Poisson Regression

Contents

1 Introduction

Introduction

In this lecture we discuss the Poisson regression model and some applications.

Poisson regression deals with situations in which the dependent variable is a count. In our earlier discussion of the Poisson distribution, we mentioned that it is a limiting case of the binomial distribution when the number of trials becomes large while the expectation remains stable, i.e., the probability of success is very small.

An important additional property of the Poisson distribution is that sums of independent Poisson variates are themselves Poisson variates, i.e., if Y_1 and Y_2 are independent with Y_i having a $P(\mu_i)$ distribution, then

$$Y_1 + Y_2 \; P(\mu_1 + \mu_2) \tag{1}$$

As we shall see, the key implication of this result is that individual and grouped data can both be analyzed with the Poisson distribution.

2 An Introductory Example

An Introductory Example

On his superb website at *data.princeton.edu*(which I strongly recommend as a source for reading and examples), Germán Rodríguez presents an introductory example involving data from the World Fertility Study.

The Children Ever Born (ceb) Data

The dataset has 70 rows representing grouped individual data. Each row has entries for:

- The cell number (1 to 71, cell 68 has no observations)

- Marriage duration (1=0–4, 2=5–9, 3=10–14, 4=15–19, 5=20–24, 6=25–29)
- Residence (1=Suva, 2=Urban, 3=Rural)
- Education (1=none, 2=lower primary, 3=upper primary, 4=secondary+)
- Mean number of children ever born (e.g. 0.50)
- Variance of children ever born (e.g. 1.14)
- Number of women in the cell (e.g. 8)

Reference: Little, R. J. A. (1978). Generalized Linear Models for Cross-Classified Data from the WFS. *World Fertility Survey Technical Bulletins, Number 5*.

An Introductory Example

A tabular presentation shows data on the number of children ever born to married Indian women classified by duration since their first marriage (grouped in six categories), type of place of residence (Suva, other urban and rural), and educational level (classified in four categories: none, lower primary, upper primary, and secondary or higher). Each cell in the table shows the mean, the variance and the number of observations.

Introductory Example

TABLE 4.1: Number of Children Ever Born to Women of Indian Race
By Marital Duration, Type of Place of Residence and Educational Level
(Each cell shows the mean, variance and sample size)

Marr.	Suva				Urban				Rural			
Dur.	N	LP	UP	S+	N	LP	UP	S+	N	LP	UP	S+
0–4	0.50	1.14	0.90	0.73	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	1.14	0.73	0.67	0.48	1.06	1.59	0.73	0.54	0.88	0.81	0.80	0.59
	8	21	42	51	12	27	39	51	62	102	107	47
5–9	3.10	2.67	2.04	1.73	4.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	1.66	0.99	1.87	0.68	3.44	1.51	0.97	0.81	1.93	1.36	1.30	1.19
	10	30	24	22	13	37	44	21	70	117	81	21
10–14	4.08	3.67	2.90	2.00	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	1.72	2.31	1.57	1.82	2.97	2.99	1.96	1.52	3.52	3.31	3.28	2.50
	12	27	20	12	18	43	29	15	88	132	50	9
15–19	4.21	4.94	3.15	2.75	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	2.03	1.46	0.81	0.92	7.40	2.97	3.83	0.70	4.91	3.23	3.29	–
	14	31	13	4	23	42	20	5	114	86	30	1
20–24	5.62	5.06	3.92	2.60	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	4.15	4.64	4.08	4.30	7.19	4.44	4.33	0.33	8.20	5.72	5.20	0.50
	21	18	12	5	22	25	13	3	117	68	23	2
25–29	6.60	6.74	5.38	2.00	6.52	7.51	7.54	–	7.48	7.81	5.80	–
	12.40	11.66	4.27	–	11.45	10.53	12.60	–	11.34	7.57	7.07	–
	47	27	8	1	46	45	13	–	195	59	10	–

Introductory Example

The unit of analysis is the individual woman, the response variable is the number of children given birth to, and the potential predictor variables are

1. Duration since her first marriage
2. Type of place where she resides
3. Her educational level, classified in four categories.

3 The Poisson Regression Model

The Poisson Regression Model

The Poisson regression model assumes that the sample of n observations y_i are observations on independent Poisson variables Y_i with mean μ_i.

