Modelling Proportions and Count Data

Rick White

May 4, 2016

Outline

- Analysis of Count Data
- Binary Data Analysis
- Categorical Data Analysis
- Generalized Linear Models
- Questions

Types of Data

Continuous data: any value in a specified range is possible.

- Normal distribution: entire real line
- Regression and ANOVA models

Count data: non-negative integer valued

- Poisson distribution
- Negative Binomial distribution

Categorical data: non numeric

- ordinal or nominal (binary is a special case)

Count Data

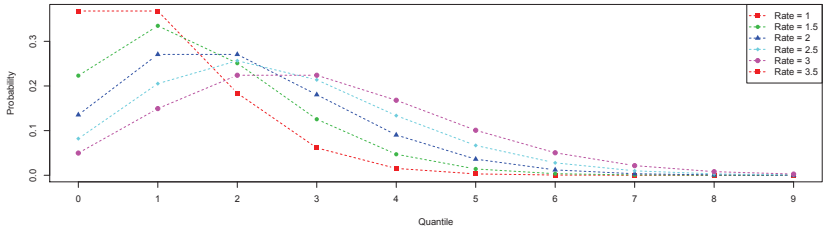
Data are non-negative integers arising from counting not ranking.

We are interested in modelling the rate (count/unit time). If data are observed over varying time periods then we need to standardize the counts to make them comparable. Any analysis must adjust for these varying times.

When making comparisons we usually talk about the relative rate. For adverse events this is usually referred to as the relative risk.

Main distributions are Poisson and Negative Binomial.

Poisson Distribution



Rate need not be an integer but only integer values have a positive probability.

Mean and variance both equal the rate.

The sum of Poisson's is Poisson, just sum the rates.

If the rate is ≥ 20 then Poisson $\sim$ Normal.

Estimating the Rate with a 95% CI for Poisson data.

We know the variance equals the mean for Poisson data.

If the sum of the data is ≥ 20 then we can use a normal approximation.

Estimate the rate with $\bar{x} = \sum_1^n x_i/n$.

Then a 95% CI is given by $\bar{x} \pm 1.96\sqrt{\bar{x}/n}$

If the sum is < 20 then normal approximation may not be very good and a more exact method should be used.

Comparing 2 Poisson rates

We summarize the groups by the sample means and we compute the overall mean of the data.

If the sum in each group is ≥ 20 then we do the following.

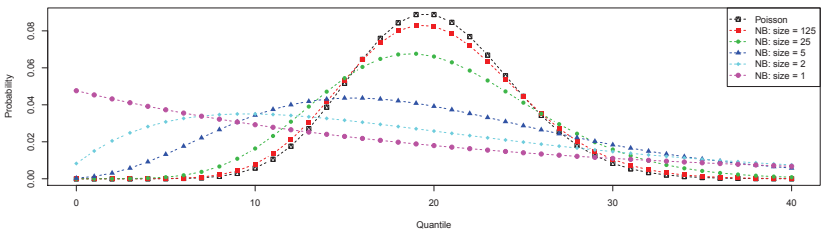
$\bar{x} = \sum_1^n x_i/n \qquad \bar{y} = \sum_1^m y_j/m \qquad \bar{z} = \frac{\sum_1^n x_i + \sum_1^m y_j}{n+m}$

Then under the null hypothesis

$\frac{\bar{x}-\bar{y}}{\sqrt{\bar{z}}} / \sqrt{\frac{1}{n} + \frac{1}{m}} \sim N(0,1)$

There are exact methods to compare the rate ratio of 2 groups.

Negative Binomial Distribution (NB)



All of the above have a rate of 20. Note the high variation in the shapes of the curves.

May have seen NB described as the number of failures until the rth success, where a success occurs with probability p.

Sounds like r is an integer but r can be any positive number.

$\mu = r\frac{1-p}{p} \qquad \sigma^2 = r\frac{1-p}{p^2}$

Negative Binomial (Continued)

In most analyses we are interested in the mean of the distribution. Therefore we prefer to use the mean (μ) as one of the parameters of the distribution. The second parameter can be either r or p.

If we use p then we define an over dispersion parameter $\tau^2 = 1/p$ and $\sigma^2 = \tau^2\mu$. If $\tau = 1$, we have a Poisson. This is quite often referred to as over-dispersed Poisson Regression.

If we use r then $\sigma^2 = \mu + \mu^2/r$. As $r \to \infty$ the distribution becomes Poisson.

These are the parametrizations typically used in Negative Binomial Regression.

Estimating the Rate and 95% CI for NB data

Most count data we encounter in practice has $\sigma^2 > \mu$.

Compute the sample mean and variance.

