Chapter 3　　Count Regression

Count Regression：
Response variable（応答変数: y）が正の整数もしくは0
・数に上限がある場合→　Binomial Regression
・数に上限がない場合→　Poisson Regression
・数(平均値？)が十分に大きい場合→　Normal Liner Regression も可能
この章では Poisson と（ちょっとマイナーな）Negative binomial について解説する。

3.1　Poisson Regression　＝Poisson な世界へようこそ＝

例１：　ある地域における特殊な病気の患者数
ある地域の患者数なので、数に上限（ある地域の人口）があるため、Binomial distribution への近似が適しているが、事象の発生件数（ある特殊な病気の患者数）がとても少なく、考えられる最大値がとても大きい場合は Poisson distribution に近似する事も可能

例２：　起こる間隔、確率が互いに独立な事象が一定時間に起こる数
ある期間（１時間、午後とか）にサービスセンターにかかってくる電話の数
一定期間におこる地震の回数
→現実にはそれぞれの事象がおこる確率は互いに独立ではないが、Poisson distribution で十分に近似できる場合が多い。

例３：　事象が起こる間隔がそれぞれ独立で指数関数的に分布している場合

血液型が人種、性別によって異なるか？というように response variable がカテゴリーに分かれる場合は Poisson distribution では近似できない。
→multinominal response model, categorical data analysis を利用しよう

☆　Poisson distribution の特殊な性質
- 平均値μi とμj の Poisson distribution の和は平均値（μi＋μj）の Poisson distribution になる。（p=56　l=7）
- Poisson の確率変数のパラメーターは平均値μの一つ。縛りがつよい…。
- 平均値＝μ＝分散（平均と分散は比例するはず…）（p=59, Fig3.2）
- over dispersion かどうかは Deviance をＸ二乗値の近似として利用して計算できる。φ＞１ならば Negative binomial も考えよう？(p=60)

### 3.2　Rate Model ＝offset を上手に使おう＝

ここの疑問：　異常染色体の観察数（ca）は実験に利用している細胞の数(cells)に比例するため、ca を単純に response variable としたモデルでは不適切では？

**良くある例：　response variable を ca ではなく、ca/cells として直線回帰する。**
問題点？：　djustedR^2 は悪くないが、等分散ではないようだ…。
ca/cells>0 なのに正規分布仮定している…。
Response variable は 1/10 も 100/1000 も同じ 0.1 になってしまう。

**時々ある例：　カウントデータなので glm( family=poisson, link function=log)とする。**
cells は ca に対して比例(multiplicative effect)すると仮定するので log(cells)を説明変数に入れる
問題点？：　実際の log(cells)の係数は 1.0025 となったので良い感じ♪
良い感じだけど、そもそも係数を１としたモデルを作ればいいのでは？

**オススメ例：　log(cells)は offset 項として glm( family=poisson, link function=log)する。**
問題点？：　Poisson 分布で ca が cells に比例することも表現できた！
リンク関数が logit の時は解釈が難しい…（実用性なし？？）ので、分布が Poisson か Negative Binomial でリンク関数が log の時に有効。

### 3.3　Negative binomial　＝Poisson に収まらない時は…＝

Negative binomial distribution:
統計的に独立なベルヌーイ試行（成功／失敗、表／裏などの 1／0 で表せるような試行）を行ったとき、r 回の「成功」を得るのに必要な試行回数の分布。
Poisson distribution を一般化したもので、Poisson は Negative binomial の特別な形。

Link parameter $k$ を指定する必要がある。→glm(, negative.binomial($k$))
glm.nb()を使えば、$k$ は最尤推定されるので、指定する必要はない。

私は…、、Poisson では過分散となってしまう時に利用していますが、おそらくもっと奥が深いはずです。glmmML()では negative.binomial は使えないので random effect を考慮する場合は、glmmML(family=poisson)としています。どのモデルが適しているかは AIC で評価できます。