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 شمارشی

فصل
دوازدهم

Count data Regression Model

مقدمه

مدل رگرسیون پواسون

مدل رگرسیون دوجمله‌ای منفی

مدل رگرسیون شمارشی با صفر زیاد

مدل رگرسیون شمارشی منقطع/سانسور شده

❑ انواع خاصی از متغیرهای گسسته؛

- مقادیر ۰ و ۱ (دو انتخابی)
- مقادیر محدود (مثلا ۴ گروه: ۰، ۱، ۲، ۳)
- موارد بالا را در فصل مدل‌های رگرسیون لوجیت و پروبیت بررسی کردیم.

- متغیرهای تصادفی شمارشی؛
  - نوع عمومی‌تر متغیرهای گسسته شامل مواردی است که مقادیر ۰، ۱، ۲، ... داشته باشد.
  - این نوع متغیرها معمولاً بیانگر فراوانی وقوع یک رویداد (در یک بازه زمانی) هستند.
  - به دلیل گسسته بودن داده‌ها، نمی‌توان از فرض نرمال بودن توزیع استفاده کرد.
  - وقتی $Y=0,1,2,\ldots$ باشد، باید از توزیع‌های گسسته استفاده کرد.
  - بنابراین، استفاده از روش OLS مناسب نیست! چرا؟

مقدمه

❑ متغیرهای تصادفی شمارشی؛

زیرا فروض اصلی روش OLS، مانند نرمال بودن و همسانی واریانس، نقض می‌شوند و نیز ممکن است مقادیر پیش‌بینی‌شده منفی به دست آید.

- **بنابراین، استفاده از روش OLS مناسب نیست! چرا؟**
- **در این شرایط، برای توصیف Y از توزیع‌های احتمال گسسته مانند دوجمله‌ای، پواسون، دوجمله‌ای منفی، هندسی و غیره استفاده می‌شود.**
- **دو توزیع پواسون، دوجمله‌ای منفی پرکاربردتر هستند.**

- **توزیع پواسون** معمولاً برای مدل‌سازی تعداد وقوع یک رویداد در یک بازه زمانی یا مکانی مشخص به کار می‌رود. این توزیع به‌ویژه زمانی مناسب است که وقوع رویدادها مستقل بوده و نرخ متوسط وقوع آنها ثابت باشد.
- همچنین، توزیع پواسون برای توصیف حوادث نادر (احتمال موفقیت (وقوع رویداد موردنظر) نسبتاً کم باشد) نیز به کار می‌رود.

مقدمه

- در توزیع پواسون میانگین و واریانس با هم برابر است.
- در بسیاری موارد از دنیای واقعی، این ویژگی نقض می‌شود.
- در این موارد از سایر توزیع‌ها (به خصوص دوجمله‌ای منفی) استفاده می‌کنیم.
- در برخی دیگر از موارد که در بین مشاهدات نمونه، **تعداد صفرها بسیار زیاد است**، از مدل‌های صفر-انباشته بهره می‌بریم.

مدل رگرسیون داده‌های شمارشی

فصل
دوازدهم

Count data Regression Model

مقدمه

مدل رگرسیون پواسون

مدل رگرسیون دوجمله‌ای منفی

مدل رگرسیون شمارشی با صفر زیاد

مدل رگرسیون شمارشی منقطع/سانسور شده

مدل پواسون

مدرس: محمدمهدی بشارتی

❖ **فرض کنید؛**

- رویدادهای موردنظر به صورت تصادفی در طول زمان رخ می‌دهند؛
- و متوسط تعداد وقوع آنها طی فاصله زمانی موردنظر (مثلا یک ماه)، معلوم است (λ).

مدل پواسون

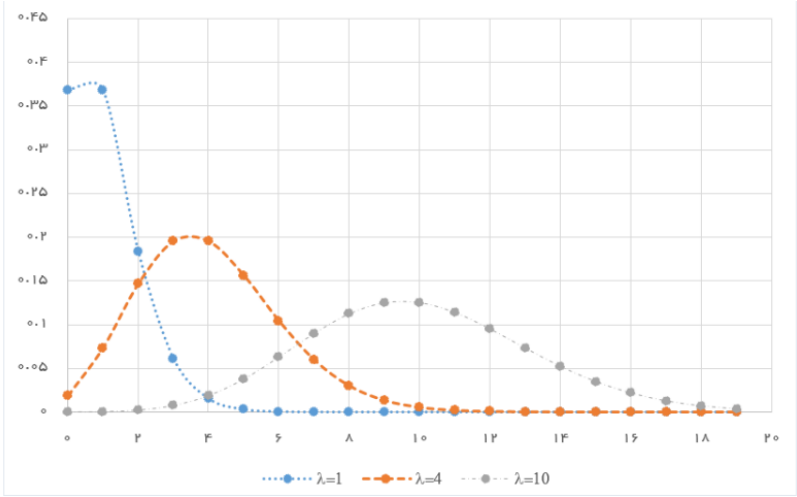
❖ یادآوری؛

- هنگامی که اندازه نمونه افزایش یابد، و احتمال وقوع رویداد موردنظر کم باشد، آنگاه، توزیع پواسون تقریب مناسبی برای توزیع دوجمله‌ای است.
  - ◁ ((کم بودن احتمال وقوع به معنای نادر بودنِ رویداد موردنظر است))
- بین داده‌های شمارشی و دیرش_Duration_ (مدت زمان انتظار بین وقوع دو رویداد) نیز رابطه زیر وجود دارد؛
  - ◁ برای هر $t > 0$ فراوانی وقوع در فاصله زمانی $[0,t]$ از توزیع پواسون با میانگین λt پیروی می‌کند.

- برای توزیع پواسون فرض بر این است که مقدار مشاهده شده Y_i در حالت i مستقل از مقدار مشاهده شده در حالت j است.
- همچنین، مهمترین ویژگی توزیع پواسون آن است که میانگین و واریانس با هم برابر است.

مدل پواسون– توزیع احتمال پواسون

مدرس: محمدمهدی بشارتی



- **تابع جرم احتمال توزیع پواسون،**

$$P(Y_i|\mathbf{x}_i) = \frac{e^{-\lambda_i}\ {\lambda_i}^{Y_i}}{Y_i\ !} \qquad ;\ Y_i = 0,1,2,\ldots$$

- **توزیع پواسون کاملا وابسته به پارامتر λ_i است.**
- **به‌گونه‌ای که به ازای λ_i های کوچک، دارای چولگی شدید است؛ ولی با بزرگ‌شدن λ_i به یک توزیع قرینه تبدیل می‌شود.**
- **در توزیع پواسون امیدریاضی و واریانس با هم برابر هستند؛** $E(Y_i|\mathbf{x}_i) = var(Y_i|\mathbf{x}_i) = \lambda_i$
- **به عبارت دقیق‌تر: در مدل پواسون، میانگین شرطی و واریانس شرطی متغیر پاسخ برابر هستند.**
- **بنابراین، هر عاملی که موجب تغییر در میانگین شرطی شود، واریانس شرطی را نیز به همان نسبت تغییر می‌دهد.**
- **مثلا اگر با افزایش X مقدار Y افزیش می‌یابد، میانگین و واریانس آن نیز باید افزایش یابند.**

مدل پواسون– رگرسیون پواسون

- امیدریاضی Y_i برابر با λ_i است که تحت تاثیر عوامل مختلفی قرار دارد $(E(Y_i|\mathbf{x}_i) = \lambda_i)$.
- بنابراین، می‌توان برای λ_i یک معادله رگرسیون به صورت $E(Y_i|\mathbf{x}_i)$ تعریف نمود.
- چون λ_i غیرمنفی است $(\lambda_i > 0)$، و Y_i ها نیز غیرمنفی هستند، بنابراین، آن را به صورت نمایی تعریف می‌کنند؛

$$E(Y_i|\mathbf{x}_i) = \lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}} = e^{\beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}}$$

- **نکته:** استفاده از تابع نمایی تضمین می‌کند که مقدار λ_i همواره مثبت باشد.
- لگاریتم این معادله عبارتست از؛

$$\ln \lambda_i = \mathbf{x}_i'\boldsymbol{\beta} = \beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}$$

- تابع درستنمایی مدل پواسون:

$$Y_i \sim Poisson\ (\lambda_i) \quad \Rightarrow \quad L = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\ {\lambda_i}^{Y_i}}{Y_i\ !}$$

مدل پواسون– تفسیر نتایج مدل پواسون

- اثر نهایی X_k بر $\mathbf{Y}$ در مدل پواسون برابر است با؛

$$\ln E(Y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad \Rightarrow \quad \frac{\partial \ln E(Y|\mathbf{x}_i)}{\partial\, \mathrm{X}_k} = \beta_k$$

- بنابراین، رابطه تقریبی روبرو برقرار است؛

$$\frac{\Delta E(Y|\mathbf{x}_i)}{E(Y|\mathbf{x}_i)} \cong \beta_k\, \Delta X_k$$

- به این معنا که اگر X_k یک واحد تغییر کند، آنگاه امیدریاضی شرطی $\mathbf{Y}$ $E(Y|\mathbf{x}_i)$ به اندازه $100\beta_k$ درصد تغییر خواهد کرد (اگر مقدار β_k کوچک باشد.).
- مثلا اگر $\beta_k = 0.03$ باشد، آنگاه به ازای یک واحد تغییر در X_k، امیدریاضی $\mathbf{Y}$ در حدود ۳٪ تغییر خواهد کرد.
- تفسیر دقیق‌تر بر اساس رابطه $e^{\beta_k} - 1$ انجام می‌شود (اسلاید بعد).

مدل پواسون– تفسیر نتایج مدل پواسون

- **ریسک نسبی (نرخ وقوع رویداد Incident Rate Ratio (IRR))**

- وقتی که با داده‌های شمارشی کار می‌کنیم، به ریسک نسبی «**نسبت نرخ وقوع (Incident Rate Ratio (IRR))**» می‌گویند.

$$\text{IRR}=e^{\beta_k}$$

- یعنی یک واحد افزایش در مقدار متغیر X_k می‌تواند **ریسک نسبی وقوع رخداد موردنظر** را به اندازه e^{β_k} افزایشدهد (در شرایط ثابت ماندن سایر متغیرها)
- یعنی به ازای یک واحد افزایش در مقدار متغیر X_k ، مقدار موردانتظار متغیر پاسخ، e^{β_k} برابر می‌شود.
- مثلا اگر $e^{\beta_k} = 1.20$ باشد، نرخ مورد انتظار وقوع رویداد ۲۰٪ افزایش می‌یابد.

مدل پواسون– تفسیر نتایج مدل پواسون

- **متوسط اثر حاشیه‌ای (Average marginal effect)**

- با استفاده از داده‌های مشاهده شده X_k و مقادیر برآورد شده برای پارامترهای مدل، می‌توان متوسط اثرات حاشیه‌ای را برای هر متغیر X_k به دست آورد:

$$\frac{\partial E(Y|\mathbf{x}_i)}{\partial \, \mathrm{X}_k} = e^{\mathbf{x}_i'\boldsymbol{\beta}}\beta_k$$

- **شکل ساده‌تر:**

$$\frac{\partial E(Y|\mathbf{x}_i)}{\partial \, \mathrm{X}_k} = \lambda_i \beta_k$$

مدل پواسون– تفسیر نتایج مدل پواسون

مدرس: محمدمهدی بشارتی

○ مثال

فرض کنید هدف پژوهش، مدل‌سازی **تعداد تصادفات ماهانه در تقاطع‌های شهری** باشد.

