# 数理統計 補助資料
# ～ポアソン回帰モデル～

2023年度2学期：　月曜1限, 水曜3限
担当教員：　石垣　司



## ポアソン分布(離散分布)

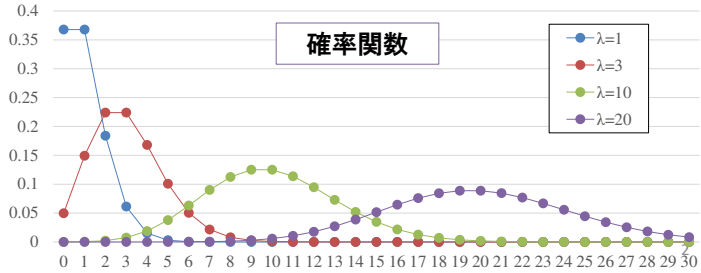
Siméon Denis Poisson
(1781-1840)

### 稀にしか起きない事象の発生回数の分布

例：1日に起こる交通事故や倒産の回数
例：1時間以内の電話や来店客の回数

- 確率関数：　$\mathrm{Poisson}(X=x)=\frac{\lambda^x}{x!}\exp(-\lambda)$
- パラメータ：　$\lambda$（単位時間当たりの平均発生回数）
- 平均と分散：　$E[X]=V[X]=\lambda$

※ポアソン分布の意義と使い方は単位時間(1秒, 1時間, 1日など)に依存する。単位時間の扱いに注意






## ポアソン回帰モデル #1

### 目的変数がカウントデータ(非負の整数)の回帰モデル

| 【説明変数】 | 【目的変数】 |
|---|---|
| 空気の汚染度 | 1年間の発病者数 |
| 広告視聴時間 | 1か月間の商品の販売個数 |
| 気温の累計 | 1シーズンの野生植物の発芽個体数 |
| 総合学習授業の時間 | ある年度の志望校への合格者数 |

※　線形回帰モデルを利用すると発病者数の予測がマイナスとなり得る

### ポアソン分布を利用した回帰モデル

- 事象の発生回数がポアソン分布に従うと仮定した回帰モデル
- ポアソン分布の平均パラメータを説明変数 $x$ へ回帰
  説明変数の増減によってパラメータ $\mu$ が増減するモデル
- 目的変数の予測値が必ず非負



## ポアソン回帰モデル #2

### 説明変数が1つのポアソン回帰モデル

- データ：$\{x_i, y_i\}\ (i=1,\ldots,N)$
- 目的変数：$y\in\{0,\mathbb{Z}^+\}$
- 説明変数：$x\in\mathbb{R}$
- パラメータ：$\{a,b\}$
- ポアソン分布の確率関数：$\Pr(y)=\frac{\mu^y}{y!}\exp(-\mu)$
- ポアソン回帰モデル

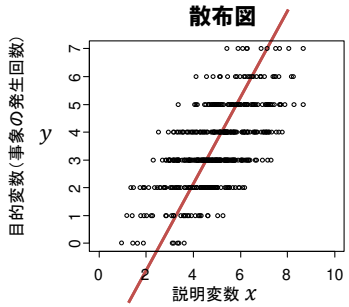
$$\mu=\exp(a+bx)$$

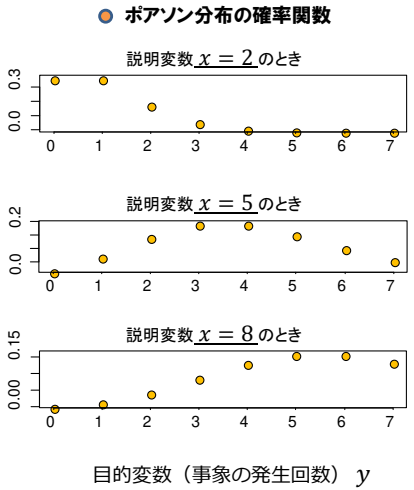
$\mu>0$ が必要なため線形モデルの指数関数を利用

$$\Pr(y|x)=\frac{\exp(a+bx)^y}{y!}\exp\{-\exp(a+bx)\}$$

# ポアソン回帰モデル #3

## ポアソン回帰モデルによる予測




**線形回帰モデルによる予測**
$x = 2$ **のとき**, $\hat{y} = -0.8 < 0$

**ポアソン回帰モデルによる予測**
$x = 2$ **のとき**,
$\hat{y} = E[y] = \exp(\hat{a} + 2\hat{b}) > 0$



# ポアソン回帰モデル #4

## 説明変数が複数あるポアソン回帰モデル

- **目的変数：変数** $y$
- **説明変数：変数** $\boldsymbol{x} = [x_1, x_2, \cdots, x_P]^T$
- **データ：** $D = \{y_i, \boldsymbol{x}_i\}(i = 1, \cdots, N)$
  $\boldsymbol{x}_i = [1 \quad x_{i1} \quad \cdots \quad x_{iP}]^T$（**行列** $X$ **の** $i$ **番目の行**）
- **回帰係数：係数** $b_0, b_1, \cdots, b_P$（**パラメータ**）

