Chapter 4

Poisson Models for Count Data

In this chapter we study log-linear models for count data under the assumption of a Poisson error structure. These models have many applications, not only to the analysis of counts of events, but also in the context of models for contingency tables and the analysis of survival data.

4.1 Introduction to Poisson Regression

As usual, we start by introducing an example that will serve to illustrative regression models for count data. We then introduce the Poisson distribution and discuss the rationale for modeling the logarithm of the mean as a linear function of observed covariates. The result is a generalized linear model with Poisson response and link log.

4.1.1 The Children Ever Born Data

Table 4.1, adapted from Little (1978), comes from the Fiji Fertility Survey and is typical of the sort of table published in the reports of the World Fertility Survey. The table shows data on the number of children ever born to married women of the Indian race 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.

In our analysis of these data we will treat the number of children ever

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	–

born to each woman as the response, and her marriage duration, type of place of residence and level of education as three discrete predictors or factors.

4.1.2 The Poisson Distribution

A random variable Y is said to have a Poisson distribution with parameter μ if it takes integer values $y = 0, 1, 2, \ldots$ with probability

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!} \tag{4.1}$$

for $\mu > 0$. The mean and variance of this distribution can be shown to be

$$E(Y) = \text{var}(Y) = \mu.$$

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data.

The classic text on probability theory by Feller (1957) includes a number of examples of observations fitting the Poisson distribution, including data on the number of flying-bomb hits in the south of London during World War II. The city was divided into 576 small areas of one-quarter square kilometers each, and the number of areas hit exactly k times was counted. There were a total of 537 hits, so the average number of hits per area was 0.9323. The observed frequencies in Table 4.2 are remarkably close to a Poisson distribution with mean $\mu = 0.9323$. Other examples of events that fit this distribution are radioactive disintegrations, chromosome interchanges in cells, the number of telephone connections to a wrong number, and the number of bacteria in different areas of a Petri plate.

TABLE 4.2: Flying-bomb Hits on London During World War II

Hits	0	1	2	3	4	5+
Observed	229	211	93	35	7	1
Expected	226.7	211.4	98.6	30.6	7.1	1.6

The Poisson distribution can be derived as a limiting form of the binomial distribution if you consider the distribution of the number of successes in a very large number of Bernoulli trials with a small probability of success in each trial. Specifically, if $Y \sim B(n, \pi)$ then the distribution of Y as $n \to \infty$ and $\pi \to 0$ with $\mu = n\pi$ remaining fixed approaches a Poisson distribution with mean μ. Thus, the Poisson distribution provides an approximation to the binomial for the analysis of rare events, where π is small and n is large.

In the flying-bomb example, we can think of each day as one of a large number of trials where each specific area has only a small probability of being hit. Assuming independence across days would lead to a binomial distribution which is well approximated by the Poisson.

An alternative derivation of the Poisson distribution is in terms of a stochastic process described somewhat informally as follows. Suppose events occur randomly in time in such a way that the following conditions obtain:

- The probability of at least one occurrence of the event in a given time interval is proportional to the length of the interval.
- The probability of two or more occurrences of the event in a very small time interval is negligible.
- The numbers of occurrences of the event in disjoint time intervals are mutually independent.

Then the probability distribution of the number of occurrences of the event in a fixed time interval is Poisson with mean $\mu = \lambda t$, where λ is the rate of occurrence of the event per unit of time and t is the length of the time interval. A process satisfying the three assumptions listed above is called a Poisson process.

In the flying bomb example these conditions are not unreasonable. The longer the war lasts, the greater the chance that a given area will be hit at least once. Also, the probability that the same area will be hit twice the same day is, fortunately, very small. Perhaps less obviously, whether an area is hit on any given day is independent of what happens in neighboring areas, contradicting a common belief that bomb hits tend to cluster.

The most important motivation for the Poisson distribution from the point of view of statistical estimation, however, lies in the relationship between the mean and the variance. We will stress this point when we discuss our example, where the assumptions of a limiting binomial or a Poisson process are not particularly realistic, but the Poisson model captures very well the fact that, as is often the case with count data, the variance tends to increase with the mean.

A useful property of the Poisson distribution is that the sum of independent Poisson random variables is also Poisson. Specifically, if Y_1 and Y_2 are independent with $Y_i \sim P(\mu_i)$ for $i = 1, 2$ then

$$Y_1 + Y_2 \sim P(\mu_1 + \mu_2).$$

This result generalizes in an obvious way to the sum of more than two Poisson observations.

