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

مقدمه

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، نمی‌توان از روش OLS بهره برد.

مقدمه

❑ متغیرهای تصادفی شمارشی؛

- 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 برای توصیف حوادثی به کار می‌رود که احتمال موفقیت (وقوع رویداد موردنظر) نسبتاً کم باشد.
- بر این مبنا، توزیع پواسون برای توصیف حوادث نادر 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

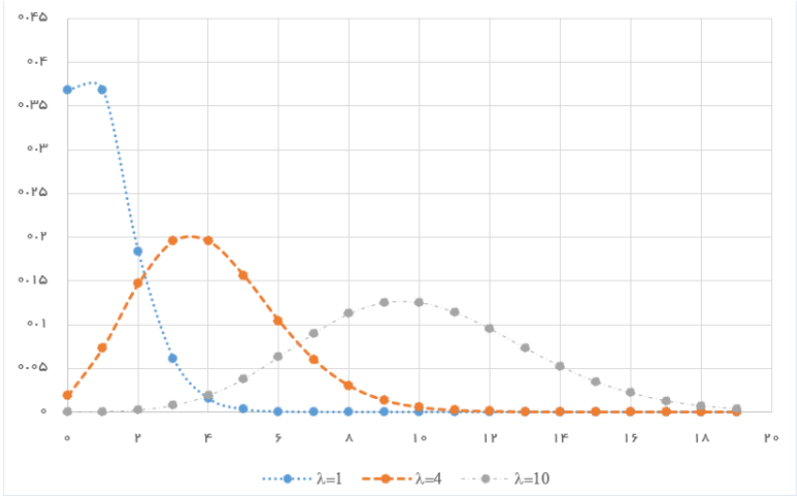
مدرس: محمدمهدی بشارتی

❖ 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



- **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 این معنا که هنگامی که 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، و Y_i ها نیز غیرمنفی هستند، بنابراین، آن را به صورت نمایی تعریف می‌کنند؛

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

- 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:

$$L = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\ {\lambda_i}^{Y_i}}{Y_i\ !}$$

مدل پواسون– تفسیر نتایج مدل پواسون

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

$$\frac{\partial \ln E(Y|\mathbf{x}_i)}{\partial \mathrm{X}_k} = \beta_k \quad \Longrightarrow \quad \frac{\partial \ln E(Y|\mathbf{x}_i)}{\partial \mathrm{X}_k} = \frac{1}{E(Y|\mathbf{x}_i)} = \beta_k$$

- 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 $E(Y|\mathbf{x}_i)$ به اندازه $100 \times \beta_k$ درصد تغییر خواهد کرد.
- 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.

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.

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

- یعنی یک واحد افزایش/کاهش در مقدار متغیر X_k می‌تواند **ریسک نسبی وقوع رخداد موردنظر** را به اندازه e^{β_k} افزایش/کاهش دهد (در شرایط ثابت ماندن سایر متغیرها)

- یعنی به ازای یک واحد افزایش/کاهش در مقدار متغیر X_k ، مقدار متغیر پاسخ، e^{β_k} برابر می‌شود.

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

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

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.
- 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 بیش از حد واقعی، بزرگ یا کوچک شوند؛ و بنابراین، آزمون معناداری پارامترها را مخدوش می‌کند.

مدل پواسون– محدودیت‌های مدل پواسون

مدرس: محمدمهدی بشارتی

❖ 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.))

❖ 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 $\hat{\beta}$ در مدل پواسون: $SE_p(\hat{\beta})$

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

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

- مدل رگرسیون پواسون از توزیع احتمال پواسون به دست می‌آید.
- مدل رگرسیون لوجیت از توزیع احتمال لوجستیک به دست می‌آید.
- مدل رگرسیون خطی نرمال از توزیع احتمال نرمال به دست می‌آید.
- مدل رگرسیون دوجمله‌ای منفی از یک تابع توزیع احتمال خاص به دست می‌آید.
- 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).
- این ترکیب، تنها یکی از راه‌های استخراج توزیع دوجمله‌ای منفی است.
- ویژگی مهم این روش آن است که به ما امکان مدل‌سازی ناهمگنی را در مدل پواسون می‌دهد..