$\bar{x} = \sum_1^n x_i/n$ $\qquad$ $s^2 = \sum_1^n (x_i - \bar{x})^2/(n-1)$

$\tau^2 = \sigma^2/\mu$ $\qquad$ is an estimate of the over dispersion.

$r = \mu^2/(\sigma^2 - \mu)$ $\qquad$ is an estimate of the size parameter.

We use the normal approximation for the 95% CI

$\bar{x} \pm 1.96 s/\sqrt{n}$.

This approximation may not be very good.

Poisson and Negative Binomial Regression

Both model the mean or rate as a function of other variables.

Model: $\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$

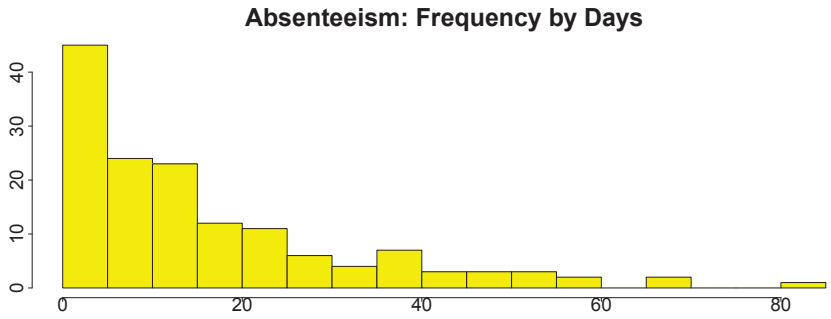
Predictors can be categorical or continuous. Interactions terms can be in the model.

The Poisson assumption ($\sigma^2 = \mu$) needs to be checked. If the assumption is violated, the model can dramatically overstate the significance of the predictors.

Poisson regression can include an over dispersion parameter. This is similar but not identical to negative binomial regression.

Negative Binomial regression will estimate either a single value for r or τ in addition to the rate. This allows σ^2 to be greater than the rate.

Absenteeism from School in Rural New South Wales




```
##                Eth  N  Mean    Var   V/M    r
## 1 (N)ot Aboriginal 77 12.18 183.89 15.10 0.86
## 2     (A)boriginal 69 21.23 313.95 14.79 1.54
```

Negative Binomial model

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Days
##     LR Chisq Df Pr(>Chisq)
## Eth    12.09  1   0.000507

## RR (A/N)    2.5 %   97.5 %
##    1.743    1.275    2.383

## r = 1.157
```

Overdispersed Poisson model

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Days
##     LR Chisq Df Pr(>Chisq)
## Eth    12.14  1   0.000492

## RR (A/N)    2.5 %   97.5 %
##    1.743    1.270    2.393

## Overdispersion = 13.14
```

Categorical Data

Two main types Nominal and Ordinal

Nominal data is differentiated by label but otherwise there is no logical order. (Gender, Ethnicity, Species)

Ordinal data is differentiated by a label that allows a logical order but the magnitude of difference cannot be established. (Likert Scales)

Binary data can be either ordinal or nominal. With only two possible outcomes, it is very easy to deal with. We can code the two outcomes as 0 or 1 but this is only an indicator that an outcome has occurred not an indication of order or a real number.

Binary Data 1/0 (special case of categorical data)

Binary data need not be coded as 1/0. It can be be coded as any binary indicator such as True/False, Success/Failure, etc.

We are interested with estimating the probability of each outcome. Although knowing one completely defines the other.

$P(X = 1) = p \qquad P(X = 0) = 1 - p$

Another parametrization is the odds $= p/(1 - p)$

Or log odds (called the logit function)

$\eta = logit(p) = log(\frac{p}{1-p}) = log(p) - log(1 - p).$

We saw something similar with the negative binomial distribution.

Binomial Distribution

If we have n independent samples of a binary variable then x_i is 0 or 1 for each trial.

If we sum our variable the $X = \sum x$ is the number of times our sample gave us a value of 1.

Our variable $X \sim Binomial(n, p)$.

The expected value of $X = np$

The Variance of $X = np(1 - p)$

X can take any integer value between 0 and n.

We can calculate the probability for any value of X.

Estimating p or $logit(p)$

Usually we are not interested in the number of successes but the chance of a success.

If we have n trial with n_1 observed successes and n_0 observed failures then we estimate the probability of a success by $\hat{p}$.

$\hat{p} = \bar{x} = n_1/n \qquad \text{se}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

If we wish to estimate the log odds then

$\hat{\eta} = logit(\hat{p}) = log(n_1) - log(n_0)$

$\text{se}_{\hat{\eta}} = \sqrt{1/n_1 + 1/n_0}$

Testing the value of p

To test a specific value of p or $\eta = logit(p)$ we use a Wald test. Estimate the parameter from the data then plug the hypothesized value into the following.