متغیر پاسخ (Y_i): "تعداد تصادفات در تقاطع i طی یک ماه"

چون: تعداد تصادفات یک متغیر شمارشی است،
مقادیر آن ۰،۱،۲،... می‌باشد،
و معمولاً توزیع نرمال ندارد،
می‌توان از مدل رگرسیون پواسون استفاده کرد.

فرض کنید مدل زیر برآورد شده است:

$$\ln \lambda_i = -1.20 + 0.015\ \mathrm{Traffic_i} + 0.4\mathrm{Signal_i} - 0.3\mathrm{Camera_i}$$

که در آن:

$$\lambda_i = E(Y_i|\mathbf{x}_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}} = e^{\beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}}$$

یعنی λ_i= تعداد مورد انتظار تصادفات در ماه.

Traffici: حجم ترافیک روزانه (هزار وسیله نقلیه)
$Signal_i$: وجود چراغ راهنمایی (۱=دارد، ۰=ندارد)
$Camera_i$: وجود دوربین کنترل سرعت (۱=دارد، ۰=ندارد)

مدل پواسون– تفسیر نتایج مدل پواسون

مدرس: محمدمهدی بشارتی

- **مثال**

تفسیر ضرایب:

۱. متغیر حجم ترافیک

$$\beta_i = 0.015$$

تفسیر تقریبی درصدی

با افزایش یک واحد در حجم ترافیک (یعنی ۱۰۰۰ وسیله نقلیه بیشتر در روز):

$$100 \times 0.015 = 1.5\%$$

تعداد مورد انتظار تصادفات تقریباً ۱.۵٪ افزایش می‌یابد.

تفسیر دقیق با IRR

$$IRR = e^{0.015} = 1.015$$

یعنی:

هر ۱۰۰۰ وسیله نقلیه اضافه، نرخ مورد انتظار تصادفات را حدود ۱/۵٪ افزایش می‌دهد.

مدل پواسون– تفسیر نتایج مدل پواسون

مدرس: محمدمهدی بشارتی

- **مثال**

تفسیر ضرایب:

۲. متغیر چراغ راهنمایی

$$\beta_i = 0.4$$

چون متغیر مجازی (Dummy) است:

$$IRR = e^{0.4} = 1.49$$

تفسیر:

تقاطع‌هایی که چراغ راهنمایی دارند، به‌طور متوسط حدود ۴۹٪ تصادف بیشتری نسبت به تقاطع‌های بدون چراغ دارند (با ثابت بودن سایر عوامل).

سوال: «پس چراغ راهنمایی خطرناک است!» ؟

مدل پواسون– تفسیر نتایج مدل پواسون

مدرس: محمدمهدی بشارتی

- **نکته:**

✓ در مدل پواسون، ضرایب مستقیماً اثر خطی بر مقدار Y ندارند، بلکه اثر آنها بر **لگاریتم** میانگین شرطی متغیر پاسخ است.

✓ به همین دلیل معمولاً ضرایب با استفاده از IRR یا تغییرات درصدی تفسیر می‌شوند.

مدل پواسون– محدودیت‌های مدل پواسون

مدرس: محمدمهدی بشارتی

- در توزیع پواسون امیدریاضی و واریانس با هم برابر است (این ویژگی با نام «پراکندگی یکسان» شناخته می‌شود)
- این ویژگی یک محدودیت برای مدل پواسون است.
- در پدیده‌های دنیای واقعی، در اکثر موارد واریانس بزرگتر از امیدریاضی است. $Var(Y_i|\mathbf{x}_i) > E(Y_i|\mathbf{x}_i)$
- این پدیده با نام «بیش‌پراکندگی –Over-dispersion» شناخته می‌شود.
- در شرایطی که بیش‌پراکندگی در میان مشاهدات وجود دارد، برآوردهای مدل پواسون دارای انحراف معیار با تورش منفی بوده و ناکارا خواهند بود.
- بیش‌پراکندگی موجب می‌شود که انحراف معیار پارامترهای برآورد شده معمولا کمتر از مقدار واقعی برآورد شوند؛ و بنابراین، آزمون معناداری پارامترها را مخدوش می‌کند (آماره‌های t و z بیش از حد بزرگ شده و ممکن است متغیرها به‌اشتباه معنادار تشخیص داده شوند.).

مدل پواسون– محدودیت‌های مدل پواسون

❖ آزمون‌های بررسی بیش‌پراکندگی؛

- آزمون‌های مختلفی برای آزمودن بیش‌پراکندگی وجود دارد.
- یکی از آزمون‌های بررسی بیش‌پراکندگی به صورت گام‌های زیر است؛

1. مدل پواسون را برآورد کرده و مقادیر برآوردشده برای متغیر وابسته ($\hat{Y}_i$) را حساب می‌کنیم.
2. مقادیر خطاها (e_i) را حساب می‌کنیم (اختلاف $\hat{Y}_i$ و Y_i)
3. عبارت $e_i^2 - Y_i$ را حساب می‌کنیم.
4. مقادیر $e_i^2 - Y_i$ را بر روی $\hat{Y}_i$ برازش می‌کنیم.
5. اگر ضریب $\hat{Y}_i$ معنادار باشد، فرض پراکندگی یکسان رد می‌شود و بنابراین استفاده از مدل رگرسیون پواسون اشتباه است.

((اگر ضریب $\hat{Y}_i$ مثبت باشد، به معنای وجود بیش‌پراکندگی و اگر منفی باشد، به معنای وجود کم‌پراکندگی است.))

مدل پواسون– محدودیت‌های مدل پواسون

❖ آزمون‌های بررسی بیش‌پراکندگی؛

○ آزمون پیرسون برای بررسی بیش‌پراکندگی؛

- فرض کنید واریانس متناسب با میانگین باشد (رابطه روبرو): $var(Y) = \Phi E(Y) = \Phi\lambda$
- اگر $\Phi = 1$ باشد، برابری واریانس و میانگین برقرار است؛ بنابراین می‌توان از مدل پواسون استفاده کرد.
- اگر $\Phi > 1$ باشد، بیش‌پراکندگی وجود دارد.
- اگر $\Phi < 1$ باشد، کم‌پراکندگی وجود دارد (به ندرت رخ می‌دهد).
- برای آزمون برابری واریانس و میانگین از آماره کای‌دو پیرسون استفاده می‌کنیم؛

$$\chi^2_{n-K} = \sum_{i=1}^{n} \frac{(Y_i - \lambda_i)^2}{\lambda_i}$$

- از آنجا که امیدریاضی χ^2 برابر با درجه آزادی آن است، بنابراین، $E(\chi^2_{n-K}) = n - K$
- به عبارت دیگر، مقدار انتظاری برای χ^2_{n-K} برابر است با $n - K$

مدل پواسون– محدودیت‌های مدل پواسون

❖ آزمون‌های بررسی بیش‌پراکندگی؛

○ آزمون پیرسون برای بررسی بیش‌پراکندگی؛

- مقدار انتظاری برای χ^2_{n-K} برابر است با $n-K$
- بنابراین، اگر مدل موردنظر برای برآورد λ_i به درستی تصریح شده باشد، باید مقدار آماره پیرسون برابر با $n-K$ باشد.
- بنابراین، انتظار داریم که نسبت $\frac{\chi^2_{n-K}}{n-K}$ تقریبا برابر با ۱ باشد.
- اما اگر Y_i دچار بیش‌پراکندگی باشد، دراینصورت این نسبت بزرگتر از ۱ خواهد بود.
- به عبارت دیگر، $\frac{\chi^2_{n-K}}{n-K}$ برآوردی از Φ است؛ $\widehat{\Phi} = \frac{\chi^2_{n-K}}{n-K}$
- بر این مبنا، اگر مقدار آماره پیرسون بزرگتر از $\chi^2_{\alpha\,,\,n-K}$ باشد، بیش‌پراکندگی رد نمی‌شود.

مدل پواسون– محدودیت‌های مدل پواسون

مدرس: محمدمهدی بشارتی

- اگر بیش‌پراکندگی رد نشود، بدین معناست که نمی‌توان از مدل پواسون استفاده کرد.
- در این شرایط نیاز به تعدیل یا تغییر مدل داریم
- **تعدیل مدل (modify):** لحاظ کردن یک ضریب برای درنظر گرفتن بیش‌پراکندگی در مدل پواسون (منجر به مدل‌های شبه‌پواسون (Quasi Poisson) و پواسون تعمیم‌یافته (Generalized Poisson) می‌شود.)
- **تغییر مدل:** جایگزینی مدل پواسون با سایر مدل‌ها مانند دوجمله‌ای منفی.

مدل پواسون– مدل شبه‌پواسون (Quasi-Poisson Regression)

مدرس: محمدمهدی بشارتی

- یکی از روش‌های لحاظ نمودن بیش‌پراکندگی در مدل پواسون است.
- در این مدل فرض می‌شود که واریانس، تابع خطی از میانگین است؛ $var(Y) = \Phi E(Y)$
- در مدل شبه‌پواسون، پارامترها مانند مدل پواسون برآورد می‌شوند.
- اما ضریب Φ را برای تعدیل واریانس برآورد می‌کنیم؛ و معنادار بودن پارامترها نیز براساس واریانس تعدیل‌شده،آزموده می‌شود.
- بنابراین، پارامترهای مدل‌ها یکسان خواهد بود، اما آماره آزمون معناداری آن‌ها متفاوت است.

انحراف معیار $\hat{\boldsymbol{\beta}}$ در مدل پواسون: $\boldsymbol{SE_p(\hat{\beta})}$

انحراف معیار $\hat{\boldsymbol{\beta}}$ در مدل شبه‌پواسون: $\boldsymbol{SE_{Qp}(\hat{\beta}) = \sqrt{\hat{\Phi}} SE_p(\hat{\beta})}$

مدل پواسون– مدل پواسون تعمیم‌یافته

مدرس: محمدمهدی بشارتی

- مدل‌های پواسون تعمیم‌یافته انواع مختلفی داشته و به روش‌های مختلف و با فرمول‌بندی‌های متفاوت ارائه شده‌اند.
- مدل رگرسیون پواسون معمولی، یک پارامتر λ دارد که میانگین و واریانس را توصیف می‌کند.
- اما در مدل رگرسیون پواسون تعمیم‌یافته نیاز به ضریب دیگری داریم تا پدیده بیش‌پراکندگی را توصیف کند
- این ضریب را معمولاً با α نمایش داده و به آن ضریب پراکندگی (Dispersion Parameter) می‌گویند.
- هر یک از انواع مدل‌های پواسون تعمیم‌یافته، علاوه بر λ_i ، پارامتر دیگری نیز برای توصیف پراکندگی درنظر می‌گیرند.

