تعریف می‌کنیم.

$$g_1(\mathbf{x}_i') = ln\frac{P(Y=1|\mathbf{x}_i')}{P(Y=0|\mathbf{x}_i')} = \beta_{11}+\beta_{12}X_{2i}+\ldots+\beta_{1p}X_{pi} + u_i = \mathbf{x}_i'\boldsymbol{\beta}_1$$

برای Y=2 را به صورت زیر تعریف می‌کنیم.

$$g_1(\mathbf{x}_i') = ln\frac{P(Y=2|\mathbf{x}_i')}{P(Y=0|\mathbf{x}_i')} = \beta_{21}+\beta_{22}X_{2i}+\ldots+\beta_{2p}X_{pi} + u_i = \mathbf{x}_i'\boldsymbol{\beta}_2$$

❖ **مدل لوجیت سه گانه**

○ **احتمال‌های شرطی برای هر پیامد عبارت است از:**

$$P(Y=0|\mathbf{x}_i') = \frac{1}{1+e^{g_1(\mathbf{x}_i')}+e^{g_2(\mathbf{x}_i')}}$$

$$P(Y=1|\mathbf{x}_i') = \frac{e^{g_1(\mathbf{x}_i')}}{1+e^{g_1(\mathbf{x}_i')}+e^{g_2(\mathbf{x}_i')}}$$

$$P(Y=2|\mathbf{x}_i') = \frac{e^{g_2(\mathbf{x}_i')}}{1+e^{g_1(\mathbf{x}_i')}+e^{g_2(\mathbf{x}_i')}}$$

برای سادگی و اختصار، از نام‌گذاری‌های زیر استفاده می‌کنیم؛

$$\pi_j(\mathbf{x}_i') = P(Y=j|\mathbf{x}_i') = \frac{e^{g_j(\mathbf{x}_i')}}{\sum_{k=0}^{2} e^{g_k(\mathbf{x}_i')}} \quad ; \quad j=0,1,2 \quad ; g_0(\mathbf{x}_i') = \mathbf{0}$$

بنابراین، هر احتمال تابعی از بردار $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1' , \boldsymbol{\beta}_2')$ است؛ که تعداد ضرایب برابر با $2p$ است.

ضرایب این مدل با روش حداکثر درستنمایی برآورد می‌شود.

❖ **مدل لوجیت سه گانه**

- **برای تشکیل تابع درستنمایی، سه متغیر دوتایی می‌سازیم که هر یک شامل 0 و ۱ می‌شود.**
- **توجه:** این متغیرها را فقط برای تشکیل تابع درستنمایی استفاده می‌کنیم و در تفسیر مدل لوجیت کاربردی ندارد.
- متغیرها به صورت روبرو تعریف می‌شوند؛

$$Y=0 \qquad Y_0=1,\ Y_1=0,\ Y_2=0$$

$$Y=1 \qquad Y_0=0,\ Y_1=1,\ Y_2=0$$

$$Y=2 \qquad Y_0=0,\ Y_1=0,\ Y_2=1$$

- جمع این متغیرها باید برابر با ۱ باشد؛

$$\sum_{j=0}^{2} Y_j = 1$$

- تابع درستنمایی (برای n مشاهده مستقل)؛

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\,\pi_0(\mathbf{x}_i')^{Y_{0i}}\pi_1(\mathbf{x}_i')^{Y_{1i}}\pi_2(\mathbf{x}_i')^{Y_{2i}}]$$

❖ **نسبت شانس**

○ **فرض کنید ۳ حالت برای متغیر تصادفی Y متصور است؛ که عبارتند از،** $Y = j \quad ; \quad j = 0,1,2$

$Y = j = 0$ را به عنوان حالت مرجع در نظر گرفته و دو حالت دیگر را با آن مقایسه می‌کنیم.

اگر X تغییر کند موجب تغییر در $P(Y = j)$ می‌شود.

فرض کنید X=a باشد. آنگاه؛

$$X = a \Rightarrow \begin{cases} P(Y = 0 \mid X = a) \\ P(Y = j \mid X = a) \quad\quad ; j = 1,2 \end{cases}$$

بنابراین، تغییر در احتمال به معنای تغییر از $P(Y = 0 \mid X = a)$ به $P(Y = j \mid X = a)$ است.

حال اگر X تغییر کند (مثلا از a به b) در این‌صورت $P(Y = j \mid X = b)$ را داریم.

بنابراین، نسبت شانس برابر است با؛

$$OR_j(a,b) = \frac{P(Y = j \mid X = b)/P(Y = 0 \mid X = b)}{P(Y = j \mid X = a)/P(Y = 0 \mid X = a)} \quad ; j = 1,2$$

❖ **نسبت شانس**

نسبت شانس برابر است با؛

$$OR_j(a,b) = \frac{P(Y=j \mid X=b)/P(Y=0 \mid X=b)}{P(Y=j \mid X=a)/P(Y=0 \mid X=a)} \quad ; \; j=1,2$$

○ بنابراین، **پیامد $\mathbf{Y}=\boldsymbol{j}$ را در مقایسه با $\mathbf{Y=0}$ به ازای $\mathbf{X}=\boldsymbol{b}$ در مقابل $\mathbf{X}=\boldsymbol{a}$ می‌سنجیم**.