An important practical consequence of this result is that we can analyze individual or grouped data with equivalent results. Specifically, suppose we have a group of n_i individuals with identical covariate values. Let Y_{ij} denote the number of events experienced by the j-th unit in the i-th group, and let Y_i denote the total number of events in group i. Then, under the usual assumption of independence, if $Y_{ij} \sim P(\mu_i)$ for $j = 1, 2, \ldots, n_i$, then $Y_i \sim P(n_i\mu_i)$. In words, if the individual counts Y_{ij} are Poisson with mean μ_i, the group total Y_i is Poisson with mean $n_i\mu_i$. In terms of estimation, we obtain exactly the same likelihood function if we work with the individual counts Y_{ij} or the group counts Y_i.

4.1.3 Log-Linear Models

Suppose that we have a sample of n observations $y_1, y_2, \ldots, y_n$ which can be treated as realizations of independent Poisson random variables, with

$Y_i \sim P(\mu_i)$, and suppose that we want to let the mean μ_i (and therefore the variance!) depend on a vector of explanatory variables $\mathbf{x}_i$.

We could entertain a simple linear model of the form

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta},$$

but this model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represents an expected count, has to be non-negative.

A straightforward solution to this problem is to model instead the *logarithm* of the mean using a linear model. Thus, we take logs calculating $\eta_i = \log(\mu_i)$ and assume that the transformed mean follows a linear model $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Thus, we consider a generalized linear model with link log. Combining these two steps in one we can write the log-linear model as

$$\log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}. \tag{4.2}$$

In this model the regression coefficient β_j represents the expected change in the *log* of the mean per unit change in the predictor x_j. In other words increasing x_j by one unit is associated with an increase of β_j in the log of the mean.

Exponentiating Equation 4.2 we obtain a multiplicative model for the mean itself:

$$\mu_i = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}.$$

In this model, an exponentiated regression coefficient $\exp\{\beta_j\}$ represents a multiplicative effect of the j-th predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor $\exp\{\beta_j\}$.

A further advantage of using the log link stems from the empirical observation that with count data the effects of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to a much simpler model.

4.2 Estimation and Testing

The log-linear Poisson model described in the previous section is a generalized linear model with Poisson error and link log. Maximum likelihood estimation and testing follows immediately from the general results in Appendix B. In this section we review a few key results.

4.2.1 Maximum Likelihood Estimation

The likelihood function for n independent Poisson observations is a product of probabilities given by Equation 4.1. Taking logs and ignoring a constant involving $\log(y_i!)$, we find that the log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum\{y_i \log(\mu_i) - \mu_i\},$$

where μ_i depends on the covariates $\mathbf{x}_i$ and a vector of p parameters $\boldsymbol{\beta}$ through the log link of Equation 4.2.

It is interesting to note that the log is the canonical link for the Poisson distribution. Taking derivatives of the log-likelihood function with respect to the elements of $\boldsymbol{\beta}$, and setting the derivatives to zero, it can be shown that the maximum likelihood estimates in log-linear Poisson models satisfy the estimating equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}, \tag{4.3}$$

where $\mathbf{X}$ is the model matrix, with one row for each observation and one column for each predictor, including the constant (if any), $\mathbf{y}$ is the response vector, and $\hat{\boldsymbol{\mu}}$ is a vector of fitted values, calculated from the m.l.e.'s $\hat{\boldsymbol{\beta}}$ by exponentiating the linear predictor $\boldsymbol{\eta} = \mathbf{X}'\hat{\boldsymbol{\beta}}$. This estimating equation arises not only in Poisson log-linear models, but more generally in any generalized linear model with canonical link, including linear models for normal data and logistic regression models for binomial counts. It is not satisfied, however, by estimates in probit models for binary data.

To understand equation 4.3 it helps to consider a couple of special cases. If the model includes a constant, then one of the columns of the model matrix $\mathbf{X}$ is a column of ones. Multiplying this column by the response vector $\mathbf{y}$ produces the sum of the observations. Similarly, multiplying this column by the fitted values $\hat{\boldsymbol{\mu}}$ produces the sum of the fitted values. Thus, in models with a constant one of the estimating equations matches the sum of observed and fitted values. In terms of the example introduced at the beginning of this chapter, fitting a model with a constant would match the total number of children ever born to all women.