مدل پواسون– مدل پواسون تعمیم‌یافته

- یکی از انواع **مدل‌های رگرسیون پواسون تعمیم‌یافته (موسوم به GP-1)**
- در این مدل فرض می‌شود که متغیر وابسته y یک متغیر تصادفی با توزیع احتمال زیر است؛

$$P(Y_i|\mathbf{x}_i) = \frac{e^{-(\lambda_i+\alpha Y_i)}(\lambda_i+\alpha Y_i)^{Y_i}}{Y_i\,!} \qquad ; Y_i = 0,1,2,\ldots$$

امیدریاضی توزیع: $E(Y_i|\mathbf{x}_i) = \frac{\lambda_i}{1-\alpha}$

واریانس توزیع: $Var(Y_i|\mathbf{x}_i) = \frac{\lambda_i}{(1-\alpha)^3}$

سوال: درصورتیکه در این مدل، $\alpha = 0$ قرار داده شود، چه اتفاقی می‌افتد؟

مدل پواسون– مدل پواسون تعمیم‌یافته

مدرس: محمدمهدی بشارتی

- یکی از انواع **مدل‌های رگرسیون پواسون تعمیم‌یافته (موسوم به GP-2)**
- در این مدل فرض می‌شود که متغیر وابسته y یک متغیر تصادفی با توزیع احتمال زیر است؛

$$P(Y_i|\mathbf{x}_i) = \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)\frac{e^{-\left(\frac{\lambda_i(1+\alpha Y_i)}{1+\alpha\lambda_i}\right)}(\lambda_i+\alpha Y_i)^{Y_i-1}}{Y_i\,!}$$

امیدریاضی توزیع: $E(Y_i|\mathbf{x}_i) = \lambda_i$

واریانس توزیع: $Var(Y_i|\mathbf{x}_i) = \lambda_i(1+\alpha\lambda_i)^2$

سوال: درصورتیکه در این مدل، $\alpha = 0$ قرار داده شود، چه اتفاقی می‌افتد؟

مدل پواسون– مدل پواسون تعمیم‌یافته

مدرس: محمدمهدی بشارتی

❖ **رابطه محاسبه ضریب پراکندگی:**

$$\alpha = \frac{\sum_{i=1}^{N}\left(\frac{|y_i - \hat{y}_i|}{\sqrt{\hat{y}_i}} - 1\right) * (\hat{y}_i)^{(1-p)}}{N - k - 1}$$

N= اندازه نمونه

k= تعداد متغیرهای مدل

y_i : i-امین مقدار مشاهده شده

$\hat{y}_i$: λ_i پیش‌بینی شده توسط مدل برای i-امین مشاهده

p: برابر با ۱ برای مدل GP-1 و برابر با ۲ برای مدل GP-2

❖ **برآورد مدل رگرسیون پواسون تعمیم‌یافته همانند رگرسیون پواسون است؛ که در آن، برای امیدریاضی (λ_i) یک معادله به صورت $\lambda_i = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ و یک ضریب برای تعدیل واریانس درنظر گرفته می‌شود.**

❖ **سپس با روش MLE برآورد پارامترهای مدل انجام می‌شود.**

مدل رگرسیون داده‌های شمارشی

فصل
دوازدهم

Count data Regression Model

مقدمه

مدل رگرسیون پواسون

مدل رگرسیون دوجمله‌ای منفی

مدل رگرسیون شمارشی با صفر زیاد

مدل رگرسیون شمارشی منقطع/سانسور شده

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

- **در مدل‌های خطی تعمیم‌یافته (GLM)، توزیع متغیر وابسته تعیین‌کننده شکل تابع درستنمایی مدل است. به عنوان مثال:**
- **مدل رگرسیون پواسون بر مبنای توزیع احتمال پواسون** تعریف می‌شود.
- **مدل رگرسیون لوجیت بر مبنای توزیع احتمال لوجستیک (برای خطای u)** تعریف می‌شود.

(((البته، در مدل لوجیت، فرض می‌شود جمله خطای مدل نهفته از توزیع لوجستیک پیروی می‌کند که منجر به تابع احتمال لوجستیک می‌شود.)))

- **مدل رگرسیون خطی کلاسیک بر مبنای توزیع احتمال نرمال** تعریف می‌شود.
- **مدل رگرسیون دوجمله‌ای منفی بر مبنای یک تابع توزیع احتمال خاص** تعریف می‌شود.
- **یکی از راه‌های استخراج توزیع دوجمله‌ای منفی که به اختصار NB2 نامیده می‌شود، از توزیع ترکیبی پواسون-گاما به دست می‌آید (که با نام Poisson-Gamma mixture نیز شناخته می‌شود).**

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

- **یکی از راه‌های استخراج توزیع دوجمله‌ای منفی که به اختصار NB2 نامیده می‌شود، از توزیع ترکیبی پواسون-گاما به دست می‌آید (که با نام Poisson-Gamma mixture نیز شناخته می‌شود).**

- در مدل پواسون فرض می‌شود (λ_i ثابت است): $E(Y_i|\mathbf{x}_i) = var(Y_i|\mathbf{x}_i) = \lambda_i$

اما در داده‌های واقعی:

✓ نرخ وقوع رویدادها ممکن است بین افراد/مشاهدات متفاوت باشد،

✓ یعنی ناهمگنی مشاهده‌نشده وجود داشته باشد.

اگر فرض کنیم:

λ_i خود یک متغیر تصادفی با توزیع گاما باشد، آنگاه توزیع نهایی Y_i به توزیع دوجمله‌ای منفی تبدیل می‌شود.

این دقیقاً ایده Poisson-Gamma mixture است. یعنی Y_i متغیر تصادفی دارای توزیع پواسونی است که λ_i آن از توزیع گاما پیروی می‌کند. آنگاه توزیع نهایی Y_i به توزیع دوجمله‌ای منفی تبدیل می‌شود.

- **این ترکیب، تنها یکی از راه‌های استخراج توزیع دوجمله‌ای منفی است.**
- **ویژگی مهم این روش آن است که به ما امکان مدل‌سازی ناهمگنی مشاهده‌نشده** (Unobserved Heterogeneity) **را در مدل پواسون می‌دهد.**

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ **تابع احتمال دوجمله‌ای منفی،**

○ **تابع احتمال توزیع دوجمله‌ای،**

$$P(x,k,p) = \frac{(x-1)!}{(x-k)!\ (k-1)!} p^k (1-p)^{x-k}$$

$$x = k, k+1, k+2, \dots$$

X: تعداد کل آزمون‌ها (مشاهدات) تا وقوع k-امین موفقیت باشد،

امیدریاضی X
$$E(X) = \frac{k}{p}$$

واریانس X
$$Var(X) = \frac{k(1-p)}{p^2}$$

این تابع احتمال بیانگر احتمال آن است که k-اُمین موفقیت در x-اُمین آزمایش رخ دهد.

مثلا: تعداد روزهایی که یک دستگاه کار می‌کند تا برای سومین مرتبه دچار مشکل شود از توزیع دوجمله‌ای منفی پیروی می‌کند.

توجه مهم: اگرچه توزیع دوجمله‌ای منفی در نظریه احتمال معمولاً به‌صورت «تعداد آزمایش‌ها تا وقوع k-اُمین موفقیت» تعریف می‌شود، اما در اقتصادسنجی از فرم بازپارامتردهی‌شده آن برای مدل‌سازی داده‌های شمارشی دارای بیش‌پراکندگی استفاده می‌شود (پارامترهای μ,α).

$$E(Y_i|\mathbf{x}_i) = \mu_i \qquad var(Y_i|\mathbf{x}_i) = \mu_i + \alpha\mu_i{}^2$$

مدل دوجمله‌ای منفی

❖ تابع احتمال دوجمله‌ای منفی،

○ اگر از تبدیل‌های $k=r$ و $x-k=y$ استفاده کنیم، شکل دیگری از این توزیع را خواهیم داشت؛

$$P(y,r,p)=\frac{(y+r-1)!}{y!\ (r-1)!}(p)^r(1-p)^y \ ;y=0,1,2,\dots\ ;\ \mathrm{k}\ ,\ \mathrm{r}>0$$

امیدریاضی Y

$$E(Y)=\frac{(1-p)r}{p}$$

واریانس Y

$$Var(Y)=\frac{(1-p)\,r}{p^2}$$

مدل دوجمله‌ای منفی

❖ تابع احتمال دوجمله‌ای منفی،

- اگر از تبدیل $\mu = \frac{(1-p)\,r}{p}$ استفاده کنیم، شکل دیگری از این توزیع را خواهیم داشت؛

$$P(y,\mu,r) = \frac{(y+r-1)!}{y!\ (r-1)!}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^y \quad ;y = 0,1,2,\dots$$

امیدریاضی Y $\qquad E(Y) = \mu$

واریانس Y $\qquad Var(Y) = \mu + \frac{\mu^2}{r}$

- اگر r افزایش یابد، آنگاه واریانس به امیدریاضی نزدیک می‌شود (مشابه توزیع پواسون خواهد شد): $E(Y) \cong Var(Y)$
- اما برای مقادیر کوچک r، $E(Y) < Var(Y)$ است و بیش‌پراکندگی وجود دارد (نمی‌توان از توزیع پواسون استفاده کرد).

مدل دوجمله‌ای منفی

❖ **تابع احتمال دوجمله‌ای منفی،**

○ اگر از تبدیل $\alpha = \frac{1}{r}$ و $\Gamma(n+1) = n!$ استفاده کنیم، شکل دیگری از این توزیع را خواهیم داشت؛

$$P(y,\mu,\alpha) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\,\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu}\right)^{(\frac{1}{\alpha})}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \quad ; \quad y = 0,1,2,..$$

امیدریاضی Y

$$E(Y) = \mu$$

واریانس Y

$$Var(Y) = \mu + \alpha\mu^2$$

- α **ضریب بیش‌پراکندگی** است. به ازای $\alpha = 0$ **پراکندگی یکسان**، به ازای $\alpha > 0$ **بیش‌پراکندگی** و به ازای $\alpha < 0$ **کم‌پراکندگی** وجود دارد. اما در فرم استاندارد NB2 معمولا $\alpha > 0$ فرض می‌شود (تا **بیش‌پراکندگی** مدل شود). زیرا در غیر این صورت؛ برای بعضی مقادیر منفی، تابع احتمال معتبر نمی‌ماند.

مدل دوجمله‌ای منفی

- **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**
- **می‌خواهیم مدل دوجمله‌ای منفی را برحسب ترکیب پواسون-گاما به دست آوریم.**
- **فرض کنید Y_i دارای توزیع پواسون با پارامتر λ باشد؛**

$$Y_i \sim P(Y_i|\lambda) = \frac{e^{-\lambda}\lambda^{Y_i}}{Y_i\,!} \qquad ; Y_i = 0,1,2,\dots$$

- **همچنین، فرض کنید λ یک متغیر تصادفی است که از توزیع گاما با پارامترهای δ و θ تبعیت می‌کند؛**

$$\lambda \sim Gamma(\theta,\delta) = \frac{\delta^{\theta}}{\Gamma(\theta)}\lambda^{\theta-1}\,e^{-\delta\lambda} \qquad ; \lambda > 0$$

$$E(\lambda) = \frac{\theta}{\delta} \qquad Var(\lambda) = \frac{\theta}{\delta^2}$$

- **از طرف دیگر، برای λ رابطه $\lambda = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ را داریم که همان معادله رگرسیون موردنظر است.**

مدل دوجمله‌ای منفی

- **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**
- **تابع توزیع مشترک λ و Y عبارتست از؛**

$$f(Y,\lambda)= P(Y|\lambda)P(\lambda) = \left(\frac{e^{-\lambda}\lambda^{Y}}{Y!}\right)\left(\frac{\delta^{\theta}}{\Gamma(\theta)}\lambda^{\theta-1}\,e^{-\delta\lambda}\right)$$

- **اکنون تابع احتمال نهایی دوجمله‌ای منفی (توزیع حاشیه‌ای) Y با معلوم بودن x_i به صورت زیر به دست می‌آید؛**

$$f(Y|\mathbf{x}_i) = \int_0^{\infty}\left(\frac{e^{-\lambda}\lambda^{Y}}{Y!}\right)\left(\frac{\delta^{\theta}}{\Gamma(\theta)}\lambda^{\theta-1}\,e^{-\delta\lambda}\right)d\lambda$$