اگر متغیر توصیفی مقادیر دوگانه 0 و ۱ را داشته باشد؛ نسبت شانس به صورت زیر محاسبه خواهد شد؛

$$OR_j(0,1) = \frac{P(Y=j \mid X=1)/P(Y=0 \mid X=1)}{P(Y=j \mid X=0)/P(Y=0 \mid X=0)} \quad ; \; j=1,2$$

❖ **مثال: فرض کنید جدول فراوانی سطوح دو متغیر X و Y را داریم؛**

وقتی سه پیامد وجود داشته باشد، دو مدل لوجیت داریم.

این توابع را به گونه‌ای تعریف می‌کنیم که دو ضریب برآوردشده (یکی برای هر تابع لوجیت) برابر با لگاریتم نسبت شانس برای هر جدول 2*2 باشند که از طبقه‌بندی Y=j و Y=0 به دست می‌آید.

Y=0 را به عنوان مرجع درنظر می‌گیریم.

		(X)			OR
		۰	۱	جمع	
(Y)	۰	۸۰	۱۷۱	۲۵۱	$OR_0 = 1$
	۱	۲۵	۱۰۵	۱۳۰	$OR_1 = 1.965$
	2	۱۵	۱۰۴	۱۱۹	$OR_2 = 3.244$
	جمع	۱۲۰	۳۸۰	۵۰۰	

$$OR_1 = \frac{\frac{105}{171}}{\frac{25}{80}} = 1.965$$

$$OR_2 = \frac{\frac{104}{171}}{\frac{15}{80}} = 3.244$$

❖ **مثال؛**

از طرف دیگر، احتمال‌های زیر را داریم؛

$$P(Y=0)=\frac{1}{1+e^{g_1(\boldsymbol{X})}+e^{g_2(\boldsymbol{X})}}$$

$$P(Y=1)=\frac{e^{g_1(\boldsymbol{X})}}{1+e^{g_1(\boldsymbol{X})}+e^{g_2(\boldsymbol{X})}}$$

$$P(Y=2)=\frac{e^{g_2(\boldsymbol{X})}}{1+e^{g_1(\boldsymbol{X})}+e^{g_2(\boldsymbol{X})}}$$

نسبت شانس برای Y=1 که با OR_1 نشان می‌دهیم، به ازای تغییر X از a به b؛

$$OR_1(a,b)=\frac{P(Y=1\mid X=b)/P(Y=0\mid X=b)}{P(Y=1\mid X=a)/P(Y=0\mid X=a)}=\frac{e^{g_1(\boldsymbol{X=b})}}{e^{g_1(\boldsymbol{X=a})}}=e^{g_1(\boldsymbol{X=b})-g_1(\boldsymbol{X=a})}\ ;\ j=1,2$$

اگر $g_1(\boldsymbol{X})$ برابر با $g_1(X)=\alpha_1+\beta_1 X$ باشد، آنگاه، نسبت شانس برابر است با؛

$$OR_1(a,b)=e^{(\alpha_1+\beta_1\boldsymbol{b})-(\alpha_1+\beta_1\boldsymbol{a})}=e^{\beta_1(\boldsymbol{b}-\boldsymbol{a})}$$

حالت خاص: متغیر X فقط دو مقدار داشته باشد: X=0,1

$$OR_1(0,1)=e^{\beta_1}$$

در حالت کلی؛

اگر $\mathrm{g}_j(\boldsymbol{X})$ برابر با $\mathrm{g}_j(X) = \alpha_j + \beta_j X$ باشد (معادله تک‌متغیره باشد)، آنگاه، نسبت شانس برابر است با؛

$$OR_1(a,b) = e^{(\alpha_1+\beta_1\boldsymbol{b})-(\alpha_1+\beta_1\boldsymbol{a})} = e^{\beta_1(\boldsymbol{b}-\boldsymbol{a})}$$

$$OR_2(a,b) = e^{(\alpha_2+\beta_2\boldsymbol{b})-(\alpha_2+\beta_2\boldsymbol{a})} = e^{\beta_2(\boldsymbol{b}-\boldsymbol{a})}$$

حالت خاص: متغیر X فقط دو مقدار داشته باشد: X=0,1

$$OR_1(0,1) = e^{\beta_1}$$

$$OR_2(0,1) = e^{\beta_2}$$

بنابراین، **نسبت شانس در حالت کلی** برابر است با؛

$$OR_j(a,b) = e^{\mathrm{g}_j(\boldsymbol{X}=\boldsymbol{b})-\mathrm{g}_j(\boldsymbol{X}=\boldsymbol{a})}$$

درحالتی که X=0,1 باشد ($\Delta X = b - a = 1$) برآورد نسبت شانس ($\widehat{OR}_j$) برابر است با

$$\widehat{OR}_j(a,b) = e^{\hat{\beta}_j}$$

لگاریتم گیری

$$ln\widehat{OR}_j = \hat{\beta}_j$$

- **یعنی در این حالت، ضریب شیبِ برآورد شده (برای مدل لوجیت $\boldsymbol{j}$) برابر با لگاریتم شانس است.**