مدل دوجمله‌ای منفی

❖ **تابع احتمال دوجمله‌ای منفی،**

○ **تابع احتمال توزیع دوجمله‌ای،**

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

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

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 رویداد موردنظر (موفقیت/شکست) در x-اُمین آزمایش برای بار k-اُم رخ دهد.

احتمال مشاهده r-امین موفقیت/شکست در x-امین آزمایش

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؛

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

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 $\qquad E(Y) = \mu$

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

- α **ضریب بیش‌پراکندگی** است. به ازای $\alpha = 0$ **پراکندگی یکسان**، به ازای $\alpha > 0$ **بیش‌پراکندگی** و به ازای $\alpha < 0$ **کم‌پراکندگی** وجود دارد.

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؛**

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

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

- **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 (تجمعی) دوجمله‌ای منفی 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

- **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؛

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

همچنین معادله میانگین شرطی به صورت زیر خواهد بود؛

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

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؛

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

❖ **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 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

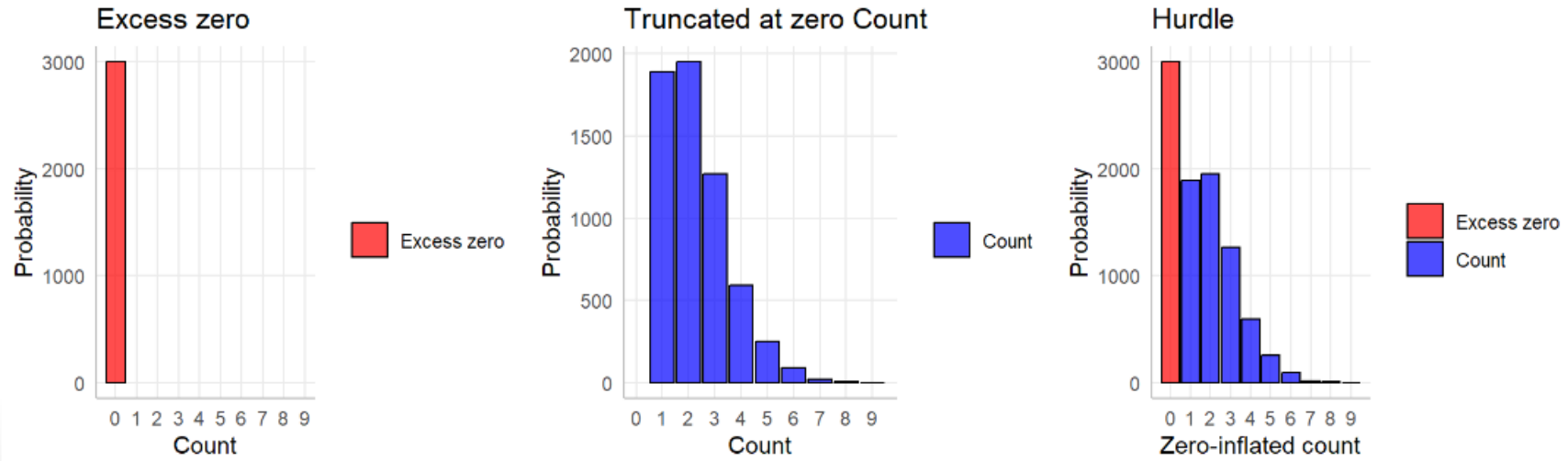
مدل‌های رگرسیون شمارشی با صفر زیاد

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– مدل دوجمله‌ای منفی هِردل

- مدل دوجمله‌ای منفی هِردل (HNB) دارای دو بخش است؛
  - بخش اول: دوجمله‌ای
  - 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)

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.

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

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

- 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,… به کار می‌رود.