$$\boldsymbol{b} = [b_0 \quad b_1 \quad \cdots \quad b_P]^T$$

$$\mu = \exp(\boldsymbol{x}^T \boldsymbol{b})$$

$$\Pr(y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T \boldsymbol{b})^y}{y!} \exp\{-\exp(\boldsymbol{x}^T \boldsymbol{b})\}$$



# ポアソン回帰係数の解釈

## パラメータの対数との線形関係としての解釈

$$\mu = \exp(a + bx) \Leftrightarrow \log(\mu) = a + bx$$

- **説明変数** $x$ **が**1**単位増減すると, 平均パラメータの対数がポアソン回帰係数** $b$ **の分だけ増減する**

## パラメータと指数関数の積として解釈

- $\mu = \exp(a + bx) \Leftrightarrow \mu = e^a e^{bx}$
- **説明変数** $x$ **が**1**単位増減すると, 平均パラメータが** $e^b$ **倍になる**

#メモ ロジスティック回帰係数の解釈と同じロジック。



# ポアソン回帰分析をやってみよう

## ポルトガルの公立高校の生徒の成績データ

- 2005**年から**2006**年の実データ** ($N = 649$)
- **家庭環境と欠席傾向の分析**

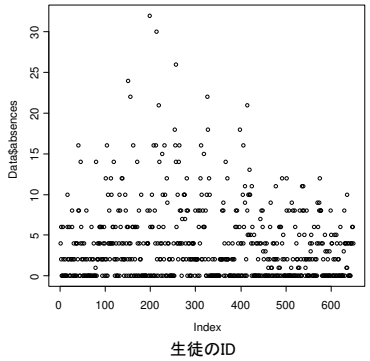
- **目的変数：**1**年間の欠席の回数**
- **説明変数：家族との関係性**(5**段階評価**), **家族の人数**(2**値変数:**4**人以上=**1), **母親の学歴**(5**段階回答**), **父親の学歴**(**同左**), **母親の職業**(**教職, ヘルスケア, 市民サービス, 在宅, その他**), **父親職業**(**同左**), **勉強時間**(5**段階回答**), **健康状態**(5**段階評価**), **放課後の自由時間**(5**段階評価**)

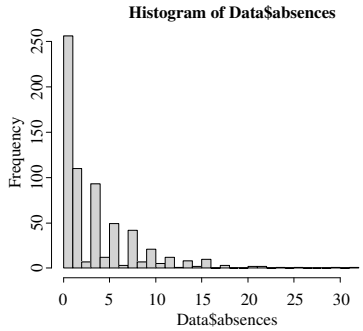
#メモ UCI Machine Learning Repository (http://archive.ics.uci.edu/ml)でStudent Performance Data Setとして無料でダウンロード可能。参照：P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. Proc. FUBUTEC 2008, 5-12

# データ内の欠席の傾向

## ポルトガルの公立高校の生徒の成績データ

- 目的変数：1年間の欠席の回数






# Rによるポアソン回帰分析

## Rのコードと出力

### Rのコード

```
Data = read.csv("student-por.csv",header=T)
pois =
glm(absences~famrel+famsize+Medu+Fedu+M
job+Fjob+studytime+freetime+health, family =
poisson, data=Data)
summary(pois)
```

### データ

```
> head(Data)
  school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
1     GP   F  18       U     GT3       A    4    4  at_home  teacher
2     GP   F  17       U     GT3       T    1    1  at_home    other
3     GP   F  15       U     LE3       T    1    1  at_home    other
4     GP   F  15       U     GT3       T    4    2   health services
5     GP   F  16       U     GT3       T    3    3    other    other
6     GP   M  16       U     LE3       T    4    3 services    other
     reason guardian traveltime studytime failures schoolsup famsup paid
1    course   mother          2         2        0       yes     no   no
2    course   father          1         2        0        no    yes   no
3     other   mother          1         2        0       yes     no   no
4      home   mother          1         3        0        no    yes   no
5      home   father          1         2        0        no    yes   no
6 reputation   mother          1         2        0        no    yes   no
  activities nursery higher internet romantic famrel freetime goout Dalc
1         no     yes    yes       no       no      4        3     4    1
2         no      no    yes      yes       no      5        3     3    1
3         no     yes    yes      yes       no      4        3     2    2
4        yes     yes    yes      yes      yes      3        2     2    1
5         no     yes    yes       no       no      4        3     2    1
6        yes     yes    yes      yes       no      5        4     2    1
  Walc health absences G1 G2 G3
1    1      3        4  0 11 11
2    1      3        2  9 11 11
3    3      3        6 12 13 12
4    1      5        0 14 14 14
5    2      5        0 11 13 13
6    2      5        6 12 12 13
```