برآورد فاصله اطمینان؛

فاصله اطمینان ۹۵درصدی برای $\hat{\beta}_j$

$$\hat{\beta}_j \pm 1.96\,SE(\hat{\beta}_j)$$

فاصله اطمینان ۹۵درصدی برای $\widehat{OR}_j$

$$Exp(\hat{\beta}_j \pm 1.96\,SE\big(\hat{\beta}_j\big))$$

مثال: فرض کنید از داده‌های جدول روبرو، مدل‌های زیر برآورد شده است؛

		(X)			OR
		۰	۱	جمع	
(Y)	۰	۸۰	۱۷۱	۲۵۱	$OR_0 = 1$
	۱	۲۵	۱۰۵	۱۳۰	$OR_1 = 1.965$
	2	۱۵	۱۰۴	۱۱۹	$OR_2 = 3.244$
	جمع	۱۲۰	۳۸۰	۵۰۰	

مقادیر OR با رابطه $OR_j = e^{\beta_j}$ به دست می‌آید.

انحراف معیار ضرایب برابر است با؛

جذر مجموع معکوس فراوانی‌ها

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{80} + \frac{1}{171} + \frac{1}{25} + \frac{1}{104}} = 0.261$$

مدل لوجیت	متغیر	ضرایب	برآورد ضرایب	انحراف معیار ضرایب	نسبت شانس	فاصله اطمینان ۹۵ برای نسبت شانس
۱	ثابت	α_1	-1.125	0.225	1.965	1.177 , 3.276
	X	β_1	0.675	0.261		
۲	ثابت	α_2	-1.545	0.25	3.244	1.1774 , 5.922
	X	β_2	1.177	0.308		

فاصله اطمینان ۹۵درصدی برای $\widehat{OR}_j$:

$$OR_1 : \ Exp\left(\hat{\beta}_1 \pm 1.96\, SE(\hat{\beta}_1)\right) = (1.177\ , 3.276)$$

❖ **حالتی که X بیش از دو وضعیت داشته باشد**

○ **مثال:** فرض کنید جدول فراوانی سطوح دو متغیر X و Y را داریم؛

		(X)				
		٥	١	٢	٣	جمع
(Y)	٥	٣٠	٥٠	۶٠	١٠٠	٢۴٠
	1	١۵	٣٠	۵٠	۴٠	١٣۵
	2	۵	٢٠	۴٠	۶٠	١٢۵
	جمع	۵٠	١٠٠	١۵٠	٢٠٠	۵٠٠

حالت Y=0 و X=0 را به عنوان مرجع درنظر می‌گیریم.

نسبت شانس‌ها به صورت زیر به دست می‌آید؛

$$OR_1(1,0)=\frac{\frac{30}{50}}{\frac{15}{30}}=1.2$$

$$OR_2(1,0)=\frac{\frac{20}{50}}{\frac{5}{30}}=2.4$$

$$OR_1(2,0)=\frac{\frac{50}{60}}{\frac{15}{30}}=1.67$$

$$OR_2(2,0)=\frac{\frac{40}{60}}{\frac{5}{30}}=4$$

$$OR_1(3,0)=\frac{\frac{40}{100}}{\frac{15}{30}}=0.8$$

$$OR_2(3,0)=\frac{\frac{60}{100}}{\frac{5}{30}}=3.6$$

❖ **حالتی که X بیش از دو وضعیت داشته باشد**

○ **مثال:** فرض کنید جدول فراوانی سطوح دو متغیر X و Y را داریم؛

		(X)				
		٥	۱	۲	۳	جمع
(Y)	٥	۳۰	۵۰	۶۰	۱۰۰	۲۴۰
	1	۱۵	۳۰	۵۰	۴۰	۱۳۵
	2	۵	۲۰	۴۰	۶۰	۱۲۵
	جمع	۵۰	۱۰۰	۱۵۰	۲۰۰	۵۰۰

از طرف دیگر، نسبت شانس برابر است با e به توان ضریب رگرسیون لوجیت

$$OR_j(i,o) = e^{\beta_{ji}} \quad ; \quad j = 1,2 \quad ; \quad i = 1,2,3$$

j حالت‌های Y و i حالت‌های X را نشان می‌دهد.

به ازای یک واحد افزایش در متغیر X (و ثابت ماندن سایر متغیرها) بخت وقوع پیامد j نسبت به حالت پایه، به میزان $e^{\beta_{ji}}$ افزایش می‌یابد

❖ **آزمون برابری دو نسبت شانس**

فرضیه صفر برابری دو نسبت شانس: $H_0\colon OR_1 - OR_2 = 0 \Rightarrow H_0\colon OR_1 = OR_2$

این فرضیه معادل آزمون فرضیه زیر است:

$$H_0\colon \ln(OR_1) = \ln(OR_2) \quad \text{یا} \quad H_0\colon \frac{OR_1}{OR_2} = 1$$

از رابطه تفاضل بین دو شیب معادلات رگرسیون لوجیت استفاده می‌کنیم؛

$$\ln(\widehat{OR}_2) - \ln(\widehat{OR}_1) = \hat{\beta}_2 - \hat{\beta}_1$$

اگر فرضیه صفر درست باشد، آنگاه این تفاضل از توزیع نرمال با میانگین و واریانس زیر پیروی می‌کند؛

$$E(\hat{\beta}_2 - \hat{\beta}_1) = \beta_2 - \beta_1$$

$$Var(\hat{\beta}_2 - \hat{\beta}_1) = Var(\hat{\beta}_2) - Var(\hat{\beta}_1) - 2Cav(\hat{\beta}_1, \hat{\beta}_2)$$