- **با اندکی تبدیل و مرتب‌سازی، این انتگرال محاسبه گردیده و مدل دوجمله‌ای منفی با ترکیب پواسون–گاما به دست می‌آید.**

$$f(Y|\mathbf{x}_i) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\ \left(\frac{\delta}{1+\delta}\right)^{\theta}\ \left(\frac{1}{1+\delta}\right)^{y}$$

مدل دوجمله‌ای منفی

❖ **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**

$$f(Y|\mathbf{x}_i) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\delta}{1+\delta}\right)^{\theta} \left(\frac{1}{1+\delta}\right)^{y}$$

❖ **اگر در تابع بالا، از تبدیل $\theta = r$ و $\delta = \frac{r}{\mu}$ استفاده کنیم، معادله زیر به دست می‌آید**

$$P(y,\mu,r) = \frac{(y+r-1)!}{y!\ (r-1)!} \left(\frac{r}{r+\mu}\right)^{r} \left(\frac{\mu}{r+\mu}\right)^{y} \quad ; y = 0,1,2,\ldots$$

❖ **اگر از تبدیل $\theta = \frac{1}{\alpha}$ و $\delta = \frac{1}{\alpha\mu}$ استفاده کنیم، معادله زیر به دست می‌آید**

$$P(y,\mu,\alpha) = \frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(y+1)\ \Gamma(\frac{1}{\alpha})} \left(\frac{1}{1+\alpha\mu}\right)^{(\frac{1}{\alpha})} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \quad ;\ y = 0,1,2,..$$

مدل دوجمله‌ای منفی

- **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**
- **شکل عمومی مدل دوجمله‌ای منفی با ترکیب پواسون-گاما، با در نظر گرفتن تبدیل $\theta = \alpha^{-1}\mu^{1-k}$ و $\delta = \alpha^{-1}\mu^{-k}$ به صورت زیر خواهد بود؛**

$$P(Y|\mathbf{x}_i) = \frac{\Gamma(y+\alpha^{-1}\mu^{1-k})}{\Gamma(y+1)\Gamma(\alpha^{-1}\mu^{1-k})} \left(\frac{1}{1+\alpha\mu^{k}}\right)^{\alpha^{-1}\mu^{1-k}} \left(\frac{\alpha\mu^{k}}{1+\alpha\mu^{k}}\right)^{y}$$

امیدریاضی Y $E(Y) = \mu$

واریانس Y $Var(Y) = \mu(1 + \alpha\mu^{k})$

این مدل به ازای مقادیر مختلف k نام‌گذاری شده است. به طور خاص، دو مدل NB1 و NB2

مدل دوجمله‌ای منفی

❖ **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**

❖ مدل NB1 که به ازای k=0 به دست می‌آید؛

امیدریاضی Y $E(Y) = \mu$

واریانس Y $Var(Y) = \mu(1 + \alpha)$

در این مدل، واریانس ضریب ثابتی از میانگین است. بنابراین، رابطه واریانس با میانگین، خطی است که مشابه مدل شبه‌پواسون است.

❖ مدل NB2 که به ازای k=1 به دست می‌آید؛

$$P(Y|\mathbf{x}_i) = \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y+1)\,\Gamma(\frac{1}{\alpha})}(\frac{1}{1+\alpha\mu})^{(\frac{1}{\alpha})}(\frac{\alpha\mu}{1+\alpha\mu})^{y}$$

امیدریاضی Y $E(Y) = \mu$

واریانس Y $Var(Y) = \mu + \alpha\mu^2$

در این مدل، رابطه واریانس با میانگین غیرخطی (درجه ۲) است.

در ادامه از این مدل استفاده می‌کنیم.

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ **تابع احتمال دوجمله‌ای منفی به عنوان ترکیب پواسون–گاما،**

مدل	فرض واریانس
Poisson	$Var = \mu$
Quasi-Poisson	$Var = \phi\mu$
NB1	$Var = \mu(1+\alpha)$
NB2	$Var = \mu + \alpha\mu^2$

- در NB2، با افزایش میانگین، واریانس سریع‌تر رشد می‌کند؛ بنابراین برای داده‌هایی با بیش‌پراکندگی شدید مناسب‌تر است.
- سوال: اگر در مدل دوجمله‌ای منفی $\alpha = 0$ شود، چه اتفاقی می‌افتد؟

مدل دوجمله‌ای منفی

❖ **برآورد مدل رگرسیون دوجمله‌ای منفی،**

❖ مشاهدات $Y_i = 0,1,2, \dots$ را در نظر بگیرید. فرض کنید که Y_i دارای توزیع دوجمله‌ای منفی است (فرم NB2)؛

$$f(Y_i|\mathbf{x}_i) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1)\,\Gamma(\frac{1}{\alpha})}(\frac{1}{1+\alpha\mu_i})^{(\frac{1}{\alpha})}(\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i} \quad ; Y_i = 0,1,2,..$$

همچنین معادله میانگین شرطی به صورت زیر خواهد بود؛

$$E(Y_i|\mathbf{x}_i) = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

در مدل رگرسیون دوجمله‌ای منفی، میانگین شرطی μ_i به‌صورت تابعی از متغیرهای توضیحی مدل‌سازی می‌شود.

پارامترهای مدل را با استفاده از روش MLE برآورد می‌کنیم.

مدل دوجمله‌ای منفی

❖ **برآورد مدل رگرسیون دوجمله‌ای منفی،**

❖ تابع درستنمایی عبارتست از؛

$$L = \prod_{i=1}^{n} \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1)\,\Gamma(\frac{1}{\alpha})} (\frac{1}{1+\alpha\mu_i})^{(\frac{1}{\alpha})} (\frac{\alpha\mu_i}{1+\alpha\mu_i})^{y_i}$$

❖ لگاریتم تابع درستنمایی عبارتست از؛

$$l = \ln L = \sum_{i=1}^{n} \left\{ \ln \Gamma(y_i + \frac{1}{\alpha}) - \ln \Gamma(y_i + 1) - \ln \Gamma(\frac{1}{\alpha}) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + y_i[\ln(\alpha\mu_i) - \ln(1+\alpha\mu_i)] \right\}$$

❖ **توجه: در معادله بالا به جای μ_i رابطه $\mu_i = \exp(x_i'\beta)$ را درون معادله قرار داده و نسبت به پارامترهای مدل (α و β) مشتق گرفته و برابر صفر قرار می‌دهیم. بدین ترتیب برآوردگرهای $\hat{\alpha}$ و $\hat{\beta}$ به دست می‌آید.**

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ برآورد مدل رگرسیون دوجمله‌ای منفی،

نکته-۱: بعد از برآورد مدل، حتماً آزمون زیر انجام شود:

آزمون وجود بیش‌پراکندگی

$$H_0: \alpha = 0$$

اگر:

فرض صفر رد شود،
مدل دوجمله‌ای منفی نسبت به پواسون مناسب‌تر است.

نکته-۲: در مدل NB2، ضرایب β همانند مدل پواسون بر لگاریتم میانگین شرطی اثر می‌گذارند؛ بنابراین تفسیر آنها نیز بر حسب IRR یا تغییرات درصدی انجام می‌شود.

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ **برآورد مدل رگرسیون دوجمله‌ای منفی،**

❖ **نکته-۳:** در برآورد مدل‌های شمارشی مانند:

- ✓ پواسون،
- ✓ دوجمله‌ای منفی،
- ✓ صفر-انباشته،
- ✓ و مدل‌های پیچیده‌تر اقتصادسنجی،

معادلات حاصل از مشتق‌گیری تابع درستنمایی معمولاً فرم بسته ندارند؛ بنابراین، برآورد پارامترها با استفاده از روش‌های بهینه‌سازی عددی انجام می‌شود.

از جمله مهم‌ترین این روش‌ها می‌توان به الگوریتم‌های نیوتن–رافسون، فیشر اسکورینگ، BFGS و EMاشاره کرد.
همچنین در مدل‌های پیچیده‌تر، از روش‌های شبیه‌سازی مانند مونت‌کارلو و MCMC نیز استفاده می‌شود.

روش‌های Markov Chain Monte Carlo مانند Gibbs Sampling و Metropolis-Hastings

مدل دوجمله‌ای منفی

❖ **توضیح بیشتر در رابطه با Closed-form solution**

مثال ساده: رگرسیون خطی OLS

در رگرسیون خطی معمولی، جواب فرم بسته داریم:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

یعنی:

بدون تکرار،

بدون الگوریتم عددی،

و فقط با چند عملیات ماتریسی،

مستقیماً جواب به دست می‌آید.

این را «جواب فرم بسته» می‌گویند.

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ **توضیح بیشتر در رابطه با Closed-form solution**

اما در مدل پواسون یا دوجمله‌ای منفی چه اتفاقی می‌افتد؟

در این مدل‌ها:

- تابع درستنمایی غیرخطی است،
- پارامترها داخل تابع نمایی، لگاریتم، و گاما قرار دارند،
- و مشتق‌ها بسیار پیچیده می‌شوند.

مثلاً در مدل پواسون:

$$\mu_i = \exp(x_i'\beta)$$

وقتی مشتق log-likelihood را مساوی صفر قرار می‌دهیم، معادله‌ای شبیه زیر ایجاد می‌شود:

$$\sum_i x_i(y_i - \exp(x_i'\beta)) = 0$$

در این معادله:

- β هم داخل نمایی است،
- هم داخل مجموع،
- و امکان جدا کردن مستقیم β وجود ندارد.

بنابراین نمی‌توان نوشت:

$$\hat{\beta} = \ldots$$

به صورت یک رابطه صریح مثل OLS

مدل دوجمله‌ای منفی

مدرس: محمدمهدی بشارتی

❖ توضیح بیشتر در رابطه با Closed-form solution

در نتیجه چه کار می‌کنیم؟

به جای حل مستقیم، از روش‌های تکراری عددی استفاده می‌کنیم.

مثلاً:

1. یک مقدار اولیه برای β حدس می‌زنیم؛
2. مقدار تابع درستنمایی را حساب می‌کنیم؛
3. پارامترها را کمی اصلاح می‌کنیم؛
4. دوباره محاسبه می‌کنیم؛
5. این کار تکرار می‌شود تا الگوریتم به نقطه بهینه برسد.

خلاصه: معادلات حاصل از مشتق‌گیری تابع درستنمایی در بسیاری از مدل‌های شمارشی، به دلیل غیرخطی بودن، دارای جواب صریح و مستقیم (Closed-form solution) نیستند؛ بنابراین، پارامترهای مدل با استفاده از الگوریتم‌های تکراری بهینه‌سازی عددی برآورد می‌شوند.

تشبیه شهودی

مثل این است که:

در OLS فرمول دقیق پاسخ را داریم؛

اما در مدل‌های غیرخطی باید با «جستجوی عددی» بهترین جواب را پیدا کنیم.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

مدرس: محمدمهدی بشارتی

○ مثال

فرض کنید هدف پژوهش، مدل‌سازی **تعداد تصادفات ماهانه در قطعه‌های بزرگراهی یک شهر** باشد.

متغیر پاسخ (Y_i): "تعداد تصادفات در قطعه i طی یک ماه"

چون: تعداد تصادفات یک متغیر شمارشی است،
مقادیر آن ۰،۱،۲،... می‌باشد،
و معمولاً توزیع نرمال ندارد،
می‌توان از مدل رگرسیون پواسون استفاده کرد.

مشکل مدل پواسون

ابتدا مدل پواسون برآورد شده است.

