Regression Models for Risks(Proportions) and Rates

Proportions

E.g. [Changes in] Sex Ratio: Canadian Births in last 60 years

Parameter of Interest: π(male) ... just above 0.5

or π(male)/ π(female) ... "ODDS" ... just above 1.0

or log [π(male)] ... just above ln[0.5]=-0.699

FIRST ANALYSIS : Assuming π is not changing over time

```
data canada;
infile 'Macintosh HD:work:626:sex_can.dat';   /* variables = #in that gender_province */
input
year
      fc fnf fpei fns fnb fq fo fm fs fa fbc
      mc mnf mpei mns mnb mq mo mm ms ma mbc ;  mfc = mc+fc; ratio=mc/fc;

proc means n min max mean sum ; var mc fc mfc;

proc genmod ; model mc/mfc = /         /* if don't put x's in model, "null" model is fitted */
   LINK = identity dist = binomial;    /* "safe" here:- proportions all far from 0 and 1    */

proc genmod ; model mc/mfc = /
   LINK = logit    dist = binomial;    /* canonical link: always "safe" */

proc genmod ; model mc/mfc = /
   LINK = log      dist = binomial;    /* "safe" here:- proportions all well inside (0,1) */

run;
```

```
                        Var   N    Min       Max        Mean          Sum
 ---------------------------------------------------------------------------
 Males (Canada)          MC   60  116 731   246 073   184 449.87   11 066 992
 Females (Canada)        FC   60  110 642   233 202   174 623.97   10 477 438
 Males&Females (Canada) MFC   60  227 575   479 275   359 073.83   21 544 430
 ---------------------------------------------------------------------------
```

```
                The GENMOD Procedure

Model Information

        Description                Value

        Distribution               BINOMIAL
        Link Function              IDENTITY
        Dependent Variable         MC
        Dependent Variable         MFC
        Observations Used          60
        Number Of Events           11066992
        Number Of Trials           21544430

Criteria For Assessing Goodness Of Fit

Criterion           DF       Value   Value/DF

Deviance            59       80.2519  1.3602
Scaled Deviance     59       80.2519  1.3602
Pearson Chi-Square  59       80.2510  1.3602
Scaled Pearson X2   59       80.2510  1.3602
Log Likelihood       .   -14925393.46      .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq     Pr>Chi

INTERCEPT  1  0.5137  0.0001 22756720.9 0.0001
SCALE      0  1.0000  0.0000          .    .

NOTE:  The scale parameter was held fixed.
```

Check: 11066992/21544430 = 0.5137 ; SE

= $\sqrt{p(1-p)/n}$ = $\sqrt{0.5137 x 0.4863/\{21.5\ 10^6\}}$

= *0.0001*

```
                The GENMOD Procedure

Model Information

Value

BINOMIAL
LOGIT
MC
MFC
60
11066992
21544430

Criteria For Assessing Goodness Of Fit

Criterion           DF         Value Value/DF

Deviance            59       80.2519  1.3602
Scaled Deviance     59       80.2519  1.3602
Pearson Chi-Square  59       80.2510  1.3602
Scaled Pearson X2   59       80.2510  1.3602
Log Likelihood       .  -14925393.46       .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq    Pr>Chi

INTERCEPT  1 0.0547   0.0004 16128.7400 0.0001
SCALE      0 1.0000   0.0000          .       .

NOTE:  The scale parameter was held fixed.
```

Check: SEXRATIO = 11.07/10.48 = 1.0563
ln SEXRATIO = 0.0547

SE[0.0547] = $\sqrt{1/\#m + 1/\#f}$ =

$\sqrt{1/\{11.07\ 10^6\} + 1/\{10.48\ 10^6\}}$ = *0.0004*

Not advocating log LINK... out of interest...

```
Distribution       BINOMIAL
Link Function      LOG
Dependent Variable MC
Dependent Variable MFC
Observations Used  60
Number Of Events   11066992
Number Of Trials   21544430

Criteria For Assessing Goodness Of Fit

Criterion   DF         Value Value/DF

Deviance                 59    80.2519    1.3602
Scaled Deviance          59    80.2519    1.3602
Pearson Chi-Square       59    80.2510    1.3602
Scaled Pearson X2        59    80.2510    1.3602
Log Likelihood          .  -14925393.46   .


Analysis Of Parameter Estimates

Parameter    DF    Estimate     Std Err
ChiSquare  Pr>Chi

INTERCEPT 1  -0.6662  0.0002  10098433.3 0.0001
SCALE     0   1.0000  0.0000           .    .

NOTE:  The scale parameter was held fixed.
```

Notes...

