CORVINUS JOURNAL OF SOCIOLOGY AND SOCIAL POLICY Vol.5 (2014) 2, 97–114 DOI: 10.14267/cjssp.2014.02.04

THE USE OF POISSON REGRESSION IN THE SOCIOLOGICAL STUDY OF SUICIDE

Ferenc Moksony–Rita Hegedűs[1]

ABSTRACT This paper explains how Poisson regression can be used in studies in which the dependent variable describes the number of occurrences of some rare event such as suicide. After pointing out why ordinary linear regression is inappropriate for treating dependent variables of this sort, we go on to present the basic Poisson regression model and show how it fits in the broad class of generalized linear models. Then we turn to discussing a major problem of Poisson regression known as overdispersion and suggest possible solutions, including the correction of standard errors and negative binomial regression. The paper ends with a detailed empirical example, drawn from our own research on suicide.

KEYWORDS Deviant behavior, Generalized linear models, Poisson regression, Rare events, Research methods, Statistical analysis, Suicide

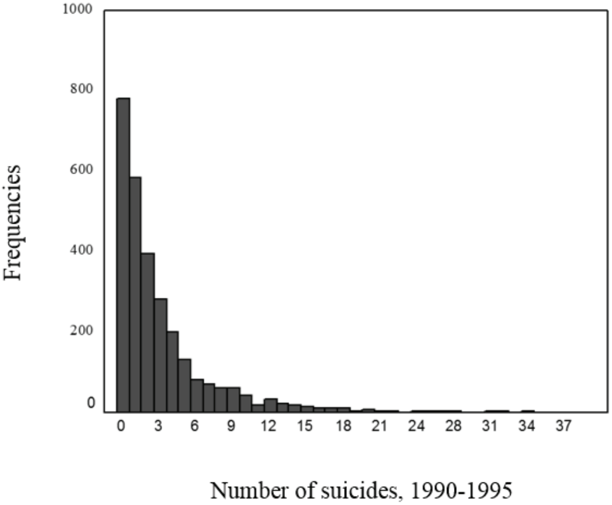
In social research, we often encounter dependent variables that describe the *frequency of occurrence* of events of various kinds. Students of deviant behavior, for example, look at whether media reporting of celebrity suicides leads to an increase in the number of self-destruction; demographers study the impact of environmental hazards on the number of birth defects; and sociologists of science examine the factors that influence the frequency with which publications are cited by fellow researchers. In cases like these, scholars frequently rely on ordinary linear regression to analyze their data. This sometimes produces acceptable results, especially when the mean frequency of occurrence of the event under study is relatively *large*, since in this situation the distribution of the dependent variable does not usually deviate substantially from the normal distribution assumed by ordinary least

[1] Ferenc Moksony is professor, Rita Hegedűs is associate professor at the Corvinus University of Budapest. Corresponding author's e-mail: *fmokson@gmail.com.* The research reported in this paper was financially supported by the Hungarian Scientific Research Fund (Grant no. T043441).

squares regression.

Often, however, the event we are interested in explaining is *rare* and the distribution of the dependent variable is highly *skewed*, with frequencies peaking at the lowest value and sharply declining toward the upper end of the scale. Figure 1 illustrates this, showing the frequency distribution of suicides in Hungarian villages from 1990 to 1995.[2]

Figure 1 *Frequency distribution of suicide in Hungarian villages, 1990–1995*



What we see on this graph is a far cry from the familiar symmetrical bell-shaped curve of normal distribution: almost 800 villages had absolutely no suicide during the six years under study and another 600 had but one. The number of communities with many cases of death, in contrast, is very low, with only 38 villages registering more than twenty suicides. Variables with such asymmetric, right-skewed distributions can be approximated with an important class of discrete distribution, Poisson distribution.[3]

[2] The graph displays data for areas officially recognized as villages for the whole period from 1990 to 1995.

[3] As the mean frequency of occurrence of the event under study increases, the shape of the Poisson distribution gradually approaches that of normal distribution (see, e.g., Sváb 1981: 498). The Poisson distribution, then, is of greatest importance for the study of rare events (cf., Land - McCall 1996).

PROBLEMS OF ORDINARY LEAST SQUARES AND POSSIBLE SOLUTIONS

One important characteristic of Poisson distribution is that the *mean is equal to the variance*. This is easy to understand if we regard the Poisson distribution as an extreme case of binomial distribution in which the probability of one of two possible outcomes (e.g., suicide/no suicide, birth defect/no birth defect) is *very small*. The mean of a binomially distributed variable is

$$\bar{x} = np,$$

and its variance is

$$s^2 = np(1-p),$$

where *n* is the number of observations and *p* is the relative frequency of one of the two possible outcomes. As *p* gets smaller and smaller, the term in parentheses on the right hand side of the equation approaches 1, and thus s^2, the variance, approaches *np*, the mean.