نتایج نشان می‌دهد:

شاخص	مقدار
میانگین تعداد تصادفات	2.4
واریانس تعداد تصادفات	9.1

چون:

$$Var(Y) > E(Y)$$

بنابراین داده‌ها دارای بیش‌پراکندگی هستند و مدل پواسون مناسب نیست.

در نتیجه از مدل رگرسیون دوجمله‌ای منفی استفاده می‌کنیم.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

مدرس: محمدمهدی بشارتی

○ مثال

فرض کنید هدف پژوهش، مدل‌سازی **تعداد تصادفات ماهانه در قطعه‌های بزرگراهی یک شهر** باشد.

متغیر پاسخ (Y_i): "تعداد تصادفات در قطعه i طی یک ماه"

فرض کنید مدل زیر برآورد شده است:

$$\ln \mu_i = -1.80 + 0.06\ \mathrm{Traffic_i} + 0.75\mathrm{Rain_i} - 0.4\mathrm{Camera_i} + \ 0.03\mathrm{Sp}eed_{\mathrm{i}}$$

که در آن:

$$\mu_i = E(Y_i|\mathbf{x}_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}} = e^{\beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}}$$

یعنی μ_i= تعداد مورد انتظار تصادفات در ماه.

تعریف متغیرها

متغیر	توضیح
$Traffic_i$	حجم ترافیک (هزار وسیله در روز)
$Rain_i$	بارندگی (1=بارانی، 0=خشک)
$Camera_i$	وجود دوربین کنترل سرعت
$Speed_i$	متوسط سرعت خودروها (km/h)

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

- **مثال**

۱. حجم ترافیک

ضریب:

$$0.06$$

تفسیر لگاریتمی

با افزایش ۱۰۰۰ وسیله نقلیه در روز:

$$100 \times 0.06 = 6\%$$

تعداد مورد انتظار تصادفات تقریباً ۶٪ افزایش می‌یابد.

تفسیر IRR

$$IRR = e^{0.06} = 1.062$$

یعنی:

هر ۱۰۰۰ وسیله اضافه،

نرخ مورد انتظار تصادفات را حدود ۶.۲٪ افزایش می‌دهد.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

- **مثال**

۲. بارندگی

ضریب:

$$0.75$$

$$IRR = e^{0.75} = 2.12$$

تفسیر

در شرایط بارانی، نرخ مورد انتظار تصادفات حدود ۲.۱ برابر شرایط خشک است.

یا:

$$112\%$$

افزایش دارد.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

- **مثال**

۳. دوربین کنترل سرعت

ضریب:

$$-0.40$$

$$IRR = e^{-0.40} = 0.67$$

تفسیر

وجود دوربین کنترل سرعت:

نرخ مورد انتظار تصادفات را به ۰.۶۷ برابر کاهش می‌دهد،

یعنی حدود ۳۳٪ کاهش ایجاد می‌کند.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

- **مثال**

۴. سرعت متوسط

ضریب:

$$0.03$$

$$IRR = e^{0.03} = 1.03$$

تفسیر

هر ۱ کیلومتر-بر-ساعت افزایش در سرعت متوسط:

فراوانی تصادفات را حدود ۳٪ افزایش می‌دهد.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

مثال

نقش پارامتر بیش‌پراکندگی

فرض کنید مدل NB2 مقدار زیر را برآورد کرده است:

$$\hat{\alpha} = 0.85$$

و معنادار نیز هست.

تفسیر

چون:

$$\alpha > 0$$

داده‌ها دارای بیش‌پراکندگی هستند و استفاده از مدل دوجمله‌ای منفی نسبت به پواسون مناسب‌تر است.

مدل دوجمله‌ای منفی – تفسیر نتایج مدل

مدرس: محمدمهدی بشارتی

- **مثال**

واریانس شرطی در NB2

$$Var(Y_i \mid x_i) = \mu_i + \alpha\mu_i^2$$

اگر برای یک قطعهi:

$$\mu_i = 2$$

باشد، آنگاه:

$$Var(Y_i \mid x_i) = 2 + 0.85(2^2) = 2 + 3.4 = 5.4$$

مشاهده می‌شود که:

واریانس بسیار بزرگ‌تر از میانگین است،

بنابراین فرض پواسون نقض می‌شود.

در مدل پواسون فرض می‌شود همه قطعات بزرگراهی مشابه (براساس Xها) دارای نرخ تصادف ثابتی هستند.
اما در دنیای واقعی، برخی مقاطع ذاتاً خطرناک‌ترند:

- کیفیت روسازی،
- طراحی هندسی،
- دید راننده،
- رفتار رانندگان،
- و عوامل مشاهده‌نشده دیگر

موجب می‌شوند نرخ تصادف بین مقاطع متفاوت باشد.

مدل دوجمله‌ای منفی این ناهمگنی مشاهده‌نشده را با استفاده از ترکیب پواسون-گاما مدل می‌کند.

مدل رگرسیون داده‌های شمارشی

فصل
دوازدهم

Count data Regression Model

مقدمه

مدل رگرسیون پواسون

مدل رگرسیون دوجمله‌ای منفی

مدل رگرسیون شمارشی با صفر زیاد

مدل رگرسیون شمارشی منقطع/سانسور شده

مدل‌های رگرسیون شمارشی با صفر زیاد

مدرس: محمدمهدی بشارتی

- تعداد صفرهای موجود در داده‌های شمارشی یکی از موضوعات مورد بررسی در مطالعات است.
- در مدل‌های پواسون و دوجمله‌ای منفی (در حالت معمول)، تعداد صفرها با تعداد مقادیر غیرصفر یک تناسب معقول دارند.
- بدین معنا که به عنوان مثال، در مدل پواسون، نسبت (احتمال) صفرها برابر با $e^{-\lambda_i}$ است.
- بنابراین، اگر نسبت صفرهای مشاهده شده در نمونه، خیلی بیشتر از این مقدار باشد، در این صورت نمی‌توان با مدل‌های مرسوم پواسون و دوجمله‌ای منفی، این داده‌ها را توصیف (مدلسازی) نمود.
- در این شرایط، مدل‌های مذکور را با استفاده از روش‌هایی تعدیل می‌کنیم تا بدین‌ترتیب این پدیده را مدلسازی لحاظ نماییم.
- دو نوع از رایج‌ترین مدل‌های رگرسیون شمارشی که این پدیده را در نظر می‌گیرند؛
  - مدل‌های هِردل (Hurdle models)
  - مدل‌های صفرانباشته یا پُرصفر (Zero-inflated models)

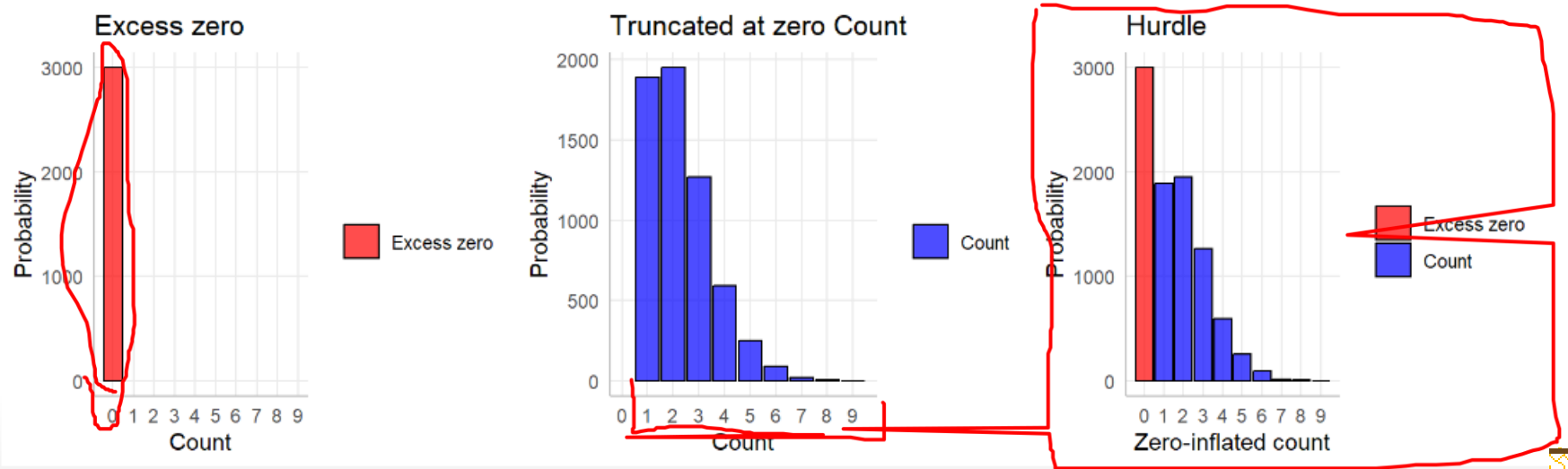
مدل‌های رگرسیون شمارشی با صفر زیاد

- ✓ **مدل‌های هِردل** زمانی مناسب‌اند که فرض شود **عبور از صفر و ورود به مقادیر مثبت**، یک فرآیند مجزا است؛ به‌طوری‌که تمام صفرها از یک فرآیند دوگزینه‌ای ایجاد می‌شوند.
- ✓ این مرز و آستانه **عبور از صفر و ورود به مقادیر مثبت** را hurdle (threshold) می‌نامند.
- ✓ **مدل‌های صفرانباشته** زمانی مناسب‌اند که فرض شود صفرها از **دو منبع متفاوت** ایجاد می‌شوند:

1. صفرهای ساختاری (همیشه صفر)
2. صفرهای تصادفی ناشی از فرآیند شمارشی

مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

- مدل دوجمله‌ای منفی هِردل (HNB) دارای دو بخش است؛
  - بخش اول: دوجمله‌ای (برای مدلسازی صفرها)
  - بخش دوم: دوجمله‌ای منفیِ منقطع از صفر (Zero-truncated negative binomial)
- از آنجا که مدل دوجمله‌ای منفی هِردل دارای دو بخش مجزاست؛ بنابراین، آن‌ها به طور همزمان برآورد نمی‌شوند.
- تفسیر پارامترها برای بخش صفر مانند یک مدل رگرسیون دوجمله‌ای است.



مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

❖ اجزای مدل دوجمله‌ای منفی هِردل (HNB)؛

- در بخش اول: یک توزیع دوجمله‌ای با پارامتر π برای $Y=0$ تعریف می‌شود (π: احتمال مشاهده صفر).
  - این بخش توسط یک مدل دوگزینه‌ای مانند رگرسیون لوجستیک (پروبیت، ...) برآورد می‌شود.
- در بخش دوم: یک توزیع دوجمله‌ای منفیِ منقطع از صفر با پارامتر μ برای $Y=1,2,\ldots$ تعریف می‌شود.
  - این بخش (مقادیر مثبت شمارشی) توسط یک مدل شمارشی منقطع از صفر (مانند مدل پواسون منقطع از صفر) برآورد می‌شود (در حالت وجود بیش‌پراکندگی: مدل دوجمله‌ای منفیِ منقطع از صفر)
  - مدل شمارشی منقطع از صفر: truncated-at-zero count mode

مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

❖ **اجزای مدل دوجمله‌ای منفی هِردل (HNB)؛**

۱- احتمال $P(y=0)$ را با $f_1(0)$ نشان می‌دهیم (همان π). بنابراین $1-f_1(0)$ بیانگر احتمال $Y>0$ است.

پس این بخش از مدل بیانگر یک مدل دو گزینه‌ای است، که

- به $y=0$ احتمال $f_1(0)$ را می‌دهد.
- به سایر مقادیر Y احتمال $1-f_1(0)$ را می‌دهد.