- Sample Sizes are very large, and π is close to 0.5, so the Central Limit Theorem holds in all scales!!

- Check that if transform CI's in one scale to those in another, will get practically same CI's as if worked in the target scale to begin with...

  CI[π] = "Inverse Logit" of CI[logit π] etc..

- Notice that the Log likelihood is the same for all 3 models, since we have simply re-parametrized the 1 parameter (π).

- Do not confuse the SEXRatio with ODDSRSATIO

  The SEXRATIO is an ODDS

  If we wish to compare the SexRatio in say the 1930's with the SexRatio in the 1980's then we might wish to work in the ODDSRatio scale

  i.e. $\frac{\pi \text{ male 1980's}}{\pi \text{ female1980'3}} / \frac{\pi \text{ male 1930'3}}{\pi \text{ female 1930'3}}$

  or...

  We might want to model a smooth temporal evolution in π or in log[π**(1-**π**)**].

- With very little evolution over X = calendar year, expect very similar regression fits using identity, logit and log links. Try the identity and logit links and see.

- SexRatio 1+δ <=> π = (1+δ)/(2+δ)   0.5 + δ/4;

  Log[1 + δ]   δ

Regression Models for Risks(Proportions) and Rates

Rates

E.g. 1 [Changes in] Frequency: U.S. Hurricanes in last 110 years

Parameter of Interest: RATE = expected no. of # storms / unit of "country-time"

```
PROC MEANS;

proc genmod ; model n_H_HIT = /
   LINK = identity dist = Poisson;  /* Counts have Poisson variation around mean */

proc genmod ; model n_H_HIT = /
   LINK = identity dist = Normal;   /* Allowing variance of counts to be independent of mean */

proc genmod ; model n_H_HIT = /
   LINK = log dist = Poisson;       /* Change from estimating mean to estimating log[mean] */

proc genmod ; model n_H_HIT = /
   LINK = log dist = Normal;       /* counts have Gaussian variation around exp[β0] */
```

```
   Variable   N      Mean     Std Dev       Min    Max
  --------------------------------------------------
   YEAR      110  1940.50     31.90        1886   1995
   N_STORMS  110     8.52      3.69           1     21
   N_CAT     110     1.80      1.43           0      7
   N_HS_HIT  110     3.34      1.81           0      8
   N_HR_1_9  110     1.74      1.43           0      7
   N_HR_1_5  110     1.43      1.32           0      6
   N_W1      110     0.53      0.73           0      3
   N_W2      110     0.33      0.57           0      3
   N_W3      110     0.40      0.66           0      3
   N_W4      110     0.13      0.37           0      2
   N_W5      110     0.01      0.13           0      1

=> N_H_HIT   110     1.43      1.32           0      6

   N_HR      110     4.95      2.48           0     12
   N_TS      110     3.31      1.99           0      9
   N_SS      110     0.16      0.56           0      4
   ------------------------------------------------
```

```
                 The GENMOD Procedure

Model Information

Description          Value

Data Set             WORK.A
Distribution         POISSON
Link Function        IDENTITY
Dependent Variable   N_H_HIT
Observations Used    110


Criteria For Assessing Goodness Of Fit

Criterion            DF     Value    Value/DF

Deviance            109   149.9483   1.3757
Scaled Deviance     109   149.9483   1.3757
Pearson Chi-Square  109   133.0127   1.2203
Scaled Pearson X2   109   133.0127   1.2203
Log Likelihood        .  -100.7859     .

Analysis Of Parameter Estimates

Parameter   DF Estimate Std Err  ChiSq    Pr>Chi

INTERCEPT   1   1.4364   0.1143 158.0000  0.0001
SCALE       0   1.0000   0.0000     .        .

NOTE:  The scale parameter was held fixed.
```

CHECK:

$\bar{y}$ = 1.4364;

SE[$\bar{y}$] = "Poisson_SD" / $\sqrt{110}$

= $\sqrt{1.4364}$ / $\sqrt{110}$ = *0.1143*

```
                 The GENMOD Procedure

Model Information

Value

WORK.A
NORMAL
IDENTITY
N_H_HIT
110


Criteria For Assessing Goodness Of Fit

Criterion             DF       Value     Value/DF

Deviance             109    191.0545      1.7528
Scaled Deviance      109    110.0000      1.0092
Pearson Chi-Square   109    191.0545      1.7528
Scaled Pearson X2    109    110.0000      1.0092
Log Likelihood         .   -186.4476        .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq    Pr>Chi

INTERCEPT  1  1.4364  0.1257  130.6643 0.0001
SCALE      1  1.3179  0.0889        .        .

Scale estimated by maximum likelihood.
```

CHECK:

$\bar{y}$ = 1.4364;

SE[$\bar{y}$] = "Gaussian_SD" / $\sqrt{110}$

= 1.3179 / $\sqrt{110}$ = *0.1257*

SCALE estimate of SD(counts)

```
             The GENMOD Procedure

Model Information

Description            Value

Data Set              WORK.A
Distribution          POISSON
Link Function         LOG
Dependent Variable    N_H_HIT
Observations Used     110

Criteria For Assessing Goodness Of Fit

Criterion              DF       Value  Value/DF

Deviance              109    149.9483    1.3757
Scaled Deviance       109    149.9483    1.3757
Pearson Chi-Square    109    133.0127    1.2203
Scaled Pearson X2     109    133.0127    1.2203
Log Likelihood          .   -100.7859         .


Analysis Of Parameter Estimates

Parameter   DF Estimate  Std Err  ChiSq Pr>Chi

INTERCEPT    1  0.3621   0.0796  20.7181 0.0001
SCALE        0  1.0000   0.0000     .      .

NOTE:  The scale parameter was held fixed.
```

INTERCEPT = estimate of log[μ], so

$\hat{\mu}_{count}$ = exp[0.3621]= 1.4363.

SCALE=1 => SD(y) not estimated separately.

```
The GENMOD Procedure

Model Information

Value

WORK.A
NORMAL
LOG
N_H_HIT
110

Criteria For Assessing Goodness Of Fit

Criterion              DF       Value  Value/DF

Deviance              109    191.0545    1.7528
Scaled Deviance       109    110.0000    1.0092
Pearson Chi-Square    109    191.0545    1.7528
Scaled Pearson X2     109    110.0000    1.0092
Log Likelihood          .   -186.4476         .


Analysis Of Parameter Estimates

Parameter  DF Estimate Std Err  ChiSq  Pr>Chi

INTERCEPT  1  0.3621   0.0875  17.1336  0.0001
SCALE      1  1.3179   0.0889      .        .

NOTE:  The scale parameter was estimated by
maximum likelihood.
```

SE of *0.0875* for estimate of log[μ] is greater than SE of *0.0796*, since Gaussian-based estimate of SD(counts) is slightly larger than Poisson-based one.

Rates

E.g. 2 Accident Rates before/after Time Change

Parameter of Interest: **RATE DIFFERENCE (RD)**

RATE RATIO (RR)

```
DATA accident;

INPUT    week accidnts   p_t;
LINES;
          0    2575   40000000
          1    2800   40000000
;
proc genmod ; model accidnts = week /   /* additive model=> difference in average COUNT */
   LINK = identity dist = Poisson;

proc genmod ; model accidnts = week /   /* additive model=> difference in average COUNT */
   LINK = identity dist = Normal;       /* but asking that SD's be estimated separately */

proc genmod ; model accidnts = week /   /* multiplicative model= ratio of average COUNTS */
   LINK = log      dist = Poisson;

proc genmod ; model accidnts = week /   /* multiplicative model=> ratio of average COUNTS */
       LINK = log      dist = Normal;       /* but asking that SD's be estimated separately */
```

```
                The GENMOD Procedure

Model Information

Description                     Value

Distribution                    POISSON
Link Function                   IDENTITY
Dependent Variable              ACCIDNTS
Observations Used               2

Criteria For Assessing Goodness Of Fit

Criterion             DF        Value Value/DF

Deviance               0       0.0000        .
Scaled Deviance        0       0.0000        .
Pearson Chi-Square     0       0.0000        .
Scaled Pearson X2      0       0.0000        .
Log Likelihood         .   37072.6815        .

Analysis Of Parameter Estimates

Parameter  DF  Estimate Std Err  ChiSq Pr>Chi

INTERCEPT  1 2575.0000 50.7445 2575.0000 0.0001
WEEK       1  225.0000 73.3144    9.4186 0.0021
SCALE      0    1.0000  0.0000         .

NOTE:  The scale parameter was held fixed.
```