آماره t را به صورت روبرو تعریف می‌کنیم

$$t = \frac{\hat{\beta}_2 - \hat{\beta}_1}{\sqrt{var(\hat{\beta}_2 - \hat{\beta}_1)}}$$

❖ **آزمون برابری دو نسبت شانس**

فاصله اطمینان ۹۵درصدی برای $(\beta_2 - \beta_1)$

$$(\hat{\beta}_2 - \hat{\beta}_1) \pm 1.96\sqrt{Var(\hat{\beta}_2 - \hat{\beta}_1)}$$

اگر حد بالا و پایین این فاصله اطمینان برابر با L_1 و L_2 باشد، آنگاه

فاصله اطمینان ۹۵درصدی برای تفاضل $OR_1 - OR_2$ برابر است با؛

$$(e^{L_1}, e^{L_2})$$

سوال:

در چه صورتی می‌گوییم OR_1 و OR_2 تفاوت معناداری ندارند؟

❖ آزمون معنادار بودن شیب رگرسیون لوجیت

- برای آزمون معنادار بودن شیب رگرسیون لوجیت (ضریب X) از نسبت درستنمایی استفاده می‌کنیم.
- این آزمون معادل با مقایسه دو رگرسیون است؛
  - رگرسیونی که شامل X است (رگرسیون غیرمقید)
  - رگرسیونی که فاقد X است (رگرسیون مقید)

این کار را برای هر یک از مدل‌های لوجیت انجام می‌دهیم.

- برای رگرسیون غیرمقید، نسبت درستنمایی را با L_1 نشان می‌دهیم.
- برای رگرسیون مقید، نسبت درستنمایی را با L_0 نشان می‌دهیم.

فرضیه صفر: ضریب X صفر است. مثلا برای رگرسیون لوجیت ۱ فرضیه صفر: $H_0: \beta_1 = 0$

درصورت درست بودن فرضیه صفر، نسبت درستنمایی λ (رابطه زیر) از توزیع χ^2 با ۲ درجه آزادی پیروی می‌کند.

$$\lambda = -2(L_0 - L_1)$$

درحالت کلی، اگر تعداد حالت‌های متغیر وابسته برابر با s و تعداد گروه‌های متغیر X برابر با k باشد، درجه آزادی برابر با $(k-1)(s-1)$ است.

مدل رگرسیون داده‌های گسسته

فصل یازدهم

Discrete data Regression Model

- مقدمه
- پروبیت دوگانه
- لوجیت دوگانه (لوجستیک)
- معیارهای نیکویی برازش
- ریسک و شانس
- لوجیت چندگانه
- **لوجیت شرطی**

❖ تصور کنید افراد برای رفت و آمد روزانه می‌توانند از چهار وسیله‌نقلیه استفاده کنند؛

- خودرو شخصی
- تاکسی
- اتوبوس
- مترو

بنابراین فرد i دارای چهار گزینه است **j=1,2,3,4**

اگر Y_i بیانگر انتخاب فرد i باشد، آنگاه؛

Y_i می‌تواند شامل چهار مقدار ۱، ۲، ۳، ۴ باشد:

$$Y_{i1} = 1 \quad , \quad Y_{i2} = Y_{i3} = Y_{i4} = 0$$

$$Y_{i2} = 1 \quad , \quad Y_{i1} = Y_{i3} = Y_{i4} = 0$$

$$Y_{i3} = 1 \quad , \quad Y_{i1} = Y_{i2} = Y_{i4} = 0$$

$$Y_{i4} = 1 \quad , \quad Y_{i1} = Y_{i2} = Y_{i3} = 0$$

احتمال انتخاب گزینه j بستگی به خصوصیات فرد i دارد.

اگر بردار متغیرهای توصیفی برای انتخاب فرد i را با x نشان دهیم، شامل متغیرهایی مانند درآمد، سن، شغل، جنسیت و غیره خواهد بود.

اگر بردار متغیرهای توصیفی برای انتخاب فرد i را با $\mathbf{x}$ نشان دهیم، شامل متغیرهایی مانند درآمد، سن، شغل، جنسیت و غیره خواهد بود.

در این حالت، یک مدل لوجیت چندگانه را به صورت زیر تعریف می‌کنیم؛

$$\pi_j(\mathbf{x}_i') = P(Y = j|\mathbf{x}_i') = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}_j}}{\sum_{j=1}^{4} e^{\mathbf{x}_i'\boldsymbol{\beta}_j}} \qquad ; \quad i = 0,1, \dots, n$$

توجه شود که در اینجا فقط ویژگی‌های فردی است که اهمیت دارند، و هیچ توجهی به ویژگی‌های گزینه‌ها (در اینجا ویژگی‌های شیوه حمل‌ونقلی) نداریم.

اکنون فرض کنید بخواهیم ویژگی‌های خاص گزینه‌ها نیز درون مدل قرار دهیم.

مثلا فرض کنید دو متغیر هزینه (C_{ij}) و زمان سفر (T_{ij}) برای گزینه‌های موردنظر، متفاوت است.

بدین ترتیب، تصمیم‌گیری و انتخاب هر گزینه برای فرد i براساس ویژگی‌های گزینه‌هاست.