How does this characteristic of Poisson distribution affect the use of ordinary linear regression? In regression analysis, we assume the conditional mean of the dependent variable is some function of the explanatory variables; that is, we assume that the former changes in a systematic way with the values of the latter. With Poisson-distributed dependent variables, this assumption implies that the conditional *variance* of the dependent variable, being equal to the mean, also changes with the values of the explanatory variables, violating an important requirement of ordinary least squares regression, namely, that the conditional variance of the dependent variable is the *same* for all levels of the independent variables (homoscedasticity).

Conventional regression methods, then, are not generally appropriate for dependent variables that describe the frequency of rare events such as suicides or birth defects. What next, then? One option is to *transform the dependent variable* in order to make it satisfy the assumptions underlying ordinary least squares, and then use the transformed data in place of the original ones. Researchers taking this approach commonly employ the square-root transformation, which has the advantage of eliminating the mean-variance identity, since the variance of the square-root of a Poisson variable is independent of its mean (Chatterjee & Price 1977: 39). An additional benefit is that the distribution of the new variable will be more similar to the normal distribution, which is important for significance tests.

While transforming the dependent variable is a common practice that is widely adopted by researchers (e.g., Moksony 2001), another possibility is to modify the regression model, making it suitable for the analysis of Poisson variables. This approach, in some sense, is the reverse of the previous one: rather than tailoring the distribution of the dependent variable to the requirements of ordinary least squares, here we proceed the other way around and *tailor the regression model to the distribution of the dependent variable*. This "custom-made" version of regression analysis is known as Poisson regression.

Poisson regression belongs to the family of generalized linear models (Hoffman 2004; Agresti 1996: Ch. 4). These models extend the scope of ordinary linear regression in two ways. First, they describe *transformations* of the conditional mean of the dependent variable, rather than the mean itself, as linear functions of explanatory variables; second, they allow the dependent variable to have conditional distributions *other than the normal*. Various forms of generalized linear models differ from each other in the particular type of transformation applied and the specific distribution assumed for the dependent variable. Table 1, adapted from Agresti (1996: 97), lists the most important of these models.

Table 1 *Generalized linear models*

Model	Transformation applied to the mean	Distribution of Dependent Variable	Type of explanatory variables
Poisson regression	Logarithm	Poisson	Numerical and Categorical
Logistic regression	Logit	Binomial	Numerical and Categorical
Linear regression	Identity	Normal	Numerical
Analysis of variance	Identity	Normal	Categorical
Analysis of covariance	Identity	Normal	Numerical and Categorical

Source: Agresti 1996: 97 (Table 4.5).

In Poisson regression, as can be seen, logarithmic transformation is used and the dependent variable is taken to follow a Poisson distribution. In contrast, in logistic regression, a logit transformation is employed and the dependent variable is assumed to have a binomial distribution. The table also includes ordinary linear regression, analysis of variance and analysis of covariance, which together comprise what is known as the *general* linear model (Fennessey 1968; Cohen 1968). In these three types of models, the normal distribution is assumed for the dependent variable and the identity transformation is used; that is, the conditional mean itself is directly described as a linear function of the explanatory variables. All these variants of the general linear model, then, are just special cases of the *generalized* linear model.[4]

THE POISSON REGRESSION MODEL AND INTERPRETATION OF ITS COEFFICIENTS

In the simple bivariate case, the Poisson regression model has the following form:

$$\ln \lambda = \beta_0 + \beta_1 X \qquad (1)$$

where λ denotes the mean or expected value of the dependent variable, X is the independent variable, and β_0 and β_1 are the regression coefficients to be estimated. These coefficients have essentially the same meaning as in ordinary linear regression; β_1, in particular gives the change in the natural logarithm of the mean frequency of the dependent variable per one unit change in the explanatory variable. This interpretation is not very user-friendly, however; researchers, after all, do not usually like to think in terms of logarithms. It is, therefore, useful to reformulate Eq. (1) by taking the antilogarithm of both sides:

$$\lambda = \exp(\beta_0 + \beta_1 X) = \exp(\beta_0) * \exp(\beta_1 X) \qquad (2)$$