NOTE: Fitting saturated model. Scale will not be estimated.
NOTE: The scale parameter was held fixed.

$\hat{\beta}$ for "week" = c1 - c0 = 2800 - 2575 = 225;

SE(diff) = $\sqrt{var(c1) + var(c0)}$

= $\sqrt{2575 + 2800} = \sqrt{5375} = 73.31$

```
                The GENMOD Procedure

Model Information

Value

NORMAL
IDENTITY
ACCIDNTS
2

Criteria For Assessing Goodness Of Fit

Criterion            DF     Value  Value/DF

Deviance              0    0.0000         .
Scaled Deviance       0    0.0000         .
Pearson Chi-Square    0    0.0000         .
Scaled Pearson X2     0    0.0000         .
Log Likelihood        .   -1.8379         .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq  Pr>Chi

INTERCEPT 1 2575.0000 1.0000 6630625.00 0.0001
WEEK      1  225.0000 1.4142 25312.5000 0.0001
SCALE     0    1.0000 0.0000     .         .

NOTE:  The scale parameter was held fixed.
```

NOTE: Fitting saturated model. Scale will not be estimated.
NOTE: The scale parameter was held fixed.

SE(diff)= $\sqrt{var(c1)+var(c0)} = \sqrt{1 + 1} = 1.4142$!!

GENMOD GOT CONFUSED : **0** df to estimate SD(Y|X)!

```
              The GENMOD Procedure

Model Information

Description                     Value

Distribution                    POISSON
Link Function                   LOG
Dependent Variable              ACCIDNTS
Observations Used               2

Criteria For Assessing Goodness Of Fit

Criterion           DF        Value Value/DF

Deviance             0       0.0000        .
Scaled Deviance      0       0.0000        .
Pearson Chi-Square   0       0.0000        .
Scaled Pearson X2    0       0.0000        .
Log Likelihood       .   37072.6815        .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq      Pr>Chi

INTERCEPT  1  7.8536   0.0197 158823.705 0.0001
WEEK       1  0.0838   0.0273     9.4131  0.0022
SCALE      0  1.0000   0.0000      .        .

NOTE:  The scale parameter was held fixed.
```

```
NOTE: Fitting saturated model. Scale will not be estimated.
NOTE:  The scale parameter was held fixed.
```

0.0838 = estimate of $\log[\mu_1]$ - $\log[\mu_0]$

=> estimate of μ_1/μ_0 = exp[0.0838]= 1.09

SE**[est of log RR]** = $\sqrt{1/c1 + 1/c0}$

= $\sqrt{1/2575 + 1/2800}$ = ***0.0273***

```
              The GENMOD Procedure

Model Information

Value

NORMAL
LOG
ACCIDNTS
2

Criteria For Assessing Goodness Of Fit

Criterion           DF        Value   Value/DF

Deviance             0       0.0000          .
Scaled Deviance      0       0.0000     .
Pearson Chi-Square   0       0.0000     .
Scaled Pearson X2    0       0.0000     .
Log Likelihood       .      -1.8379     .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err  ChiSq   Pr>Chi

INTERCEPT 1   7.8536  0.0004  408971039  0.0001
WEEK      1   0.0838  0.0005 25209.1997  0.0001
SCALE     0   1.0000  0.0000          .       .

NOTE:  The scale parameter was held fixed.
```

```
NOTE: Fitting saturated model. Scale will not be estimated.
NOTE:  The scale parameter was held fixed.
```

"SE" = 0.0005: GENMOD GOT CONFUSED again :
0 df for SD(Y|X)! it used SD(Y|X) = 1.