### ポアソン回帰分析の結果

```
Call:
glm(formula = absences ~ famrel + famsize + Medu + Fedu + studytime +
    Mjob + Fjob + freetime + health, family = poisson, data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5625  -2.4738  -0.7692   0.9041   8.8098

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.193001   0.141995  15.444  < 2e-16 ***
famrel      -0.115086   0.020706  -5.558 2.73e-08 ***
famsizeLE3   0.011960   0.045209   0.265 0.791360
Medu         0.005896   0.028144   0.209 0.834073
Fedu         0.098157   0.025949   3.783 0.000155 ***
studytime   -0.207607   0.026528  -7.826 5.04e-15 ***
Mjobhealth  -0.601397   0.120507  -4.991 6.02e-07 ***
Mjobother    0.073978   0.057545   1.286 0.198594
Mjobservices 0.206945   0.068591   3.017 0.002552 **
Mjobteacher -0.173329   0.098331  -1.763 0.077950 .
Fjobhealth  -0.185097   0.149156  -1.241 0.214618
Fjobother   -0.142269   0.079298  -1.794 0.072798 .
Fjobservices -0.277581  0.085560  -3.244 0.001177 **
Fjobteacher -0.404071   0.126437  -3.196 0.001394 **
freetime    -0.018380   0.019500  -0.943 0.345919
health      -0.026193   0.014344  -1.826 0.067837 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3464.7  on 648  degrees of freedom
Residual deviance: 3277.1  on 633  degrees of freedom
AIC: 4682.2

Number of Fisher Scoring iterations: 6
```



# ポアソン回帰係数の解釈 #1

## ポアソン回帰係数の推定結果（最尤推定）

| | Estimate (推定値) | Std.Error (標準誤差) | t value (t値) | Pr(>\|t\|) (p値) |
|---|---|---|---|---|
| 切片 ($b_0$) | 2.193 | 0.142 | 15.444 | 0.000*** |
| 家族との関係性 ($b_1$) | -0.115 | 0.021 | -5.558 | 0.000*** |
| 家族の人数 ($b_2$) | 0.012 | 0.045 | 0.265 | 0.791 |
| 母親学歴 ($b_3$) | 0.006 | 0.028 | 0.209 | 0.834 |
| 父親学歴 ($b_4$) | 0.098 | 0.026 | 3.783 | 0.000*** |
| 母職ヘルスケア ($b_5$) | -0.601 | 0.121 | -4.991 | 0.000*** |
| 母職その他 ($b_6$) | 0.074 | 0.058 | 1.286 | 0.199 |
| 母職市民サービス ($b_7$) | 0.207 | 0.069 | 3.017 | 0.003** |
| 母職教職 ($b_8$) | -0.173 | 0.098 | -1.763 | 0.078. |
| 父職ヘルスケア ($b_9$) | -0.185 | 0.149 | -1.241 | 0.215 |
| 父職その他 ($b_{10}$) | -0.142 | 0.079 | -1.794 | 0.073. |
| 父職市民サービス ($b_{11}$) | -0.278 | 0.086 | -3.244 | 0.001** |
| 父職教職 ($b_{12}$) | -0.404 | 0.126 | -3.196 | 0.001** |
| 勉強時間 ($b_{13}$) | -0.208 | 0.027 | -7.826 | 0.000*** |
| 放課後自由時間 ($b_{14}$) | -0.018 | 0.020 | -0.943 | 0.346 |
| 健康状態 ($b_{15}$) | -0.026 | 0.014 | -1.826 | 0.068. |