As a second example suppose the model includes a discrete factor represented by a series of dummy variables taking the value one for observations at a given level of the factor and zero otherwise. Multiplying this dummy variable by the response vector $\mathbf{y}$ produces the sum of observations at that level of the factor. When this is done for all levels we obtain the so-called *marginal* total. Similarly, multiplying the dummy variable by the fitted values $\hat{\boldsymbol{\mu}}$ produces the sum of the expected or fitted counts at that level. Thus, in models with a discrete factor the estimating equations match the observed and fitted marginals for the factor. In terms of the example introduced at the outset, if we fit a model that treats marital duration as a discrete factor we would match the observed and fitted total number of children ever born in each category of duration since first marriage.

This result generalizes to higher order terms. Suppose we entertain models with two discrete factors, say A and B. The additive model $A+B$ would reproduce exactly the marginal totals by A or by B. The model with an interaction effect AB would, in addition, match the totals in each combination of categories of A and B, or the AB margin. This result, which will be important in the sequel, is the basis of an estimation algorithm known as *iterative proportional fitting*.

In general, however, we will use the iteratively-reweighted least squares (IRLS) algorithm discussed in Appendix B. For Poisson data with link log, the working dependent variable $\mathbf{z}$ has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i},$$

and the diagonal matrix $\mathbf{W}$ of iterative weights has elements

$$w_{ii} = \hat{\mu}_i,$$

where $\hat{\mu}_i$ denotes the fitted values based on the current parameter estimates.

Initial values can be obtained by applying the link to the data, that is taking the log of the response, and regressing it on the predictors using OLS. To avoid problems with counts of 0, one can add a small constant to all responses. The procedure usually converges in a few iterations.

4.2.2 Goodness of Fit

A measure of discrepancy between observed and fitted values is the deviance. In Appendix B we show that for Poisson responses the deviance takes the form

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\}.$$

The first term is identical to the binomial deviance, representing 'twice a sum of observed times log of observed over fitted'. The second term, a sum of differences between observed and fitted values, is usually zero, because m.l.e.'s in Poisson models have the property of reproducing marginal totals, as noted above.

For large samples the distribution of the deviance is approximately a chi-squared with $n-p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model.

An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as

$$\chi_p^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. The Pearson statistic has the same form for Poisson and binomial data, namely a 'sum of squared observed minus expected over expected'.

In large samples the distribution of Pearson's statistic is also approximately chi-squared with $n-p$ d.f. One advantage of the deviance over Pearson's chi-squared is that it can be used to compare nested models, as noted below.

4.2.3 Tests of Hypotheses

Likelihood ratio tests for log-linear models can easily be constructed in terms of deviances, just as we did in logistic regression models. In general, the difference in deviances between two nested models has approximately in large samples a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct.

One can also construct Wald tests as we have done before, based on the fact that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ has approximately in large samples a multivariate normal distribution with mean equal to the true parameter value $\boldsymbol{\beta}$ and variance-covariance matrix $\text{var}(\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{W}\mathbf{X}$, where $\mathbf{X}$ is the model matrix and $\mathbf{W}$ is the diagonal matrix of estimation weights described earlier.

4.3 A Model for Heteroscedastic Counts

Let us consider the data on children ever born from Table 4.1. The unit of analysis here is the individual woman, the response is the number of children she has borne, and the predictors are the duration since her first marriage, the type of place where she resides, and her educational level, classified in four categories.

4.3.1 The Mean-Variance Relation

Data such as these have traditionally been analyzed using ordinary linear models with normal errors. You might think that since the response is a discrete count that typically takes values such as 0, 2 or six, it couldn't possibly have a normal distribution. The key concern, however, is not the normality of the errors but rather the assumption of constant variance.

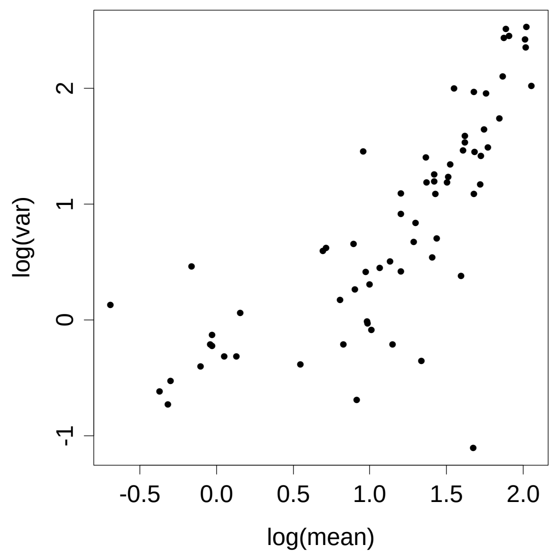

FIGURE 4.1: The Mean-variance Relationship for the CEB Data

In Figure 4.1 we explore the form of the mean-variance relationship for these data by plotting the variance versus the mean for all cells in the table with at least 20 observations. For convenience we use a log-log scale. Clearly, the assumption of constant variance is not valid. Although the variance is not exactly equal to the mean, it is not far from being proportional to it. Thus, we conclude that we can do far more justice to the data by fitting Poisson regression models than by clinging to ordinary linear models.