Rates

E.g. 2 Accident Rates before/after Time Change ... continued

Parameter of Interest: RATE DIFFERENCE (RD)

RATE RATIO (RR)

The above analyses modeled the average counts & differences/ratios in average counts; we need to refine these to estimate RATES and difference/ratio of RATES

Key :

μ = average (or expected) count = RATE x Denominator

or , other way around ...

RATE = μ_{count} / Denominator = function of regressor variables

Examples...

1. RATE = μ_{count} **/ Denominator =** $\beta_0 + \beta_1 X$ ***(additive model)***

--> **average(count) =** $\{ \beta_0 + \beta_1 X \}$ **Denominator =** β_0 **Denominator +** β_1 $\{X \times \text{Denominator}\}$

--> **Fit count as "Y" variable, with IDENTITY LINK, to the regression equation**

$$\beta_0 Z_0 + \beta_1 Z_1 ,$$

where Z_0 **= Denominator ,** $Z_2 = X \times \text{Denominator}$

i.e. MODEL Y = Z_0 Z_1 ***/ LINK = IDENTITY DIST = POISSON NOINT; (no intercept)***

---

2. RATE = μ_{count} **/ Denominator = exp [** $\beta_0 + \beta_1 X$ **]** ***(multiplicative model)***

--> **average(count)** $= \{ \exp [\beta_0 + \beta_1 X] \} \times \text{Denominator}$

$= \exp [\beta_0 + \beta_1 X + 1. \log[\text{Denominator}]]$ *{bringing Denominator inside the exp}*

--> **Fit count as "Y" variable, with LOG LINK, to the regression equation**

$$\beta_0 + \beta_1 X + 1. \log[\text{Denominator}]$$

i.e. MODEL Y = X / LINK = LOG DIST = POISSON OFFSET = Z ;

Z = log[Denominator] must be defined in dataset (doesn't have to be called Z)

An "OFFSET"in a generalized linear model is a term whose accompanying regression coefficient is not estimated, but rather is forced to be 1.

```
DATA accident;

INPUT week accidnts p_t;

p_t = p_t/1000000; /* p_t unit = 1 million person days  */

z0 = p_t; z1 = week*p_t;

z = log(p_t);

LINES;
        0  2575   40000000
        1  2800   40000000
;

proc genmod ; model accidnts = z0 z1 /
   LINK = identity dist = Poisson  NOINT;

         /* additive model to estimate rate difference */

proc genmod ; model accidnts = week /
   LINK = log      dist = Poisson  OFFSET = Z;

/* multiplicative model to estimate rate ratio */


run;
```

```
                 The GENMOD Procedure

Model Information

Description           Value

Distribution          POISSON
Link Function         IDENTITY ->Rate diff (RD)
Dependent Variable    ACCIDNTS
Observations Used     2

Criteria For Assessing Goodness Of Fit

Criterion           DF       Value Value/DF

Deviance             0     -0.0000        .
Scaled Deviance      0     -0.0000        .
Pearson Chi-Square   0      0.0000        .
Scaled Pearson X2    0      0.0000        .
Log Likelihood       .  37072.6815        .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSquare Pr>Chi

INTERCEPT 0  0.0000  0.0000          .       .
Z0        1 64.3750  1.2686   2575.0000  0.0001
Z1        1  5.6250  1.8329      9.4186  0.0021
SCALE     0  1.0000  0.0000         .       .

NOTE:  The scale parameter was held fixed.
```

NB Rates (and rate diffs) are per million person_days.

i.e. 64.37   =2575/40 million p_d and
       5.6250 =rate diff = 225/40 million p_d

(If the p_t is measured as person days, the coefficients have too may leading zeroes)

```
                 The GENMOD Procedure

Model Information

Description                     Value

Distribution                    POISSON
Link Function                   LOG -> RR
Dependent Variable              ACCIDNTS
Offset Variable                 Z
Observations Used               2

Criteria For Assessing Goodness Of Fit

Criterion           DF       Value Value/DF

Deviance             0      0.0000        .
Scaled Deviance      0      0.0000        .
Pearson Chi-Square   0      0.0000        .
Scaled Pearson X2    0      0.0000        .
Log Likelihood       .  37072.6815        .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err ChiSq  Pr>Chi

INTERCEPT  1  4.1647  0.0197 44663.2136  0.0001
WEEK       1  0.0838  0.0273  9.4131  0.0022
SCALE      0  1.0000  0.0000   .       .

NOTE:  The scale parameter was held fixed.
```

0.0838 is the log of the estimated RATE RATIO, as above,

so estimated RATERATIO = exp[0.0838] = 1.09.