❖ چنین مدلی معروف به **مدل لوجیت شرطی (Conditional Logit)** است.

توجه شود که ویژگی‌های گزینه j (مثلا هزینه گزینه j) می‌تواند برای فرد i متفاوت با سایر افراد باشد (مثلا قیمت متفاوت اتوبوس در مناطق مختلف شهری و یا تخفیف برای دانش‌آموزان)

اگر قیمت‌ها برای تمام افراد یکسان باشد، آنگاه $C_{ij} = C_j$ خواهد شد.

تصور کنید که Y_i بیانگر انتخاب بین m گزینه باشد.

از طرف دیگر، U_{ij} را مطلوبیت انتخاب گزینه j برای فرد i در نظر می‌گیریم.

U_{ij} یک متغیر تصادفی است که شامل دو جزء است:

V_{ij} جزء غیرتصادفی

ε_{ij} جزء تصادفی

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

فرض: فرد عاقلانه عمل می‌کند و گزینه‌ای را انتخاب می‌کند که بیشترین مطلوبیت را برای او دارد.

پس گزینه j درصورتی انتخاب می‌شود که U_{ij} برابر با بزرگترین مقدار از بین $(U_{i1}, \ldots, U_{im})$ باشد.

$$p_{ij} = P(Y_i = j) = P(U_{ij} = \max(U_{i1}, \ldots, U_{im}))$$

بر این مبنا، ثابت می‌شود که متغیر تصادفی ε_{ij} دارای تابع چگالی گامبل است، که عبارتست از؛

$$f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{\varepsilon_{ij}}} = e^{(-\varepsilon_{ij} - e^{\varepsilon_{ij}})}$$

از طرف دیگر، p_{ij} برای مقایسه دو گزینه $\boldsymbol{j}$ و $\boldsymbol{k}$ توسط فرد $\boldsymbol{i}$ عبارتست از؛

$$\begin{aligned} p_{ij} = P(Y_i = j) &= P(U_{ij} \geq U_{ik}) \\ &= P(V_{ij} + \varepsilon_{ij} \geq V_{ik} + \varepsilon_{ik}) \\ &= P(\varepsilon_{ij} - \varepsilon_{ik} \geq V_{ik} - V_{ij}) = P(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik}) \end{aligned}$$

یعنی احتمال اینکه تفاضل $\varepsilon_{ik} - \varepsilon_{ij}$ کوچکتر از $V_{ij} - V_{ik}$ باشد.

چون ε_{ij} و ε_{ik} دو متغیر تصادفی هستند که از توزیع گامبل پیروی می‌کنند، ثابت می‌شود که تفاضل آنها (یعنی $\varepsilon_{ij} - \varepsilon_{ik}$) از توزیع لوجستیک پیروی می‌کند.

$$p_{ij} = \frac{e^{V_{ij}-V_{ik}}}{1 + e^{V_{ij}-V_{ik}}} = \frac{e^{V_{ij}}}{e^{V_{ij}} + e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_{j=1}^{2} e^{V_{ij}}}$$

حالت کلی:
مقایسه U_{ij} با $(U_{i1}, \ldots, U_{im})$

$$p_{ij} = \frac{e^{V_{ij}}}{\sum_{k=1}^{m} e^{V_{ik}}}$$

اگر V_{ij} را بر حسب ویژگی‌های گزینه‌ها تعریف کنیم، خواهیم داشت؛

$$V_{ij} = \mathbf{x}_i'\boldsymbol{\beta} = \alpha_j + \beta_1 C_{ij} + \beta_2 T_{ij}$$

$$p_{ij} = \frac{e^{\alpha_j+\beta_1 C_{ij}+\beta_2 T_{ij}}}{\sum_{j=1}^{4} e^{\alpha_j+\beta_1 C_{ij}+\beta_2 T_{ij}}}$$

نکته: از آنجا که پایه مدل لوجیت شرطی، نظریه مطلوبیت تصادفی و مقایسه مطلوبیت حاصل از گزینه‌هاست، و p_{ij} برحسب تفاضل مطلوبیت گزینه‌ها بیان شده است.

بنابراین، فقط می‌توانیم «تفاضل α_j ها» را حساب کنیم؛ و امکان برآورد هر یک از α_j ها به صورت جداگانه نیست.

بر این اساس، اگر گزینه ۱ را به عنوان مبنا در نظر بگیریم، می‌توان α_1 را نرمال‌سازی نمود و برابر با صفر در نظر گرفت؛ سپس، بقیه α_j ها برآورد نمود.

❖ داده‌های مورد نیاز

در مدل لوجیت شرطی با مسأله تجربه یا آزمایش انتخاب گسسته (**Discrete Choice Experiment**) مواجه هستیم.

داده‌های موردنیاز در این حوزه مربوط به ترجیحات افراد است.

○ بر این مبنا، دو نوع داده قابل جمع‌آوری و استفاده است؛

۱- داده‌هایی که به طور مستقیم جمع‌آوری می‌شود؛ و بنابراین، نشان دهنده انتخاب‌ها (رفتارها)ی واقعی افراد است که در شرایط واقعی از خود نشان داده‌اند.