$z_1 = (\hat{p} - p)/se_{\hat{p}}$ $\qquad$ $z_2 = (\hat{\eta} - \eta)/se_{\hat{\eta}}$

If $n_1 \geq 5$ and $n_0 \geq 5$ both z_1 and z_2 are approximately $N(0, 1)$.

If the sample size is too small then exact methods based on binomial distributions are needed.

Comparing a binary response between 2 groups

The 2 groups can be indicated by a binary variable. So we are really comparing 2 binary variables.

We begin by creating a 2x2 table of the data

	Group 1	Group 2	Total
False	n_{11}	n_{12}	n_{1+}
True	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

The 4 numbers in the table are all we need.

Comparing the numbers

Method 1 is a Pearson χ^2 test.

- should apply a continuity correction.
- requires expected counts ≥ 5

Method 2 is fisher's exact test.

- Method is available in most software
- valid no matter what the counts in each cell are.

Method 3 is a z test for the log odds ratio.

- good for estimation the size of the effect.
- comes up naturally in logistic regression.

Predictors with more than 2 levels

	Group 1	Group 2	Group 3	Total
False	n_{11}	n_{12}	n_{13}	n_{1+}
True	n_{21}	n_{22}	n_{23}	n_{2+}
	n_{+1}	n_{+2}	n_{+3}	n_{++}

Method 1 is a Pearson χ^2 test.

- requires expected counts ≥ 5

Method 2 is fisher's exact test.

- valid no matter what the counts are in each cell.

Example: UC Berkeley Admissions by Gender

Is there gender bias in admission practices at Berkeley?

```
##          Admitted Rejected %Admitted
## Gender
## Male         1198     1493        45
## Female        557     1278        30

## Chisq = 91.6096   DF = 1    Pval = 0.0000
```

Log Odds $= 0.6104 = log(1198) + log(1278) - log(1493) - log(557)$

SE $= 0.06389 = \sqrt{1/1198 + 1/1278 + 1/1493 + 1/557}$

Example: UC Berkeley Admissions by 6 largest Dept.

Do admission rates differ by department?

```
##        Admitted Rejected %Admitted
## Dept
## A           601      332        64
## B           370      215        63
## C           322      596        35
## D           269      523        34
## E           147      437        25
## F            46      668         6

## Chisq = 778.9065   DF = 5    Pval = 0.0000
```

We can calculate individual odds ratios.

There are 15 pairwise odds ratios to consider.

Logistic Regression

With a binary response variable, we model the probability.

$logit(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$

- logit converts a unit interval to the entire real line
- can have continuous or categorical predictors
- can have many predictors including interaction terms

Estimated effects are log odds ratio for a unit change in the predictor.

Most statistical software can do this analysis.

Example: Student Admissions at UC Berkeley

```
##              Admitted Rejected %Admitted
## Dept Gender
## A    Male         512      313        62
##      Female        89       19        82
## B    Male         353      207        63
##      Female        17        8        68
## C    Male         120      205        37
##      Female       202      391        34
## D    Male         138      279        33
##      Female       131      244        35
## E    Male          53      138        28
##      Female        94      299        24
## F    Male          22      351         6
##      Female        24      317         7
## Sum  Male        1198     1493        45
##      Female       557     1278        30
```

Analysis by Gender only

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##        LR Chisq Df Pr(>Chisq)
## Gender    93.45  1     <2e-16

##  OR (M/F)     2.5 %    97.5 %
##      1.84      1.62      2.09
```

There appears to be a gender bias.

Analysis by Dept and Gender

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##        LR Chisq Df Pr(>Chisq)
## Dept      763.4  5     <2e-16
## Gender      1.5  1      0.216

##  OR (M/F)     2.5 %    97.5 %
##      0.90      0.77      1.06
```

There does not appear to be a gender bias.

Analysis of Gender within Dept.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##             LR Chisq Df Pr(>Chisq)
## Dept           855.3  5    < 2e-16
## Dept:Gender     21.7  6    0.00135
```

```
##        OR (M/F) 2.5 % 97.5 %
## DeptA      0.35  0.21   0.58
## DeptB      0.80  0.34   1.89
## DeptC      1.13  0.85   1.50
## DeptD      0.92  0.69   1.24
## DeptE      1.22  0.83   1.81
## DeptF      0.83  0.46   1.51
```

Categorical variables with more than 2 levels

If our response has more than 2 levels then the models are more complicated.

If we have a single categorical predictor we can do Pearson's χ^2 test or Fisher's exact test if some counts in the cross tabulation are small.