...

Regression Models for Risks(Proportions) and Rates

Rates

E.g. 3: Cancer Rates in Exposed and Unexposed Cohorts

c_1 = 41 in n_1 = 28,010 person years vs. c_0 = 15 in n_0 = 19,017 person years

(Rothman & Boice e.g.)

Denominators (Person years) are *<u>different</u>* *in* Exposed and Unexposed Cohorts

<u>Parameters</u> of Interest: RATE DIFFERENCE (RD)

RATE RATIO (RR)

Again :

μ = average (or expected) count = RATE x Denominator

or , other way around ...

RATE = μ_{count} / Denominator = function of regressor variables

```
DATA R_and_B; /* Rothman and Boice data */

INPUT exposure  cases  p_t;

z0 = p_t; z1 = exposure*p_t;

z = log(p_t);

LINES;
         1       41   28010
         0       15   19017
;

proc genmod ; model cases = z0 z1 /
   LINK = identity dist = Poisson  NOINT;

           /* additive model to estimate rate difference */

proc genmod ; model cases = exposure /
   LINK = log      dist = Poisson  OFFSET = Z;

          /* multiplicative model to estimate rate ratio */

run;
```

```
                                 The GENMOD Procedure
Model Information

Description                 Value
Distribution                POISSON
Link Function               IDENTITY ->Rate diff (RD)
Dependent Variable          CASES
Observations Used           2

Criteria For Assessing Goodness Of Fit
Criterion                   DF        Value Value/DF

Deviance                     0       0.0000      .
Scaled Deviance              0       0.0000      .
Pearson Chi-Square           0       0.0000      .
Scaled Pearson X2            0       0.0000      .
Log Likelihood               .     136.8772      .

Analysis Of Parameter Estimates

Parameter DF Estimate Std Err  ChiSq  Pr>Chi

INTERCEPT 0  0.0000   0.0000    .       .
Z0        1  0.0008   0.0002  15.0000  0.0001
Z1        1  0.0007   0.0003   4.8607  0.0275
SCALE     0  1.0000   0.0000    .       .

NOTE:  The scale parameter was held fixed.
```

The coefficient b_1 = 0.0007 accompanying Z1 is the estimated RATE DIFFERENCE.

Check: 41/28010 - 15/19017 = 0.00067 or an excess of 67 cases per 100000 p_y.

Check also that the SE of *0.0003* lines up with

SE(41/28010 - 15/19017)

$= \sqrt{41/28010^2 + 15/19017^2}$ in earlier exercise.

```
Model Information

Description        Value
Distribution         POISSON
Link Function        LOG -> RATERATIO
Dependent Variable   CASES
Offset Variable      Z
Observations Used    2

Criteria For Assessing Goodness Of Fit
Criterion                DF        Value   Value/DF

Deviance                  0       0.0000          .
Scaled Deviance           0       0.0000          .
Pearson Chi-Square        0       0.0000          .
Scaled Pearson X2         0       0.0000          .
Log Likelihood            .     136.8772          .

Analysis Of Parameter Estimates

Parameter  DF Estimate Std Err ChiSq Pr>Chi

INTERCEPT  1 -7.1450   0.2582 765.7736 0.0001
EXPOSURE   1  0.6183   0.3018   4.1983 0.0405
SCALE      0  1.0000   0.0000        .      .

NOTE:  The scale parameter was held fixed.
```

0.6183 is the log of the estimated RATE RATIO,

so estimated RATERATIO = exp[0.6183] = 1.86.

Check: RR = (41/28010 /(15/19017) = 1.85

Also check that SE = *0.3018 agrees with*

$\sqrt{1/41 + 1/15}$ for the log of an estimate of a rate ratio