4.3.2 Grouped Data and the Offset

At this point you may wonder whether we need the individual observations to be able to proceed further. The answer is no; all the information we need is available in Table 4.1. To see this point let Y_{ijkl} denote 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, and let $Y_{ijk} = \sum_l Y_{ijkl}$ denote the group total. 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.

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

$$\log \mathrm{E}(Y_{ijkl}) = \log(\mu_{ijk}) = \mathbf{x}'_{ijk}\boldsymbol{\beta},$$

where $\mathbf{x}_{ijk}$ is a vector of covariates. Then the log of the expected value of the group total is

$$\log \mathrm{E}(Y_{ijk}) = \log(n_{ijk}\mu_{ijk}) = \log(n_{ijk}) + \mathbf{x}'_{ijk}\boldsymbol{\beta}.$$

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, which is known beforehand, is called an *offset*, and is a frequent feature of log-linear models for counts of events. Often, when the response is a count of events the offset represents the log of some measure of exposure, in our case the number of women.

Thus, we can analyze the data by fitting log-linear models to the individual counts, or to the group totals. In the latter case we treat the log of the number of women in each cell as an offset. The parameter estimates and standard errors will be exactly the same. The deviances of course, will be different, because they measure goodness of fit to different sets of counts. Differences of deviances between nested models, however, are exactly the same whether one works with individual or grouped data. The situation is analogous to the case of individual and grouped binary data discussed in the previous chapter, with the offset playing a role similar to that of the binomial denominator.

4.3.3 The Deviance Table

Table 4.3 shows the results of fitting a variety of Poisson models to the children ever-born data. The null model has a deviance of 3732 on 69 degrees of freedom (d.f.) and does not fit the data, so we reject the hypothesis that the expected number of children is the same for all these groups.

Introducing marital duration reduces the deviance to 165.8 on 64 d.f. The substantial reduction of 3566 at the expense of only five d.f. reflects the trivial fact that the (cumulative) number of children ever born to a woman depends on the total amount of time she has been exposed to childbearing,

Table 4.3: Deviances for Poisson Log-linear Models Fitted to the Data on CEB by Marriage Duration, Residence and Education

Model	Deviance	d.f.
Null	3731.52	69
One-factor Models		
Duration	165.84	64
Residence	3659.23	67
Education	2661.00	66
Two-factor Models		
$D+R$	120.68	62
$D+E$	100.01	61
DR	108.84	52
DE	84.46	46
Three-factor Models		
$D+R+E$	70.65	59
$D+RE$	59.89	53
$E+DR$	57.06	49
$R+DE$	54.91	44
$DR+RE$	44.27	43
$DE+RE$	44.60	38
$DR+DE$	42.72	34
$DR+DE+RE$	30.95	28

as measured by the duration since her first marriage. Clearly it would not make sense to consider any model that does not include this variable as a necessary control.

At this stage one could add to the model type of place of residence, education, or both. The additive model with effects of duration, residence and education has a deviance of 70.65 on 59 d.f. (an average of 1.2 per d.f.) and provides a reasonable description of the data. The associated P-value under the assumption of a Poisson distribution is 0.14, so the model passes the goodness-of-fit test. In the next subsection we consider the interpretation of parameter estimates for this model.

The deviances in Table 4.3 can be used to test the significance of gross and net effects as usual. To test the gross effect of education one could compare the one-factor model with education to the null model, obtaining a remarkable chi-squared statistic of 1071 on three d.f. In this example it really doesn't make sense to exclude marital duration, which is an essential control for exposure time. A better test of the effect of education would therefore compare the additive model $D+E$ with both duration and education to the one-factor model D with duration only. This gives a more reasonable chi-squared statistic of 65.8 on three d.f., still highly significant. Since educated women tend to be younger, the previous test overstated the educational differential.

We can also test the net effect of education controlling for type of place of residence, by comparing the three-factor additive model $D+R+E$ with the two-factor model $D+R$ with duration and residence only. The difference in deviances of 50.1 on three d.f. is highly significant. The fact that the chi-squared statistic for the net effect is somewhat smaller than the test controlling duration only indicates that part of the effect of education may be attributed to the fact that more educated women tend to live in Suva or in other urban areas.