مدل هردل:

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

امیدریاضی مدل هردل:

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

μ_i: **میانگین توزیع داده شمارشی**

مدل‌های رگرسیون شمارشی با صفر زیاد– مدل دوجمله‌ای منفی هِردل

❖ مدل هردل دوجمله‌ای منفی

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

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

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

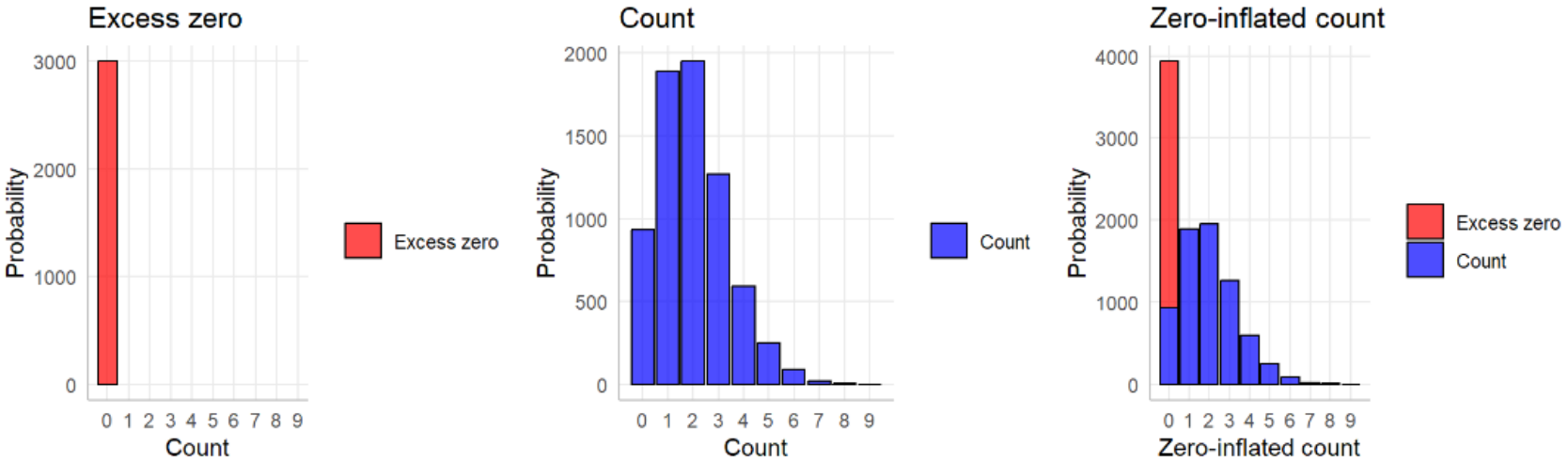
$$E(Y|y>0)=\frac{1-f_1(0)}{1-f_2(0)}\mu_i=\frac{1-\pi_i}{1-\left(\dfrac{r}{r+\mu_i}\right)^{r}}\mu_i=\left(\frac{\exp(\boldsymbol{z_i'\gamma})}{1+\exp(\boldsymbol{z_i'\gamma})}\right)\left[1-\left(\frac{r}{r+\exp(\mathbf{x_i'\beta})}\right)^{r}\right]^{-1}\exp(\mathbf{x_i'\beta})$$

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

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

- 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

مدرس: محمدمهدی بشارتی

- فرض کنید احتما 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 احتما 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؛

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

- 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؛

$$\lambda_i = \exp(\boldsymbol{x_i'\beta}) \Rightarrow \ln \lambda_i = \boldsymbol{x_i'\beta} \Rightarrow \ln Y_i = \boldsymbol{x_i'\beta} + u_i$$

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

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

مدل‌های رگرسیون شمارشی با صفر زیاد—مدل‌های صفر انباشته

مدرس: محمدمهدی بشارتی

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

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

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

- 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.

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.

- تفاوت اصلی مربوط به نحوه واکنش Y برای مقادیر ورای نقطه برش است؛
  - ✓ در مدل‌های منقطع: مقادیر ورای نقطه برش را حذف می‌کنیم.
  - ✓ در مدل‌های سانسور شده: مقادیر ورای نقطه برش را معادل با مقدار نقطه برش درنظر می‌گیریم.

- 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.**

$$P(y = 0) = (1 + \alpha\mu)^{-\frac{1}{\alpha}}$$

- در توزیع دوجمله‌ای منفی احتمال $P(y = 0)$ به صورت روبرو به دست می‌آید.
- **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 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.