# ポアソン回帰係数の解釈 #2

## ポアソン回帰係数の推定結果（5%水準で有意な変数を抜粋）

| | Estimate (推定値) | | Pr(>\|t\|) (p値) |
|---|---|---|---|
| 家族との関係性 ($b_1$) | -0.115 | $\exp(\hat{b}_1) = 0.89$ | 0.000*** |
| 父親学歴 ($b_4$) | 0.098 | $\exp(\hat{b}_4) = 1.10$ | 0.000*** |
| 母職ヘルスケア ($b_5$) | -0.601 | $\exp(\hat{b}_5) = 0.55$ | 0.000*** |
| 母職市民サービス ($b_7$) | 0.207 | $\exp(\hat{b}_6) = 1.23$ | 0.003** |
| 父職市民サービス ($b_{11}$) | -0.278 | $\exp(\hat{b}_{11}) = 0.76$ | 0.001** |
| 父職教職 ($b_{12}$) | -0.404 | $\exp(\hat{b}_{12}) = 0.67$ | 0.001** |
| 勉強時間 ($b_{13}$) | -0.208 | $\exp(\hat{b}_{13}) = 0.81$ | 0.000*** |

- 家族との関係が良好との回答であれば，$\mu$ は0.89倍になる（欠席が減る）
- 父親の学歴が高ければ，$\mu$ は1.1倍になる（欠席が増える）
- 母親がヘルスケア関連職ならば，在宅と比べて $\mu$ は0.55倍になる（欠席が減る）
- 母親が市民サービス職ならば，在宅と比べて $\mu$ は1.23倍になる（欠席が増える）
- 父親が市民サービス職ならば，在宅と比べて $\mu$ は0.76倍になる（欠席が減る）
- 父親が教職ならば，在宅と比べて $\mu$ は0.67倍になる（欠席が減る）
- 勉強時間が長いほど，$\mu$ は0.81倍になる（欠席が減る。"勉強時間"の定義については元データHP参照）

# 補足：主体によって試行回数などが異なる場合

## 適切な比較のためのオフセット項の導入

- **事象の出現回数 ≈ 試行回数**（期間や面積なども含む）**× 出現確率**
  - 1年間の発病者の人数 ⇒ 地域の居住者（例：大都市と村）
  - 1か月間の商品の販売個数 ⇒ 来店客数（例：晴天と台風の日の販売個数）
  - 1シーズンの野生植物の発芽個体数 ⇒ 調査面積
  - 先の欠席傾向の分析では，全生徒の登校可能日数は同じとみなして非導入

- **主体 $i$ の試行回数等（オフセット項 $n_i$）を含めたモデル化**

$$\mu_i = n_i \exp(a + bx_i),$$
$$\log(\mu_i) = \log(n_i) + a + bx_i$$

  - 試行回数等がそろっていなくても割り算などで誤魔化す必要はない

- **Rでのコードの書き方**

pois = glm(y~x, family = poisson, data=Data, offset = log("オフセット変数名"))



# ポアソン回帰モデルの最尤推定

## ポアソン回帰モデルの尤度関数

$$l(\boldsymbol{b}|D) = \prod_{i=1}^{N} \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{b}\right)^{y_i}}{y_i!} \exp\{-\exp\left(\boldsymbol{x}_i^T \boldsymbol{b}\right)\}$$

## ポアソン回帰モデルの対数尤度関数

$$L(\boldsymbol{b}|D) = \sum_{i=1}^{N} \{y_i \boldsymbol{x}_i^T \boldsymbol{b} - \exp\left(\boldsymbol{x}_i^T \boldsymbol{b}\right) - \log(y_i!)\}$$



# 一般化線形モデルでの説明 #1

**確率変数 $y_i$ の確率密度関数を**

$$f(y_i|\boldsymbol{b}, s, X) = \exp\left\{\frac{y_i \boldsymbol{x}_i^T \boldsymbol{b} - f_1(\boldsymbol{b})}{f_2(s)} + f_3(y_i, s)\right\}$$

**と表記すると，関数 $f_1(\boldsymbol{b}), f_2(s), f_3(y_i, s)$，定数 $s$ をそれぞれ適切に設定することで，線形回帰モデル，ロジスティック回帰モデル，ポアソン回帰モデルなどを統一的枠組みで取り扱うことができる**

- **$y_i$ の実現値に対する対数尤度関数**

$$L(\boldsymbol{b}|y_i, x_i) = \log\{f(\boldsymbol{b}|y_i, x_i)\} = \frac{y_i \boldsymbol{x}_i^T \boldsymbol{b} - f_1(\boldsymbol{b})}{f_2(s)} + f_3(y_i, s)$$

#メモ　一般化線形モデルは指数型分布族に属する確率分布を統一的に扱うための枠組みを提供する。本授業ではこれ以上詳しくは言及しないが、興味のある学生は一般化線形モデル、指数型分布族、リンク関数などのキーワードで理解を深めてほしい。