۲- برای $Y=0,1,2,\ldots$ تابع احتمال $f_2(y)$ تعریف می‌شود. اما احتمال مقادیر $Y=1,2,\ldots$ را در شرایطی محاسبه می‌کنیم که می‌دانیم y صفر نیست (منقطع). پس احتمال شرطی را به صورت $f_2(y|y>o)=\frac{f_2(y)}{1-f_2(0)}$ در نظر می‌گیریم.

$$P(Y=y)=P(y)=\begin{cases} f_1(y) & y=0 \\ \phi\, f_2(y) & y=1,2,\ldots \end{cases}$$

۳- اکنون مرحله ۱ و ۲ را ترکیب می‌کنیم. بدین صورت که؛

- احتمال $y=0$ برابر با $f_1(0)$ است و احتمال رفتن به بخش دوم مدل برابر با $1-f_1(0)$ است.
- احتمال هر یک از مقادیر $Y=1,2,\ldots$ برابر با $\frac{f_2(y)}{1-f_2(0)}$ است.
- بنابراین احتمال $Y=1,2,\ldots$ در مقابل $y=0$ برابر با $(1-f_1(0))\frac{f_2(y)}{1-f_2(0)}$ است.

$$\phi=\frac{1-f_1(0)}{1-f_2(0)}$$

مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

❖ اجزای مدل دوجمله‌ای منفی هِردل (HNB)؛

اگر تابع احتمال بخش ۱ و بخش ۲ از یک نوع باشند، در این صورت $f_1(y) = f_2(y)$ بوده و $\phi = 1$ می‌باشد.

در این حالت، مدل هردل تبدیل به مدل ساده $f(y)$ می‌شود؛ که هم برای توصیف $y = 0$ و هم برای توصیف $y = 1,2,\ldots$ به کار می‌رود.

مدل هردل:

$$P(Y = y) = P(y) = \begin{cases} f_1(y) \\ \phi\, f_2(y) \end{cases} = \begin{cases} \pi & y = 0 \\ (1-\pi)P(Y|y>0) & y = 1,2,\ldots \end{cases}$$

به گونه‌ای که :

$$P(Y = y|y > 0) = \frac{f(y)}{1 - f(0)}$$

امیدریاضی مدل هردل:

$$E(Y) = \frac{1 - f_1(0)}{1 - f_2(0)}\mu_i$$

μ_i: **میانگین توزیع داده شمارشی**

مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

❖ مدل هردل دوجمله‌ای منفی

$$P(Y=y)=P(y)=\begin{cases} f_1(y) \\ \phi\, f_2(y) \end{cases}=\begin{cases} \pi & y=0 \\ (1-\pi)P(Y|y>0) & y=1,2,\dots \end{cases}$$

$$\pi=P(y=0|\mathbf{z}_i)=f_1(0)=\frac{1}{1+\exp(\boldsymbol{z_i'\gamma})}$$

$$P(Y=y)=\begin{cases} \pi_i=\dfrac{1}{1+\exp(\boldsymbol{z_i'\gamma})} & ;\ y=0 \\ \left(\dfrac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}\right)\dfrac{\Gamma(y_i+r)}{\Gamma(y_i+1)\,\Gamma(r)}\left(\dfrac{r}{r+\exp(\mathbf{x_i'\beta})}\right)^{(r)}\left(\dfrac{\exp(\mathbf{x_i'\beta})}{r+\exp(\mathbf{x_i'\beta})}\right)^{y_i}\left[1-\left(\dfrac{r}{r+\exp(\mathbf{x_i'\beta})}\right)^{r}\right]^{-1} & ;\ y=1,2,\dots \end{cases}$$

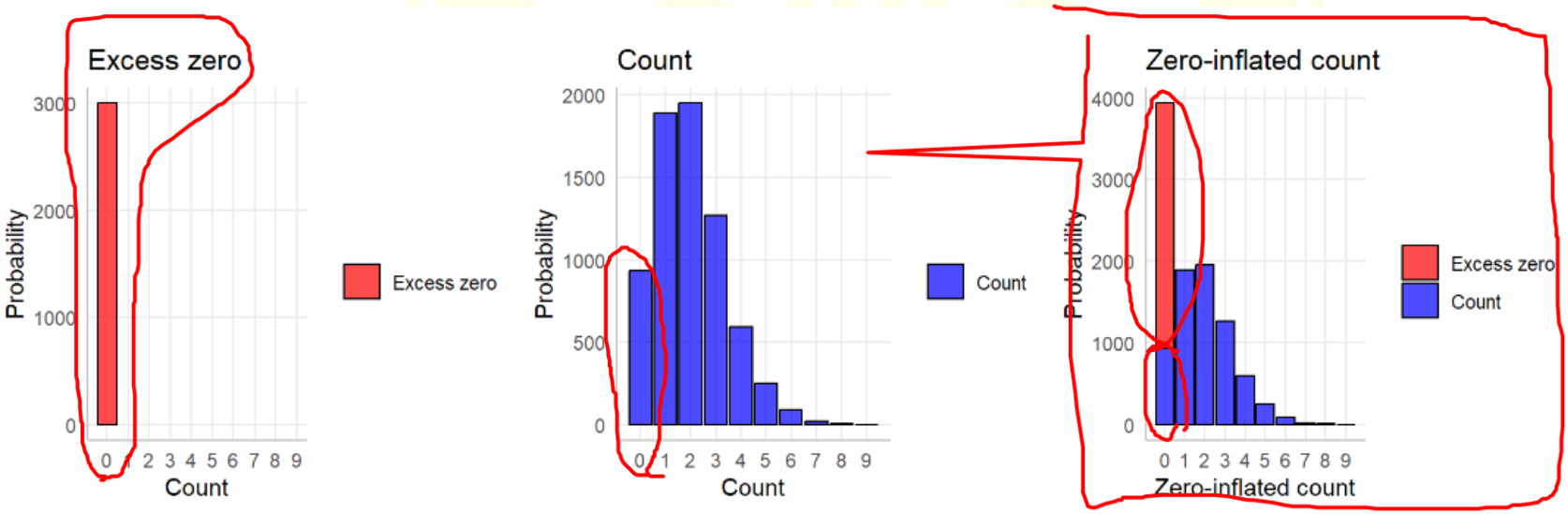
❖ پارامترهای مدل‌های بخش اول و دوم به صورت جداگانه برآورد می‌شوند (به همین دلیل، متغیرهای $\mathbf{z_i'}$ و $\mathbf{x_i'}$ می‌توانند متفاوت باشند).

$$E(Y|y>0)=\frac{1-f_1(0)}{1-f_2(0)}\mu_i=\frac{1-\pi_i}{1-\left(\frac{r}{r+\mu_i}\right)^{r}}\mu_i=\left(\frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}\right)\left[1-\left(\frac{r}{r+\exp(\mathbf{x_i'\beta})}\right)^{r}\right]^{-1}\exp(\mathbf{x_i'\beta})$$

❖ معادله بالا، رگرسیون دوجمله‌ای منفی هردل است (پارامترهای مدل: $\boldsymbol{\gamma}$ و $\boldsymbol{\beta}$).

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

- در مدل **هردل**، برای حالت $Y=0$ یک فرآیند دوگزینه‌ای و برای $Y>0$ یک فرآیند شمارشی منقطع از صفر تعریف می‌شود.
- در مدل‌های **صفرانباشته**، صفرها به دو گروه تقسیم می‌شوند؛
  - گروهی از صفرها که همیشه صفر هستند؛ و به آنها صفرهای مازاد (Excess Zeros) می‌گوییم. (واحدهایی که به طور کلی هیچگاه با رویداد مورد مطالعه، مواجه نبوده‌اند)
  - گروه دوم، صفرهای تصادفی هستند. اینها بخشی از فرآیند شمارشی هستند. (واحدهایی که با رویداد مورد مطالعه، مواجه بوده‌اند، اما مقدار مشاهده شده برای آن واحد برابر با صفر بوده است)



مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

- فرض کنید احتمال صفرهای مازاد برابر با ϕ باشد. یعنی ϕ٪ از مشاهدات را صفرهای مازاد تشکیل دهند.
- از سوی دیگر، فرآیند شمارشی پواسون را به صورت روبرو برای $y = 0,1,2, \dots$ تعریف می‌کنیم. $P(y) = \frac{e^{-\lambda}\lambda^{Y}}{y!}$
- در این فرآیند شمارشی، احتمال $Y = 0$ برابر است با؛ $P(0) = e^{-\lambda}$
- بنابراین، احتمال $Y = 0$ برای مجموع دو حالت (صفرهای مازاد و تصادفی) برابر است با؛ $P(y = 0) = \phi + (1 - \phi)e^{-\lambda}$
  - احتمال صفرهای تصادفی : $(1 - \phi)e^{-\lambda}$
  - احتمال صفرهای مازاد : ϕ

$$P(Y = y) = P(y) = \begin{cases} f_1(y) \\ \phi\, f_2(y) \end{cases} = \begin{cases} \phi + (1 - \phi)e^{-\lambda} & y = 0 \\ (1 - \phi)\frac{e^{-\lambda}\lambda^{Y}}{y!} & y = 1,2, \dots \end{cases}$$

- احتمال مقادیر غیرصفر برابر است با؛ $P(y) = (1 - \phi)\frac{e^{-\lambda}\lambda^{Y}}{y!} \;;\; y = 1,2, \dots$

- این درحالی است که در مدل هردل، احتمال همه صفرها با یکدیگر برابر بوده و به صورت زیر است؛

$$P(Y = y) = P(y) = \begin{cases} f_1(y) \\ \phi\, f_2(y) \end{cases} = \begin{cases} \pi & y = 0 \\ (1 - \pi)\frac{e^{-\lambda}\lambda^{Y}}{y!}(1 - e^{-\lambda})^{-1} & y = 1,2, \dots \end{cases}$$

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

- فرض کنید احتمال صفرهای مازاد برابر با ϕ باشد. یعنی ϕ٪ از مشاهدات را صفرهای مازاد تشکیل دهند.
- از سوی دیگر، فرآیند شمارشی پواسون را به صورت روبرو برای $y = 0,1,2, \dots$ تعریف می‌کنیم. $P(y) = \frac{e^{-\lambda}\lambda^{Y}}{y!}$
- در این فرآیند شمارشی، احتمال $Y = 0$ برابر است با؛ $P(0) = e^{-\lambda}$
- بنابراین، احتمال $Y = 0$ برای مجموع دو حالت (صفرهای مازاد و تصادفی) برابر است با؛ $P(y = 0) = \phi + (1 - \phi)e^{-\lambda}$

$$P(Y = y) = P(y) = \begin{cases} f_1(y) \\ \phi\, f_2(y) \end{cases} = \begin{cases} \phi + (1-\phi)e^{-\lambda} & y = 0 \\ (1-\phi)\frac{e^{-\lambda}\lambda^{Y}}{y!} & y = 1,2, \dots \end{cases}$$

- احتمال صفرهای تصادفی : $(1 - \phi)e^{-\lambda}$
- احتمال صفرهای مازاد : ϕ

- احتمال مقادیر غیرصفر برابر است با؛ $P(y) = (1-\phi)\frac{e^{-\lambda}\lambda^{Y}}{y!} \;;\; y = 1,2, \dots$

- امیدریاضی Y برابر است با؛

$$E(Y) = \left[\phi + (1-\phi)e^{-\lambda}\right] \times 0 + \sum_{y=1}^{\infty}(1-\phi)\frac{e^{-\lambda}\lambda^{Y}}{y!}y = (1-\phi)\lambda$$