Note that, if this model is correct, the equal variance assumption of classic linear regression is violated, since the Y_i have means equal to their variances.

So we fit the generalized linear model,

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{2}$$

We say that the Poisson regression model is a generalized linear model with Poisson error and a log link.

The Poisson Regression Model

An alternative version of Equation 2 is

$$\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{3}$$

This implies that one unit increases in an x_j are associated with a *multiplication* of μ_j by $\exp(\beta_j)$.

Grouped Data and the Offset

Note that the model of Equation 2 refers to individual observations, but the table gives summary measures. Do we need the individual observations to proceed? No, because, as Germán Rodríguez explains very clearly in his lecture notes, we can apply the result of Equation 1.

Grouped Data and the Offset

Specifically, define Y_{ijkl} to be the number of children borne by the l-th woman in the (i, j, k)-th group, where i denotes marital duration, j residence and k education. Let $Y_{ijk\bullet} = \sum_l Y_{ijkl}$ be the group total shown in the table. Then if each of the observations in this group is a realization of an independent Poisson variate with mean μ_{ijk}, then the group total will be a realization of a Poisson variate with mean $n_{ijk}\mu_{ijk}$, where n_{ijk} is the number of observations in the (i, j, k)-th cell.

Grouped Data and the Offset

Suppose now that you postulate a log-linear model for the individual means, say

$$\log(\mu_{ijkl}) = \log E(Y_{ijkl}) = \boldsymbol{x}_{ijk}\boldsymbol{\beta} \tag{4}$$

Then the log of the expected value of the group total is

$$\begin{aligned} \log(E(Y_{ijk})) &= \log(n_{ijk}\mu_{ijk}) &(5)\\ &= \log(n_{ijk}) + \boldsymbol{x}'_{ijk}\boldsymbol{\beta} &(6)\end{aligned}$$

Grouped Data and the Offset

Thus, the group totals follow a log-linear model with exactly the same coefficients $\boldsymbol{\beta}$ as the individual means, except for the fact that the linear predictor includes the term $\log(n_{ijk})$. This term is referred to as the *offset*. Often, when the response is a count of events, the offset represents the log of some measure of exposure, in this case the number of women.

4 Testing Models of the Fertility Data

Simple One-Variable Models

Let's consider some models for predicting the fertility data from our potential predictors. Our first 4 models are:

1. The null model, including only an intercept.
2. A model predicting number of children from Duration (D).
3. A model predicting number of children from Residence (R).
4. A model predicting number of children from Education (E).

To fit the models with Poisson regression, we use the `glm` package, specifying a `poisson` family (the log link is the default).

Simple One-Variable Models

Here we fit simple models that predict number of children from duration, region of residence, and education. Let's begin by looking carefully at a model that predicts number of children solely from the duration of their childbearing years.]

```
> ceb.data ← read.table("ceb.dat",header=T)
> fit.D ← glm(y~dur,family="poisson",
+     offset=log(n),data=ceb.data)
> fit.E ← glm(y~educ,family="poisson",
+     offset=log(n),data=ceb.data)
> fit.R ← glm(y~res,family="poisson",
+     offset=log(n),data=ceb.data)
```

Note that, in order to fit the model correctly, we had to specify `family ="poisson"` and `offset=log(n)`.

Predicting Children Ever Born from Duration

The `dur` variable is categorical, so R automatically codes its 6 categories into 5 variables. Each of these variables takes on a value of 1 for its respective category. The first category, `00-04`, and has no variable representing it. Consequently, it is the "reference category" and has a score of zero. All the other categories are represented by dummy predictor variables that take on the value 1 if `dur` has that category—otherwise the dummy variable has a code of zero.