در این حالت از «ترجیحات آشکارشده» (**Revealed preferences**) استفاده می‌شود. مثلا: گردآوری داده‌های واقعی استفاده از مترو، اتوبوس، تاکسی و غیره.

۲- داده‌هایی که از طریق پرسشنامه گردآوری می‌شوند. برای این منظور، فرد را در موقعیت موردنظر قرار داده و سپس از او در خصوص ترجیحاتش سوال می‌گردد.

در این حالت از «ترجیحات اظهارشده یا بیان شده» (**Stated preferences**) زیر سناریوهای مختلف استفاده می‌شود. مثلا فرد را در موقعیت انتخاب بین خودرو شخصی و اتوبوس قرار می‌دهیم و در مورد تمایل به پرداخت هزینه در نسبت با کاهش زمان جابجایی سوالات مختلفی می‌پرسیم.

❖ مفروضات در مورد ε_{ij}

فرض می‌شود ε_{ij} ها هم در میان گزینه‌ها (j) هم در میان مشاهدات (i) مستقل هستند.

بنابراین، ε_{ij}ها **iid** بوده و فرض می‌شود دارای توزیع گامبل هستند.

❖ برآورد مدل لوجیت شرطی

برآورد مدل لوجیت شرطی با استفاده از روش MLE

احتمال انتخاب گزینه‌ها برای فرد i برابر است با؛

$$L_i = P(Y_{i1} = 1, Y_{i2} = 0, \dots Y_{im} = 0) \times P(Y_{i1} = 0, Y_{i2} = 1, \dots Y_{im} = 0) \times \cdots \times P(Y_{i1} = 0, Y_{i2} = 0, \dots Y_{im} = 1) = \prod_{j=1}^{m} p_{ij}^{Y_{ij}}$$

چون n فرد و m گزینه داریم، تابع درستنمایی عبارتست از ؛

$$L = L_1 \times L_2 \times \cdots \times L_n = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{Y_{ij}}$$

لگاریتم تابع درستنمایی عبارتست از ؛

$$l = \ln L = \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij}\, ln\, p_{ij}$$

❖ تفسیر نتایج مدل لوجیت شرطی

برای محاسبه اثرات نهایی تغییر در X_{ij} (بر احتمال انتخاب گزینه ***j***)، از p_{ij} نسبت به X_{ij} مشتق می‌گیریم؛

$$\frac{\partial\, P(\mathrm{Y}_i = j)}{\partial\, \mathrm{X}_{ij}} = \frac{\partial\, p_{ij}}{\partial\, \mathrm{X}_{ij}} = p_{ij}(1 - p_{ij})\beta$$

β ضریب متغیر X است.

برای محاسبه اثرات نهایی تغییر در X_{ik} (بر احتمال انتخاب گزینه ***j***)، از p_{ij} نسبت به X_{ik} مشتق می‌گیریم؛

$$\frac{\partial\, P(\mathrm{Y}_i = j)}{\partial\, \mathrm{X}_{ik}} = \frac{\partial\, p_{ij}}{\partial\, \mathrm{X}_{ik}} = -p_{ij}p_{ik}\beta$$

❖ تفسیر نتایج مدل لوجیت شرطی

مثال؛ اثر تغییر در هزینه (قیمت) وسیله‌نقلیه شخصی، بر احتمال استفاده از وسیله‌نقلیه شخصی عبارتست از؛

$$\frac{\partial\, P(\mathrm{Y}_i = 1)}{\partial\, \mathrm{C}_{i1}} = p_{i1}(1 - p_{i1})\beta_1$$

β_1 ضریب متغیر C_{ij} است.

اثر تغییر در هزینه (قیمت) تاکسی، بر احتمال استفاده از وسیله‌نقلیه شخصی عبارتست از؛

$$\frac{\partial\, P(\mathrm{Y}_i = 1)}{\partial \mathrm{C}_{i2}} = -p_{i1}p_{i2}\beta_1$$

❖ تفسیر نتایج مدل لوجیت شرطی

مثال؛ اثر تغییر در زمان سفر وسیله‌نقلیه شخصی، بر احتمال استفاده از وسیله‌نقلیه شخصی عبارتست از:

$$\frac{\partial\, P(\mathrm{Y}_i = 1)}{\partial\, \mathrm{T}_{i1}} = p_{i1}(1 - p_{i1})\beta_2$$

β_2 ضریب متغیر T_{ij} است.

اثر تغییر در زمان سفر تاکسی، بر احتمال استفاده از وسیله‌نقلیه شخصی عبارتست از:

$$\frac{\partial\, P(\mathrm{Y}_i = 1)}{\partial \mathrm{T}_{i2}} = -p_{i1}p_{i2}\beta_2$$

❖ مثال،

مدل مطلوبیت شیوه‌های اصلی حمل‌ونقل در یک شهر به صورت زیر برآورد شده است. مطلوبست؛

الف) برآورد سهم هر شیوه جابجایی با استفاده از داده‌های موجود در جدول (احتمال انتخاب هر شیوه).