- واریانس Y برابر است با؛

$$Var(Y|\mathbf{x}_i) = (1-\phi)\lambda(1+\lambda\phi)$$

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

- اکنون فرض کنید برای توصیف λ معادله رگرسیون $\lambda_i = \exp(\boldsymbol{x_i'\beta})$ را داشته باشیم.
- برای ϕ نیز یک تابع لوجستیک در نظر می‌گیریم.
- ϕ بیانگر احتمال موفقیت (یعنی مازاد بودن صفرها) است، پس ϕ_i برابر است با؛

$$\phi_i = \frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}$$

- با تشکیل تابع درستنمایی و مشتق‌گیری از لگاریتم تابع درستنمایی نسبت به $\boldsymbol{\beta}$ و $\boldsymbol{\gamma}$، برآوردگرها به دست می‌آید.
- امیدریاضی Y برابر است با؛

$$E(Y|\mathbf{x}_i,\mathbf{z}_i) = (1-\phi_i)\lambda_i = \left(\frac{1}{1+\exp(\boldsymbol{z_i'\gamma})}\right)\exp(\boldsymbol{x_i'\beta})$$

- واریانس Y برابر است با؛

$$Var(Y|\mathbf{x}_i) = (1-\phi_i)\lambda_i(1+\lambda_i\phi_i)$$

- **معادله رگرسیون مدل پواسون با صفرانباشته (ZIP):**

$$Y_i = E(Y_i|\mathbf{x}_i,\mathbf{z}_i) + u_i = \left(\frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}\right)\exp(\boldsymbol{x_i'\beta}) + u_i$$

- با برآورد این مدل، برآوردهای $\boldsymbol{\beta}$ و $\boldsymbol{\gamma}$ به دست می‌آید.
- با برآورد $\boldsymbol{\beta}$ و $\boldsymbol{\gamma}$، می‌توان ϕ_i و λ_i و سایر احتمال‌های موردنظر را محاسبه نمود.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

- **پارامترهای $\boldsymbol{\beta}$ و $\boldsymbol{\gamma}$** را می‌توان از معادلات زیر نیز به دست آورد؛

$$\lambda_i = \exp(\boldsymbol{x_i'\beta}) \quad \Rightarrow \quad \ln \lambda_i = \boldsymbol{x_i'\beta}$$

$$\phi_i = \frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})} \quad \Rightarrow \quad \ln \frac{\phi_i}{1-\phi_i} = \boldsymbol{z_i'\gamma} + u_i$$

- با برآورد دو معادله بالا، برآوردهای $\boldsymbol{\beta}$ و $\boldsymbol{\gamma}$ به دست می‌آید.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

- **مدل رگرسیون دوجمله‌ای منفی با صفر انباشته (ZINB)**
- **اگر از توزیع دوجمله‌ای منفی استفاده کنیم، روابط زیر را می‌توان تعریف کرد؛**

$$P(Y=y_i)=\begin{cases}\phi_i+(1-\phi_i)\left(\frac{r}{r+\mu_i}\right)^r & ;\ y=0\\ (1-\phi_i)\frac{\Gamma(y_i+r)}{\Gamma(y_i+1)\,\Gamma(r)}\left(\frac{r}{r+\mu_i}\right)\left(\frac{\mu_i}{r+\mu_i}\right)^{y_i} & ;\ y=1,2,\dots\end{cases}$$

$$\mu_i=\exp(\boldsymbol{x_i'\beta}) \qquad \phi_i=\frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}$$

- **بنابراین، در حالت کلی، باید یک معادله برای برآورد ϕ_i و یک معادله برای برآورد میانگین موردنظر (در توزیع پواسون λ_i و در توزیع دوجمله‌ای منفی μ_i) تعریف کنیم.**
- **توجه: برای ϕ_i و λ_i یا μ_i می‌توان توابع مختلفی تعریف کرد.**
- **مثلا در اینجا برای ϕ_i از تابع لوجستیک و برای λ_i (μ_i) از تابع نمایی استفاده کردیم.**

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

- **خلاصه:**
- **تفاوت اصلی مدل هردل و مدل صفرانباشته در نحوه تولید صفرها است.**

✓ در مدل هردل، تمام صفرها از بخش دوگزینه‌ای مدل ایجاد می‌شوند و بخش شمارشی فقط مقادیر مثبت را تولید می‌کند.

✓ اما در مدل صفرانباشته، صفرها می‌توانند هم از فرآیند صفر ساختاری و هم از فرآیند شمارشی ایجاد شوند.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

❖ مثال حمل‌ونقلی برای مدل Hurdle

فرض کنید هدف پژوهش، مدل‌سازی «تعداد تصادفات جرحی رانندگان تاکسی در طول یک سال» باشد.

در داده‌های جمع‌آوری‌شده مشاهده می‌شود که بخش بزرگی از رانندگان هیچ تصادف جرحی نداشته‌اند ($Y = 0$)، در حالی که گروهی دیگر دارای ۱، ۲، ۳ و یا تعداد بیشتری تصادف هستند.

در مدل هردل فرض می‌شود که فرآیند ایجاد مقدار صفر، با فرآیند ایجاد مقادیر مثبت متفاوت است.

به بیان دیگر، ابتدا راننده باید از «مانع وقوع تصادف» عبور کند تا وارد بخش شمارشی شود.

بنابراین مدل شامل دو بخش است:

بخش اول: مدل دوگزینه‌ای

این بخش احتمال وقوع حداقل یک تصادف را مدل می‌کند:

$$P(Y > 0)$$

برای مثال:

- سن راننده،
- سابقه رانندگی،
- نوع شیفت کاری،
- ساعات رانندگی شبانه

می‌توانند بر احتمال ورود راننده به گروه دارای تصادف اثر بگذارند.

این بخش معمولاً با **رگرسیون لوجیت یا پروبیت** برآورد می‌شود.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

❖ مثال حمل‌ونقلی برای مدل Hurdle

فرض کنید هدف پژوهش، مدل‌سازی «تعداد تصادفات جرحی رانندگان تاکسی در طول یک سال» باشد.

در داده‌های جمع‌آوری‌شده مشاهده می‌شود که بخش بزرگی از رانندگان هیچ تصادف جرحی نداشته‌اند ($Y = 0$)، در حالی که گروهی دیگر دارای ۱، ۲، ۳ و یا تعداد بیشتری تصادف هستند.

در مدل هردل فرض می‌شود که فرآیند ایجاد مقدار صفر، با فرآیند ایجاد مقادیر مثبت متفاوت است.

به بیان دیگر، ابتدا راننده باید از «مانع وقوع تصادف» عبور کند تا وارد بخش شمارشی شود.

بنابراین مدل شامل دو بخش است:

بخش دوم: مدل شمارشی منقطع از صفر

پس از آنکه راننده وارد گروه دارای تصادف شد ($Y > 0$)، تعداد تصادفات وی مدل می‌شود.

در این مرحله:

فقط مقادیر مثبت ($1,2,3,\ldots$) مورد تحلیل قرار می‌گیرند.

اگر داده‌ها دارای بیش‌پراکندگی باشند، معمولاً از مدل دوجمله‌ای منفی منقطع از صفر استفاده می‌شود.

برای مثال:

- میزان پیمایش سالانه،
- تراکم ترافیک مسیر،
- تعداد ساعات رانندگی روزانه،
- خستگی راننده

می‌توانند تعداد تصادفات را تعیین کنند.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

❖ مثال حمل‌ونقلی برای مدل Hurdle

تفسیر مفهومی مدل Hurdle

در مدل هردل:

- تمام صفرها از یک فرآیند مستقل تولید می‌شوند،
- و تمام مقادیر مثبت از یک فرآیند شمارشی جداگانه حاصل می‌شوند.

بنابراین:

- ابتدا احتمال «وقوع یا عدم وقوع تصادف» مدل می‌شود،
- سپس «تعداد تصادفات» برای رانندگانی که حداقل یک تصادف داشته‌اند تحلیل می‌شود.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

❖ مثال حمل‌ونقلی برای مدل Zero-Inflated (ZIP / ZINB)

فرض کنید پژوهشگری قصد دارد «تعداد تصادفات موتورسواران در طول یک سال» را مدل‌سازی کند.

در داده‌ها تعداد زیادی مقدار صفر مشاهده می‌شود.
اما بررسی دقیق‌تر نشان می‌دهد که این صفرها از دو گروه متفاوت تشکیل شده‌اند:

گروه اول: صفرهای ساختاری (Excess Zeros)

برخی موتورسواران:

- تقریباً هیچ‌گاه در معرض خطر تصادف نیستند،
- مثلاً فقط در کوچه‌های محلی و با سرعت پایین رانندگی می‌کنند،
- یا بسیار کم از موتورسیکلت استفاده می‌کنند.

این افراد عملاً همیشه صفر تولید می‌کنند.

❖ مثال حمل‌ونقلی برای مدل Zero-Inflated (ZIP / ZINB)

فرض کنید پژوهشگری قصد دارد «تعداد تصادفات موتورسواران در طول یک سال» را مدل‌سازی کند.

در داده‌ها تعداد زیادی مقدار صفر مشاهده می‌شود.

اما بررسی دقیق‌تر نشان می‌دهد که این صفرها از دو گروه متفاوت تشکیل شده‌اند:

گروه دوم: صفرهای تصادفی

گروه دیگری از موتورسواران:

- در معرض واقعی خطر تصادف هستند،
- اما در سال مورد بررسی به صورت تصادفی، تصادفی نداشته‌اند.

این صفرها بخشی از فرآیند شمارشی معمول هستند.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

❖ **ساختار مدل ZIP/ZINB**

مدل صفرانباشته فرض می‌کند که داده‌ها از ترکیب دو فرآیند تولید شده‌اند:

بخش اول: مدل لوجیت

احتمال تعلق مشاهده به گروه «صفرهای ساختاری» را مدل می‌کند.

برای مثال: متغیرهای زیر می‌توانند تعیین کنند که آیا فرد اساساً در معرض تصادف قرار دارد یا خیر.

- میزان استفاده از موتورسیکلت،
- نوع مسیر،
- هدف سفر،
- سابقه رانندگی

بخش دوم: مدل شمارشی

اگر مشاهده متعلق به گروه صفرهای ساختاری نباشد، تعداد تصادفات با یک مدل شمارشی مدل می‌شود:

پواسون (ZIP)

یا دوجمله‌ای منفی (ZINB)

در صورت وجود بیش‌پراکندگی، مدل ZINB مناسب‌تر است.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

❖ تفسیر مفهومی مدل ZIP/ZINB

در مدل صفرانباشته:

- ✓ صفرها دو منشأ مختلف دارند،
- ✓ بخشی از صفرها «همیشه صفر» هستند،
- ✓ و بخشی دیگر «صفرهای تصادفی» ناشی از فرآیند شمارشی‌اند.

بنابراین برخلاف مدل هردل:

- ✓ در ZIP/ZINB صفرها می‌توانند از هر دو بخش مدل ایجاد شوند،
- ✓ اما در مدل هردل تمام صفرها فقط از بخش اول تولید می‌شوند.

تفاوت مفهومی Hurdle و ZIP در مثال حمل‌ونقل

مدل Hurdle

سؤال اصلی:

«آیا راننده وارد گروه دارای تصادف می‌شود یا خیر؟»

پس از عبور از مانع:

تعداد تصادفات مدل می‌شود.