Predicting Children Ever Born from Duration

Let's look at some output:

```
> summary(fit.D)

Call:
glm(formula = y ~ dur, family = "poisson", data = ceb.data, offset = log(n))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5626  -1.4608  -0.5515   0.6060   4.0093

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10413    0.04416  -2.358   0.0184 *
dur05-09     1.04556    0.05241  19.951   <2e-16 ***
dur10-14     1.44605    0.05025  28.779   <2e-16 ***
dur15-19     1.70719    0.04976  34.310   <2e-16 ***
dur20-24     1.87801    0.04966  37.818   <2e-16 ***
dur25-29     2.07923    0.04752  43.756   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.52  on 69  degrees of freedom
Residual deviance:  165.84  on 64  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 4
```

Predicting Children Ever Born from Duration

Consider a woman whose first marriage was in the last 0–4 years. On average, such women have $\exp(-0.1) = 0.9$ children.

Consider, on the other hand, a woman whose duration is 20–24 years. Such women have, on average $\exp(-0.1 + 1.71) = 4.97$ children.

Predicting Children Ever Born from Education

Next, let's look at education alone as a predictor.

```
> summary(fit.E)


Call:
glm(formula = y ~ educ, family = "poisson", data = ceb.data,
    offset = log(n))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-19.2952   -3.0804    0.7426    3.8574   13.1418

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.43567    0.01594  90.090   <2e-16 ***
educnone     0.21154    0.02168   9.759   <2e-16 ***
educsec+    -1.01234    0.05176 -19.557   <2e-16 ***
educupper   -0.40473    0.02951 -13.714   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.5  on 69  degrees of freedom
Residual deviance: 2661.0  on 66  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 5
```

Predicting Children Ever Born from Education

With 4 education categories, we need 3 dummy variables. Which category is the "reference" category in this case?

Consider a woman whose education was "lower" Such women have, on average, $\exp(1.44) = 4.2$ children.

Consider, on the other hand, a woman whose educational level is postsecondary. Such women have, on average, $\exp(1.44 + -0.4) = 2.8$ children.

Now — *You* Try It!

Examine the model predicting number of children solely from place of residence. What is the reference category?

What is the average number of children ever born for women in the reference category?

Two-Factor Additive Models

Next we add education as a predictor to duration. The `anova` function helps us to see that there is a significant improvement.

```
> fit.NULL <-glm(y~1,family="poisson",
+    offset=log(n),data=ceb.data)
> fit.D.E <- glm(y~dur+educ,family="poisson",
+    offset=log(n),data=ceb.data)
> anova(fit.NULL,fit.D,fit.D.E)
```

```
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ dur
Model 3: y ~ dur + educ
  Resid. Df Resid. Dev Df Deviance
1        69     3731.5
2        64      165.8  5   3565.7
3        61      100.0  3     65.8
```

Three-Factor Additive Model

Next we add residence to duration and education.

```
> fit.D.E.R <- glm(y~dur+educ+res,
+  family="poisson",offset=log(n),data=ceb.data)
> anova(fit.NULL,fit.D,fit.D.E,fit.D.E.R)
```

```
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ dur
Model 3: y ~ dur + educ
Model 4: y ~ dur + educ + res
  Resid. Df Resid. Dev Df Deviance
1        69     3731.5
2        64      165.8  5   3565.7
3        61      100.0  3     65.8
4        59       70.7  2     29.4
```

Three-Factor Additive Model

```
> summary(fit.D.E.R)
```

```
Call:
glm(formula = y ~ dur + educ + res, family = "poisson", data = ceb.data,
    offset = log(n))
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.29124  -0.66487   0.07588   0.66062   3.67903

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.05695    0.04805   1.185    0.236
dur05-09     0.99765    0.05275  18.912  < 2e-16 ***
dur10-14     1.37053    0.05108  26.833  < 2e-16 ***
dur15-19     1.61423    0.05121  31.524  < 2e-16 ***
dur20-24     1.78549    0.05122  34.856  < 2e-16 ***
dur25-29     1.97679    0.05005  39.500  < 2e-16 ***
educnone    -0.02308    0.02266  -1.019    0.308
educsec+    -0.33266    0.05388  -6.174 6.67e-10 ***
educupper   -0.12475    0.03000  -4.158 3.21e-05 ***
resSuva     -0.15122    0.02833  -5.338 9.37e-08 ***
resurban    -0.03896    0.02462  -1.582    0.114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3731.525  on 69  degrees of freedom
Residual deviance:   70.653  on 59  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 4
```

Three-Factor Additive Model

What is the predicted average number of children for women married 5–9 years, living in Suva, with post-secondary education?