# 一般化線形モデルでの説明 #2

| 線形回帰モデル | ロジスティック回帰モデル | ポアソン回帰モデル |
|---|---|---|
| $f_1(\boldsymbol{b}) = \frac{1}{2}(\boldsymbol{x}_i^T \boldsymbol{b})^2$ | $f_1(\boldsymbol{b}) = -\log\left(1 + e^{\boldsymbol{x}_i^T \boldsymbol{b}}\right)$ | $f_1(\boldsymbol{b}) = \exp\left(\boldsymbol{x}_i^T \boldsymbol{b}\right)$ |
| $f_2(s) = \sigma^2$ | $f_2(s) = 1$ | $f_2(s) = 1$ |
| $f_3(y_i, s) = -\frac{1}{2}\left\{\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}$ | $f_3(y_i, s) = 0$ | $f_3(y_i, s) = -\log(y_i!)$ |

**データ $D$ に対する対数尤度関数**　$L(\boldsymbol{b}|D) = \sum_{i=1}^{N} \left\{\frac{y_i \boldsymbol{x}_i^T \boldsymbol{b} - f_1(\boldsymbol{b})}{f_2(s)} + f_3(y_i, s)\right\}$

| 線形回帰モデルの対数尤度関数 | ロジスティック回帰モデルの対数尤度関数 | ポアソン回帰モデルの対数尤度関数 |
|---|---|---|
| $L(\boldsymbol{b}|D) = -N\log\left(\sqrt{2\pi\sigma^2}\right)$ $-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T \boldsymbol{b}\right)^2$ | $L(\boldsymbol{b}|D) = -\sum_{i=1}^{N} y_i \log\left(1 + e^{-\boldsymbol{x}_i^T \boldsymbol{b}}\right)$ $+\sum_{i=1}^{N}(1 - y_i)\log\left(1 - \frac{1}{1 + e^{-\boldsymbol{x}_i^T \boldsymbol{b}}}\right)$ | $L(\boldsymbol{b}|D)$ $= \sum_{i=1}^{N}\{y_i \boldsymbol{x}_i^T \boldsymbol{b}$ $- \exp\left(\boldsymbol{x}_i^T \boldsymbol{b}\right) - \log(y_i!)\}$ |

# 一般化線形モデルでの説明 #3

**確率変数 $y_i$ の確率密度関数 $f(y_i|\boldsymbol{b}, s, X)$ や対数尤度関数 $L(\boldsymbol{b}|D)$ で成り立つ性質は，線形，ロジスティック，ポアソン回帰のすべての統計モデルでも成り立つ**

- **例：$\theta_i = \boldsymbol{x}_i^T \boldsymbol{b}$，とし $f(y_i|\boldsymbol{b}, s, X) = f(y_i|\theta_i, s)$ と書く。また，$\log(f) = \log\{f(y_i|\theta_i, s)\}$ と書き，$\log(f)$ は $\theta_i$ で2階微分可能とする。このとき，確率変数 $y_i$ の期待値に関して次の関係が成り立つ**

$$E\left[\frac{d\log(f)}{d\theta_i}\right] = 0, E\left[\frac{d^2\log(f)}{d{\theta_i}^2}\right] = -E\left[\left(\frac{d\log(f)}{d\theta_i}\right)^2\right]$$

#メモ1　つまり、$f(y_i|\theta, s)$ や $L$ の性質を調べることで、個別の統計モデルに関する性質を調べることなく、共通する性質を見つけ出すことができる
#メモ2　上の期待値の関係の証明は煩雑なので、本授業では割愛



# 演習問題

**問題：$\theta_i = x_i b$（切片が 0，説明変数の数が1つの回帰モデル）を考える。$f_1(b)$ が2階微分可能とすると，データ $D$ が与えられた時の対数尤度関数**

$$L(b|D) = \sum_{i=1}^{N}\left\{\frac{y_i x_i b - f_1(b)}{f_2(s)} + f_3(y_i, s)\right\}$$

**は，パラメータ $b$ に対して凸関数となることを示しなさい**

**ヒント：$\frac{d^2L(b|D)}{db^2} \leq 0$ を示せばよい**

- **確率変数 $y_i$ の確率密度関数 $f$ と尤度関数 $L$ は数式上は同じ形なので，$L(b|y_i, x_i) = \log(f)$ として計算してよい**
- $E\left[\frac{d^2\log(f)}{db^2}\right] = -E\left[\left(\frac{d\log(f)}{db}\right)^2\right]$

#メモ　ここでは問題を授業時間内で解答可能とするため最も単純な回帰モデルを考えたが、重回帰モデルでも同様の結果を得ることができる。