ب) اگر هزینه پارکینگ ۱۰هزار ریال برای هر سفر افزایش یابد، سهم هر شیوه حمل‌ونقلی چه تغییری می‌کند؟

$$V_1 = -0.3 - 0.002C_1 - 0.05T_1$$

$$V_2 = -0.35 - 0.002C_2 - 0.05T_2$$

$$V_3 = -0.4 - 0.002C_3 - 0.05T_3$$

C_j : هزینه شیوه حمل‌ونقلی j(برحسب صد ریال)

T_j : زمان سفر با شیوه حمل‌ونقلی j(برحسب دقیقه)

شیوه حمل‌ونقلی	هزینه (هزار ریال)	زمان سفر
خودرو شخصی	۱۳	۲۵
اتوبوس	۷.۵	۳۵
مترو	۹	۴۰

❖ نسبت شانس

نسبت شانس (احتمال نسبی) برابر است با؛

$$OR = \frac{P(\mathrm{Y}_i = j)}{P(\mathrm{Y}_i = k)} = \frac{p_{ij}}{p_{ik}} = \frac{e^{\mathbf{x}'_{ij}\boldsymbol{\beta}}}{e^{\mathbf{x}'_{ik}\boldsymbol{\beta}}} = e^{(\mathbf{x}'_{ij}-\mathbf{x}'_{ik})\boldsymbol{\beta}}$$

با لگاریتم‌گیری، لگاریتم نسبت شانس برابر است با؛

$$\ln OR = (\mathbf{x}'_{ij} - \mathbf{x}'_{ik})\boldsymbol{\beta}$$

برای مثال در مسأله حمل‌ونقل، اگر فقط C_{ij} تغییر کند، آنگاه داریم؛

$$\ln OR = \ln \frac{e^{\alpha_1+\beta_1 C_{i1}+\beta_2 T_{i1}}}{e^{\alpha_2+\beta_1 C_{i2}+\beta_2 T_{i2}}} = (\alpha_1 - \alpha_2) + \beta_1(C_{i1} - C_{i2}) + \beta_2(T_{i1} - T_{i2})$$

❖ نرخ نهایی جانشینی (Marginal Rate of Substitution (MRS))

یکی دیگر از تحلیل‌های حاصل از مدل لوجیت شرطی، محاسبه و تحلیل نرخ نهایی جانشینی است.

نرخ نهایی جانشینی بیانگر تغییر در گزینه‌ها به ازای ثابت ماندنِ مطلوبیت است.

در مثال حمل‌ونقل مدل زیر را درنظر بگیرید،

$$U_{ij} = \alpha_j + \beta_1 C_{ij} + \beta_2 T_{ij} + \varepsilon_{ij}$$

تغییرات مطلوبیت را به ازای تغییر در هزینه حمل‌ونقل و تغییر در زمان صرف شده، حساب کرده و برابر با صفر قرار می‌دهیم؛

$$\Delta U_{ij} = \beta_1 \Delta C_{ij} + \beta_2 \Delta T_{ij} = 0$$

نرخ نهایی جانشینی برابر است با؛

$$MRS = -\frac{\Delta C_{ij}}{\Delta T_{ij}} = \frac{\beta_2}{\beta_1}$$

β_2 بیانگر اثر یک گزینه غیرپولی (زمان صرف شده) است، و β_1 بیانگر اثر یک گزینه پولی (هزینه حمل‌ونقل)

نسبت MRS بیانگر آن است که اگر یک واحد در زمان صرف شده کاسته شود، چقدر به هزینه (تمایل به پرداخت) اضافه خواهد شد (بدون آنکه مطلوبیت تغییر کند)؛ که برابر است با $\frac{\beta_2}{\beta_1}$

❖ نرخ نهایی جانشینی

نرخ نهایی جانشینی برابر است با؛

$$MRS = -\frac{\Delta C_{ij}}{\Delta T_{ij}} = \frac{\beta_2}{\beta_1}$$

β_2 بیانگر اثر یک گزینه غیرپولی (زمان صرف شده) است، و β_1 بیانگر اثر یک گزینه پولی (هزینه حمل‌ونقل)

نسبت MRS بیانگر آن است که اگر یک واحد در زمان صرف شده کاسته شود، چقدر به هزینه (تمایل به پرداخت) اضافه خواهد شد؛ که برابر است با $\frac{\beta_2}{\beta_1}$

رابطه MRS را به صورت روبرو نیز می‌توان نوشت؛

$$\Delta C_{ij} = -\frac{\beta_2}{\beta_1}\Delta T_{ij}$$

این رابطه بیانگر **میزان تمایل به پرداخت هزینه بابت تغییر در زمان صرف شده** است.

مثلا اگر $\frac{\beta_2}{\beta_1} = 1.5$ آنگاه می‌توان گفت به ازای یک واحد کاهش در زمان صرف شده ($\Delta T_{ij} = -1$)، ۵۰٪ افزایش در تمایل به پرداخت هزینه حمل‌ونقل، رخ می‌دهد.

یعنی اگر قیمت ۱۰۰۰ تومان باشد؛ به ازای یک واحد کاهش در زمان، کاربر حاضر است ۱۵۰۰ تومان بپردازد.

