Poisson Regression

Not in the text, but it's another generalized linear model

Poisson Process

- Events happening randomly in space or time
- Independent increments
- For a small region or interval,
  - Chance of 2 or more events is negligible
  - Chance of an event roughly proportional to the size of the region or interval
- Then (solve a system of differential equations), the probability of observing *x* events in a region of size *t* is

$$\frac{e^{-\lambda t}(\lambda t)^x}{x!} \text{ for } x = 0, 1, \ldots$$

Regression: Outcomes are Counts

- Poisson process model roughly applies
- Examples: Relationship of explanatory variables to
  - Number of children
  - Number of typos in a short document
  - Number of workplace accidents in a short time period
  - Number of marriages
- For large λ a normality assumption is okay, but not constant variance

Linear Model for log λ

- $\log \lambda = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$
- Implicitly for $i = 1, \ldots N$
- Everybody in the sample has a different $\lambda = \lambda_i$
- Take exponential function of both sides
- Substitute into Poisson likelihood
- Maximum likelihood as usual
- Likelihood ratio tests, Wald tests, etc.

$$\log \lambda = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

- Increase x_k with everything else held constant, and
  - Log λ increases by β_k
  - λ is multiplied by $e^{\beta k}$

Back to the job study: N=200 Students

- 106 employed in a job related to field of study
- 74 employed in a job unrelated to field of study
- 20 unemployed
- Could be independent Poisson processes

- Conditionally on the total number of students, multinomial with
  - $p_1 = \lambda_1/(\lambda_1+\lambda_2+\lambda_3)$
  - $p_2 = \lambda_2/(\lambda_1+\lambda_2+\lambda_3)$
  - $p_3 = \lambda_3/(\lambda_1+\lambda_2+\lambda_3)$

Poisson regression with dummy variables

Job Status	d_1	d_2	$\log \lambda = \beta_0 + \beta_1 d_1 + \beta_2 d_2$
Related	0	0	β_0
Unrelated	1	0	$\beta_0 + \beta_1$
Unemployed	0	1	$\beta_0 + \beta_2$

On average, we expect $e^{\beta 2}$ times as many unemployed students as students with jobs related to their fields of study.

The senseless Null Hypothesis

H_0: $p_1 = p_2 = p_3$ if and only if

$\lambda_1 = \lambda_2 = \lambda_3$ if and only if

$\beta_0 = \beta_0+\beta_1 = \beta_0+\beta_2$ if and only if

$\beta_1 = \beta_2 = 0$

Tested first hypothesis directly, got

$G^2 = 65.6$, df=2

```
> jobz = read.table(stdin()) # Read from standard input
0:      Job         Freq
1: 1   Related      106
2: 2   Unrelated     74
3: 3   Unemployed    20
4:
> # End with Ctrl-D on Unix (Mac) or Ctrl-Z on Windows
> jobz
         Job Freq
1    Related  106
2  Unrelated   74
3 Unemployed   20
> freq = jobz$Freq
> job = factor(jobz$Job)
> full0 = glm(freq~job,family=poisson) # Saturated
```

```
> summary(full0)

Call:
glm(formula = freq ~ job, family = poisson)

Deviance Residuals:
[1]  0  0  0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.66344    0.09713  48.013  < 2e-16 ***
jobUnemployed -1.66771    0.24379  -6.841 7.88e-12 ***
jobUnrelated  -0.35937    0.15148  -2.372   0.0177 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  6.5598e+01  on 2  degrees of freedom
Residual deviance: -7.9936e-15  on 0  degrees of freedom
AIC: 23.489

Number of Fisher Scoring iterations: 3

> full0$null.deviance
[1] 65.59798
```

Better H_0

$$
\begin{aligned}
H_0 \quad & p_1 = 2p_2 \\
\Leftrightarrow \quad & \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 2\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \\
\Leftrightarrow \quad & \lambda_1 = 2\lambda_2 \\
\Leftrightarrow \quad & \log \lambda_1 = \log 2 + \log \lambda_2 \\
\Leftrightarrow \quad & \beta_0 = \log 2 + \beta_0 + \beta_1 \\
\Leftrightarrow \quad & \beta_1 = -\log 2
\end{aligned}
$$

$G^2 = 4.739$, df=1

$G^2 = 4.739$, df=1

```
# Offset "can be used to specify an a priori known component
# to be included in the linear predictor during fitting. This should
# be NULL or a numeric vector of length either one or equal to the
# number of cases."
> freq
[1] 106  74  20
> d1 = c(0,1,0)
> d2 = c(0,0,1)
> red0 = glm(freq ~ d2, offset=-log(2)*d1,family=poisson)
> summary(red0)
```

```
> summary(red0)
```

$G^2 = 4.739$, df=1

```
Call:
glm(formula = freq ~ d2, family = poisson, offset = -log(2) *
    d1)

Deviance Residuals:
     1       2       3
-1.304   1.743   0.000

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.78749    0.07454  64.231  < 2e-16 ***
d2          -1.79176    0.23570  -7.602 2.92e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 95.2551  on 2  degrees of freedom
Residual deviance:  4.7395  on 1  degrees of freedom
AIC: 26.229

Number of Fisher Scoring iterations: 4
```