The question of interactions remains to be raised. Does education make more of a difference in rural areas than in urban areas? To answer this question we move from the additive model to the model that adds an interaction between residence and education. The reduction in deviance is 10.8 on six d.f. and is not significant, with a P-value of 0.096. Does the effect of education increase with marital duration? Adding an education by duration interaction to the additive model reduces the deviance by 15.7 at the expense of 15 d.f., hardly a bargain. A similar remark applies to the residence by duration interaction. Thus, we conclude that the additive model is adequate for these data.

4.3.4 The Additive Model

Table 4.4 shows parameter estimates and standard errors for the additive model of children ever born (CEB) by marital duration, type of place of residence and education.

The constant represents the log of the mean number of children for the reference cell, which in this case is Suvanese women with no education who have been married 0–4 years. Since $\exp\{-0.1173\} = 0.89$, we see that on the average these women have 0.89 children at this time in their lives. The duration parameters trace the increase in CEB with duration for any residence-education group. As we move from duration 0–4 to 5–9 the log of the mean increases by almost one, which means that the number of CEB gets multiplied by $\exp\{0.9977\} = 2.71$. By duration 25–29, women in each category of residence and education have $\exp\{1.977\} = 7.22$ times as many children as

Table 4.4: Estimates for Additive Log-Linear Model of Children Ever Born by Marital Duration, Type of Place of Residence and Educational Level

Parameter		Estimate	Std. Error	z-ratio
Constant		−0.1173	0.0549	−2.14
Duration	0–4	–		
	5–9	0.9977	0.0528	18.91
	10–14	1.3705	0.0511	26.83
	15–19	1.6142	0.0512	31.52
	20–24	1.7855	0.0512	34.86
	25–29	1.9768	0.0500	39.50
Residence	Suva	–		
	Urban	0.1123	0.0325	3.46
	Rural	0.1512	0.0283	5.34
Education	None	–		
	Lower	0.0231	0.0227	1.02
	Upper	-0.1017	0.0310	-3.28
	Sec+	-0.3096	0.0552	-5.61

they did at duration 0–4.

The effects of residence show that Suvanese women have the lowest fertility. At any given duration since first marriage, women living in other urban areas have 12% larger families ($\exp\{0.1123\} = 1.12$) than Suvanese women with the same level of education. Similarly, at any fixed duration, women who live in rural areas have 16% more children ($\exp\{0.1512\} = 1.16$), than Suvanese women with the same level of education.

Finally, we see that higher education is associated with smaller family sizes net of duration and residence. At any given duration of marriage, women with upper primary education have 10% fewer kids, and women with secondary or higher education have 27% fewer kids, than women with no education who live in the same type of place of residence. (The last figure follows from the fact that $1 - \exp\{-0.3096\} = 0.27$.)

In our discussion of interactions in the previous subsection we noted that the additive model fits reasonably well, so we have no evidence that the effect of a variable depends on the values of other predictors. It is important to note, however, that the model is additive in the *log* scale. In the original scale the model is multiplicative, and postulates relative effects which translate into different absolute effects depending on the values of the other predictors. To clarify this point we consider the effect of education. Women with secondary or higher education have 27% fewer kids than women with no education. Table 4.5 shows the predicted number of children at each duration of marriage for Suvanese women with secondary education and with no education, as well as the difference between these two groups.

Table 4.5: Fitted Values for Suvanese Women with No Education and with Secondary or Higher Education

Marital Duration	0–4	5–9	10–14	15–19	20–24	25+
No Education	0.89	2.41	3.50	4.47	5.30	6.42
Secondary+	0.65	1.77	2.57	3.28	3.89	4.71
Difference	0.24	0.64	0.93	1.19	1.41	1.71

The educational differential of 27% between these two groups translates into a quarter of a child at durations 0–4, increases to about one child around duration 15, and reaches almost one and a quarter children by duration 25+. Thus, the (absolute) effect of education measured in the original scale increases with marital duration.

If we had used an ordinary linear regression model for these data we would have ended up with a large number of interaction effects to accommodate the fact that residence and educational differentials increase with marital duration. In addition, we would have faced a substantial problem of heteroscedasticity. Taking logs of the response would ameliorate the problem, but would have required special treatment of women with no children. The Poisson log-linear model solves the two problems separately, allowing the variance to depend on the mean, and modeling the log of the mean as a linear function of the covariates.