```
##             Wine Rating
## Temperature  1  2  3  4  5
##        cold  5 16 13  2  0
##        warm  0  6 13 10  7
```

```
## Fisher's Exact test p-value = 7.366514e-05
```

Cumulative Logistic Regression Models

If the response has k levels then this model subdivides the real line into k pieces that represent the probability of each category on the logit scale.

Predictors in the model shift these cut points to change the probabilities in each category.

How these shifts occur depend on the type of response, ordinal or nominal.

With an ordinal response we can assume some structure like proportional odds.

With Nominal data we make no assumptions about structure so we fit a more general model.

Copenhagen Housing Conditions Survey

Variables are

- Sat - Satisfaction with their present housing circumstances (Low, Medium, High)
- Infl - Perceived influence on the management of the property (Low, Medium, High)
- Type - (Tower, Atrium, Apartment, Terrace)

Satisfaction is the response (ordered factor)

We will predict Satisfaction by Influence and Type

Here is the data

```
##                   Sat Low Medium High
## Infl   Type
## Low    Tower          21     21   28
##        Apartment      61     23   17
##        Atrium         13      9   10
##        Terrace        18      6    7
## Medium Tower          34     22   36
##        Apartment      43     35   40
##        Atrium          8      8   12
##        Terrace        15     13   13
## High   Tower          10     11   36
##        Apartment      26     18   54
##        Atrium          6      7    9
##        Terrace         7      5   11
```

Proportional odds logistic regession model

```
## Threshold Parameters for baseline

##  Low|Medium Medium|High
##  -0.3959673   0.6892151

## Shift parameters for predictors

##                 Estimate Std. Error Pr(>|z|)
## InflMedium     0.4901320  0.1655909   0.0031
## InflHigh       1.1934906  0.1865318   0.0000
## TypeApartment -0.5277651  0.1667091   0.0015
## TypeAtrium    -0.2377054  0.2421692   0.3263
## TypeTerrace   -0.5632012  0.2316138   0.0150
```

Predicted probabilities for proportional odds model

```
##                   Sat  Low Medium High
## Infl   Type
## Low    Tower         40.2   26.4 33.4
##        Apartment     53.3   23.9 22.8
##        Atrium        46.1   25.6 28.4
##        Terrace       54.2   23.6 22.2
## Medium Tower         29.2   25.8 45.0
##        Apartment     41.1   26.3 32.6
##        Atrium        34.3   26.4 39.3
##        Terrace       42.0   26.2 31.8
## High   Tower         16.9   20.7 62.3
##        Apartment     25.7   24.9 49.4
##        Atrium        20.6   22.8 56.6
##        Terrace       26.4   25.1 48.5
```

Nominal logistic regression model

```
## All are threshold parameters
```

```
##                                    Est   PVal
## Low|Medium.(Intercept)    -0.3967147 0.0299
## Medium|High.(Intercept)    0.7000645 0.0001
## Low|Medium.InflMedium     -0.5188023 0.0044
## Medium|High.InflMedium    -0.4707628 0.0160
## Low|Medium.InflHigh       -1.0728042 0.0000
## Medium|High.InflHigh      -1.2567069 0.0000
## Low|Medium.TypeApartment   0.5347458 0.0050
## Medium|High.TypeApartment  0.5172442 0.0051
## Low|Medium.TypeAtrium      0.1309072 0.6436
## Medium|High.TypeAtrium     0.3230569 0.2343
## Low|Medium.TypeTerrace     0.5538032 0.0327
## Medium|High.TypeTerrace    0.5696173 0.0305
```

Predicted probabilities for Nominal model

```
##                   Sat  Low Medium High
## Infl   Type
## Low    Tower         40.2   26.6 33.2
##        Apartment     53.4   23.7 22.8
##        Atrium        43.4   30.2 26.4
##        Terrace       53.9   24.1 21.9
## Medium Tower         28.6   27.1 44.3
##        Apartment     40.6   27.2 32.2
##        Atrium        31.3   32.1 36.5
##        Terrace       41.1   27.9 31.0
## High   Tower         18.7   17.7 63.6
##        Apartment     28.2   20.8 51.0
##        Atrium        20.8   23.4 55.8
##        Terrace       28.6   21.7 49.7
```

Generalized Linear Models

All the models we have fit are generalized linear models.

There is the link function which converts the mean into a linear function of the model parameters. There is a specific variance that weights each observation that is update as the mean changes.

$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$

Dist	Link	Variance	Dist	Link	Variance
Normal	μ	σ^2	Gamma	$1/\mu$	μ^2
Poisson	$\log(\mu)$	μ	NB	$\log(\mu)$	$\mu + \mu^2/r$
Binomial	$\text{logit}(\mu)$	$\mu(1-\mu)$	NB	$\log(\mu)$	$\tau^2\mu$

Questions?

We are changing our name to

Applied Statistics and Data Science Group

- www.stat.ubc.ca/SCARL
- asda.stat.ubc.ca

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research.

Hourly Projects - Professional staff. Fee-for-service consulting.

The End