مدل ZIP/ZINB

سؤال اصلی:

«آیا این راننده اساساً در معرض تصادف قرار دارد یا جزو گروه همیشه‌صفر است؟»

سپس:

تعداد تصادفات برای افراد در معرض خطر مدل می‌شود.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

مدل	منبع صفرها	بیش‌پراکندگی
Poisson	عادی	ندارد
NB	عادی	دارد
Hurdle	فقط ساختاری	ممکن است
ZIP	ساختاری + تصادفی	ممکن است
ZINB	ساختاری + تصادفی	دارد

به طور کلی:

- اگر بیش‌پراکندگی وجود نداشته باشد، مدل پواسون مناسب است.
- در صورت وجود بیش‌پراکندگی، مدل دوجمله‌ای منفی ترجیح داده می‌شود.
- اگر تعداد صفرها بیش از حد معمول باشد، مدل‌های هردل یا صفرانباشته مناسب‌تر خواهند بود.
- در صورت وجود همزمان بیش‌پراکندگی و صفرهای مازاد، مدل ZINBمعمولاً عملکرد بهتری دارد.

❖ انتخاب بین مدل‌های شمارشی

- در عمل، معمولاً چندین مدل مختلف برای داده‌های شمارشی قابل استفاده هستند؛ مانند مدل پواسون، دوجمله‌ای منفی، مدل‌های هردل، و مدل‌های صفرانباشته.
- از آنجا که هر یک از این مدل‌ها فرض‌های متفاوتی درباره پراکندگی داده‌ها و نحوه تولید صفرها دارند، لازم است مناسب‌ترین مدل با استفاده از معیارها و آزمون‌های آماری انتخاب شود.
- یکی از روش‌های متداول، مقایسه مقدار تابع درستنمایی (Likelihood) مدل‌ها است.
- به طور کلی، مدلی که مقدار لگاریتم درستنمایی بزرگ‌تری داشته باشد، برازش بهتری بر داده‌ها دارد.
- با این حال، افزایش تعداد پارامترها معمولاً موجب بهبود مصنوعی برازش می‌شود؛ بنابراین، برای مقایسه منصفانه مدل‌ها از معیارهای اطلاعاتی استفاده می‌شود.

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

❖ انتخاب بین مدل‌های شمارشی

معیار آکائیک (AIC)

$$AIC = -2\ln L + 2k$$

معیار بیزین (BIC)

$$BIC = -2\ln L + k\ln(n)$$

که در آن:

L: مقدار تابع درستنمایی،

k: تعداد پارامترهای مدل،

n: اندازه نمونه است.

- ✓ در هر دو معیار، مقادیر کوچک‌تر نشان‌دهنده مدل بهتر هستند.
- ✓ معیار BIC نسبت به AIC جریمه بیشتری برای مدل‌های پیچیده در نظر می‌گیرد.

مدل رگرسیون داده‌های شمارشی

فصل
دوازدهم

Count data Regression Model

مقدمه

مدل رگرسیون پواسون

مدل رگرسیون دوجمله‌ای منفی

مدل رگرسیون شمارشی با صفر زیاد

مدل رگرسیون شمارشی منقطع/سانسور شده

مدل‌های رگرسیون شمارشی سانسورشده و منقطع

مدرس: محمدمهدی بشارتی

- **مقدمه**

- **گاهی اوقات مجبوریم بنا به دلایل مختلف، بخشی از داده‌های موردنظر را کنار بگذاریم.**
- **مدل‌های سانسورشده (Censored) و منقطع (Truncated) برای چنین مواردی طراحی شده‌اند.**
- **هر یک از این دو نوع مدل‌ها، به دو شکل هستند؛**
  - **الف) منقطع/سانسور شده از پایین،**
  - **ب) منقطع/سانسور شده از بالا.**

✓ وقتی برخی از مقادیر متغیر شمارشی به صورت ساختاری در داده‌ها مشاهده نمی‌شوند (مثلاً مقدار صفر وجود ندارد)، باید تابع احتمال توزیع شمارشی را تعدیل کنیم تا برای این نوع داده‌ها مناسب شود.

✓ در چنین شرایطی، بخشی از فضای نمونه حذف می‌شود؛ بنابراین لازم است تابع احتمال مجدداً نرمال‌سازی شود تا مجموع احتمالات برابر با ۱ باقی بماند.

مدل‌های رگرسیون شمارشی سانسورشده و منقطع

مدرس: محمدمهدی بشارتی

- **مقدمه**

- **تفاوت اصلی مربوط به نحوه واکنش Y برای مقادیر ورای نقطه برش است؛**
  - ✓ در مدل‌های منقطع: مقادیر ورای نقطه برش را از نمونه حذف می‌کنیم.
  - ✓ در مدل‌های سانسور شده: مشاهده حذف نمی‌شود (بلکه فقط مقدار واقعی آن مشاهده نمی‌شود). یعنی مقادیر ورای نقطه برش را معادل با مقدار نقطه برش درنظر می‌گیریم. پس مشاهده در نمونه باقی می‌ماند، اما مقدار واقعی آن فراتر از نقطه برش قابل مشاهده نیست و با مقدار حدی ثبت می‌شود.

- **مدل‌های سانسورشده و منقطع شباهت زیادی دارند.**
  - ✓ از جمله اینکه در هر دو مدل، باید توابع احتمال و درستنمایی را به گونه‌ای تعدیل کنیم که تغییر در توزیع Y را لحاظ نمایند.

مدل‌های رگرسیون شمارشی سانسورشده و منقطع

مدرس: محمدمهدی بشارتی

- **مدل‌های منقطع**
- **برای توزیع پواسون،**

$$P(Y = y_i) = \frac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!} \qquad P(y = 0) = e^{-\lambda_i}$$

- **تابع احتمال توزیع پواسون منقطع:**

$$P(Y = y_i | y_i > 0) = \frac{P(y_i)}{P(y_i > 0)} = \frac{P(y_i)}{1 - P(y_i = 0)} = \frac{\frac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!}}{1 - e^{-\lambda_i}}$$

- **برای توزیع دوجمله‌ای منفی،**
- **مانند مدل پواسون منقطع (مدل بالا)، توزیع دوجمله‌ای منفی ($P(y_i)$) را بر احتمال $P(y = 0)$ تقسیم می‌کنیم ($\frac{P(y_i)}{P(y_i>0)}$) تا مدل دوجمله‌ای منفی منقطع به دست بیاید.**
- در توزیع دوجمله‌ای منفی احتمال $P(y = 0)$ به صورت روبرو به دست می‌آید. $P(y = 0) = (1 + \alpha\mu)^{-\frac{1}{\alpha}}$
- این رابطه مربوط به پارامتردهی NB2 است.
- **توجه:** مدل‌های بالا، منقطع شده در صفر است می‌توان مدل‌های منقطع در هر مقدار دلخواه A را نیز ساخت.

مدل‌های رگرسیون شمارشی سانسورشده و منقطع

مدرس: محمدمهدی بشارتی

- **مدل‌های سانسورشده**

- تفاوت اصلی مدل‌های **سانسورشده** با مدل‌های **منقطع** آن است که؛

  - برای **مدل‌های سانسورشده**، سطح برش A برابر با کمترین مقدار دلخواه برای Y خواهد بود.
    - مثلا اگر سطح برش (سانسور) را برابر با ۲ قرار دهیم، مقادیر Y که کمتر از ۲ هستند (یعنی ۰ و ۱) را با مقدار ۲ جایگزین می‌کنیم.
  - برای **مدل‌های منقطع**، سطح برش A به این معناست که مقادیر Y کمتر/مساوی A را از داده‌ها حذف می‌کنیم.
    - مثلا اگر سطح برش را برابر با ۲ قرار دهیم، مقادیر Y که کمتر/مساوی ۲ هستند (یعنی ۰ ، ۱ و ۲) را از داده‌ها حذف کرده و در برآورد مدل از آن‌ها استفاده نمی‌کنیم.

مدل‌های رگرسیون شمارشی سانسورشده و منقطع

❖ مثال ۱ — مدل منقطع (Truncated Count Model)

تحلیل تعداد تصادفات رانندگان پرخطر

فرض کنید یک شرکت بیمه قصد دارد تعداد تصادفات رانندگان پرخطر را مدل‌سازی کند.
اما داده‌های شرکت فقط شامل رانندگانی است که حداقل یک تصادف در سه سال اخیر داشته‌اند.

بنابراین:

رانندگان بدون تصادف در پایگاه داده وجود ندارند.

مقدار $Y = 0$ اساساً مشاهده نشده است.

در این حالت، استفاده از مدل پواسون معمولی منجر به تورش خواهد شد؛ زیرا داده‌ها از ابتدا در صفر منقطع شده‌اند.

بنابراین باید از:

- ✓ مدل پواسون منقطع از صفر،
- ✓ یا مدل دوجمله‌ای منفی منقطع از صفر

استفاده شود.

مثال ۲ — مدل سانسورشده (Censored Count Model)

تعداد تخلفات ثبت‌شده رانندگان

فرض کنید پلیس راهور تعداد تخلفات هر راننده را ثبت می‌کند، اما سیستم ثبت اطلاعات دارای سقف ۱۰ تخلف است.

بنابراین:

اگر راننده‌ای ۱۳ تخلف داشته باشد، در پایگاه داده مقدار ۱۰ ثبت می‌شود.

در این حالت:

- ✓ مشاهدات حذف نشده‌اند،
- ✓ اما مقادیر واقعی بزرگ‌تر از ۱۰ قابل مشاهده نیستند.

بنابراین داده‌ها در بالا سانسور شده‌اند و باید از:

- ✓ مدل شمارشی سانسورشده (Censored Poisson/NB) استفاده شود.

موضوع فصل بعد؟

❖ نکته در رابطه با برآورد مدل رگرسیون شمارشی

- در مدل‌های رگرسیون شمارشی، پارامترهای مدل معمولاً با استفاده از روش حداکثر درستنمایی (MLE) برآورد می‌شوند.
- در این رویکرد، هدف یافتن مقادیری از پارامترهاست که احتمال مشاهده داده‌های نمونه را بیشینه کند.
- اگرچه این روش در بسیاری از مسائل عملکرد مناسبی دارد، اما در مدل‌های پیچیده‌تر (مانند مدل‌های دارای ناهمگنی پنهان، پارامترهای تصادفی، ساختارهای سلسله‌مراتبی، یا مدل‌های صفر-انباشته) محاسبه تابع درستنمایی و برآورد پارامترها بسیار دشوار می‌شود.
- در چنین شرایطی، رویکرد بیزی (Bayesian Approach) چارچوب انعطاف‌پذیرتری برای برآورد مدل‌ها فراهم می‌کند.
- در روش بیزی، پارامترها به صورت متغیرهای تصادفی در نظر گرفته شده و با ترکیب اطلاعات پیشین (Prior Distribution) و اطلاعات نمونه، توزیع پسین (Posterior Distribution) پارامترها به دست می‌آید.
- با این حال، توزیع‌های پسین در اغلب مدل‌های اقتصادسنجی فرم بسته ندارند و محاسبه مستقیم آن‌ها امکان‌پذیر نیست. به همین دلیل، از روش‌های شبیه‌سازی و نمونه‌گیری عددی، به‌ویژه روش‌های زنجیره مارکوف مونت‌کارلو (Markov Chain Monte Carlo: MCMC)، برای تقریب توزیع پسین و برآورد پارامترها استفاده می‌شود.
- در فصل بعد، مبانی رویکرد بیزی و مهم‌ترین الگوریتم‌های MCMC مانند Gibbs Sampling و Metropolis-Hastings معرفی خواهند شد.