❖ ناهمگونی در ترجیحات (Preference heterogeneity)

- تا اینجا فرض این بود که همه افراد با ویژگی‌های مختلف، دارای ترجیحات و تمایل به پرداخت یکسان هستند.
- به همین دلیل ضرایب β برای همه افراد یکسان درنظر گرفته شد.
- این ضرایب می‌توانند متغیر باشند (یعنی میزان اهمیت یک متغیر برای افراد مختلف، متفاوت باشد).
- مثلا، β_1 که ضریب هزینه حمل‌ونقل است می‌تواند با افزایش درآمد، افزایش یابد.
- یعنی با افزایش درآمد، افراد شیوه حمل‌ونقل گران‌تر را انتخاب می‌کنند.
- یا با افزایش درآمد، تمایل دارند شیوه حمل‌ونقلی که زمان سفر کمتری دارد را انتخاب کنند.
- اگر فرض کنیم که افراد با افزایش درآمد به دنبال انتخاب شیوه حمل‌ونقلی با زمان سفر کمتر هستند، می‌توان این موضوع را با وارد کردن حاصلضرب زمانِ صرف شده و درآمد، لحاظ نمود.

$$U_{ij} = \alpha_j + \beta_1 C_{ij} + \beta_2 T_{ij} + \gamma_2 T_{ij} I_{ij} + \cdots$$

❖ ناهمگونی در ترجیحات

$$U_{ij} = \alpha_j + \beta_1 C_{ij} + \beta_2 T_{ij} + \gamma_2 T_{ij} I_{ij} + \cdots$$

I_{ij} درآمد فرد $\boldsymbol{i}$ را نشان می‌دهد که گزینه $\boldsymbol{j}$ را انتخاب می‌کند.

این معادله را به صورت روبرو بازنویسی می‌کنیم.

$$U_{ij} = \alpha_j + \beta_1 C_{ij} + (\beta_2 + \gamma_2 I_{ij}) T_{ij} + \cdots$$

ضریب T_{ij} برابر با $(\beta_2 + \gamma_2 I_{ij})$ است که متغیر می‌باشد، و بستگی به سطح درآمد فرد دارد.

اگر β_1 نیز تابع درآمد باشد، U_{ij} به زیر خواهد بود؛

$$U_{ij} = \alpha_j + +(\beta_1 + \gamma_1 I_{ij})\ C_{ij} + +(\beta_2 + \gamma_2 I_{ij}) T_{ij} + \cdots$$

در این حالت، تمایل به پرداخت برای کاهش زمان صرف شده (نرخ نهایی جانشینی) برابر است با؛

بنابراین، نرخ نهایی جانشینی یا تمایل به پرداخت برای کاهش زمان صرف شده،
بستگی به سطح درآمد فرد $\boldsymbol{i}$ خواهد داشت

$$MRS = \frac{\beta_2 + \gamma_2 I_{ij}}{\beta_1 + \gamma_1 I_{ij}}$$

❖ استقلال گزینه‌های نامربوط ((Independence of Irrelevant Alternatives (IIA)

یکی از الزامات مدل لوجیت شرطی این است که انتخاب از بین مجموعه گزینه‌ها باید دارای شرط **استقلال گزینه‌های نامربوط** باشد.

- فرض **استقلال گزینه‌های نامربوط** بیان می‌کند که؛
  - ✓ انتخاب بین دو گزینه A و B نباید وابسته به ویژگی‌های گزینه نامربوط C باشد.

مثلا فرض کنید افراد می‌توانند بین گزینه **اتوبوس** و **خودرو شخصی** انتخاب کنند.

با معرفی یک شیوه جدید، مثلا **مترو**، اگر فرض استقلال گزینه‌های نامربوط برقرار باشد، نباید تغییری در ترجیح نسبی افراد برای انتخاب بین دو گزینه **اتوبوس** و **خودرو شخصی** ایجاد شود.

- ✓ یعنی اگر یک نفر بین دو گزینه **اتوبوس** و **خودرو شخصی**، گزینه **اتوبوس** را انتخاب کرده بود؛ نباید بعد از معرفی گزینه **مترو**، گزینه **خودرو شخصی** را انتخاب کند!

❖ استقلال گزینه‌های نامربوط (Independence of Irrelevant Alternatives (IIA))

توجه شود که فرض کرده‌ایم که ε_{ij} ها **iid** هستند، یعنی،

- تصمیمات افراد مستقل از هم است (استقلال افراد) و
- گزینه‌ها نیز شبیه به هم نیستند (استقلال گزینه‌ها).

استقلال تصمیمات افراد، منطقی است، اما اگر گزینه‌ها شباهت داشته باشند، فرض استقلال گزینه‌ها نقض می‌شود.

برای دو گزینه j و k، لگاریتم نسبت شانس برابر است با:

$$\ln OR = \ln \frac{P(Y_i = j)}{P(Y_i = k)} = (\mathbf{x}_{ij} - \mathbf{x}_{ik})'\boldsymbol{\beta}$$

مشاهده می‌شود که در صورت برقراری فرض استقلال گزینه‌ها OR تابعی از گزینه‌های j و k است (و تابع سایر گزینه‌ها نیست).

بنابراین، در انتخاب میان دو گزینه j و k، سایر گزینه‌ها نامربوط هستند.

یعنی وقتی می‌خواهیم ترجیحات فرد i را در خصوص این دو گزینه بررسی کنیم، حضور یا عدم حضور سایر گزینه‌ها اهمیتی ندارد.

- به عبارت دیگر، احتمال نسبی دو گزینه، ربطی به ویژگی‌های سایر گزینه‌ها ندارد.
- یا ریسک نسبی دو گزینه، مستقل از سایر گزینه‌هاست.