In contrast to β_1 , which represented the *additive* effect of the explanatory variable on the *log* of the mean frequency, *exp(*β_1*)* represents its *multiplicative* effect on the mean frequency itself, indicating how many times larger (or smaller) the mean frequency of the phenomenon under study becomes as the independent variable increases by one unit. To see this, suppose X is equal to a, which can be any number. Eq. (2) then looks like this:

$$(\lambda \mid X = a) = \exp(\beta_0 + \beta_1 a) = \exp(\beta_0) * \exp(\beta_1 a),$$

where the vertical bar refers to the condition that the explanatory variable takes on one particular value, namely, a. Let us now increase the value of the

[4] Roughly speaking, while the *general* linear model extends the scope of ordinary regression by allowing categorical *independent* variables to be included in the analysis, *generalized* linear models go one step further and also lift the constraints imposed on the *dependent* variable.

explanatory variable by one unit, from *a* to *a+1*. Eq. (2) now takes this form:

$$(\lambda \mid X = a+1) = \exp\left[\beta_0 + \beta_1(a+1)\right] = \exp(\beta_0) * \exp(\beta_1 a) * \exp(\beta_1).$$

Taking the ratio of the two equations, we get:

$$\frac{(\lambda \mid X = a+1)}{(\lambda \mid X = a)} = \frac{\exp(\beta_0) * (\beta_1 a) * \exp(\beta_1)}{\exp(\beta_0) * (\beta_1 a)} = \exp(\beta_1).$$

We can see that as the value of the explanatory variable, *X*, has increased from *a* to *(a+1)*, that is, by one unit, the mean frequency of the dependent variable, λ, has indeed changed by a factor equal to *exp*(β_1).

It can be helpful to express this change in percentage form using the following formula:

$$\text{percentage change} = 100 * \left[\exp(\beta_1) - 1\right].$$

If, for example, the dependent variable is the number of suicides in a village, the explanatory variable is the unemployment rate, and the value of β_1 is, say, .055, then exp (β_1) = exp (.055) = 1.057 and 100*(exp (β_1)–*1=100*(1.057–1)=5.7,* which means each percentage point rise in the level of unemployment increases, on average, the number of suicides by 5.7 percent. In the same way, if the independent variable is type of settlement, with villages coded 0 and cities 1, and the value of β_1 is, say, –.035, then exp (β_1)=exp(–.35)=.70, and 100*(exp (β_1)–1)=100*(.70–1)= –30, which means the number of suicides is, on average, 30 percent less in cities than in villages.[5]

INCLUDING POPULATION AT RISK IN THE MODEL

One important feature of the Poisson regression model discussed thus far is that it does not take *differences in the size of the population at risk* into account. This is clearly a limitation because, with other factors remaining

[5] It should be noted, though, that with dichotomous independent variables like type of settlement (cities/villages), *exp* (β_1) can only be interpreted directly as reflecting the impact of those variables when we use *dummy* coding; that is, when we assign the values 0 and 1 to the two categories. This is because in this case, the distance or difference between the categories exactly equals one unit. With *effect* coding, in contrast, when the values –1 and +1 are used to denote the two categories, the distance or difference is two rather than one unit and thus *exp* (β_1)no longer directly indicates the influence of the explanatory variable under study. To get this influence, *exp* (β_1) has to be raised to the second power, to accommodate the greater distance between categories.

constant, a larger population at risk can obviously be expected to produce a greater frequency of the phenomenon under study. More populous cities, for instance, will in general have higher numbers of suicides simply because of their dimensions, quite regardless of the impact of other factors, such as poverty, unemployment, or residential segregation.

Differences in the size of the population at risk can be controlled for by applying some sort of standardization, dividing the mean frequency by the population at risk:

$$\ln\left(\frac{\lambda}{n}\right) = \beta_0 + \beta_1 X \qquad (3),$$

where *n* is the population at risk, such as the number of individuals living in a city or village. This modified version of the Poisson regression model describes the *rate* of occurrence of the phenomenon under study, rather than its absolute frequency, as a linear function of *X*, the independent variable. By making use of the properties of logarithms, Eq. (3) can also be written as

$$\ln(\lambda) - \ln(n) = \beta_0 + \beta_1 X \qquad (4),$$

which, in turn, can be rearranged to get:

$$\ln(\lambda) = \ln(n) + \beta_0 + \beta_1 X \qquad (5).$$

This model differs from the one given by Eq. (1) in that it contains a separate explanatory variable, *ln*(*n*), generally called an *offset*, which reflects the size of the population at risk, the coefficient of which is automatically taken to equal 1. If the size of the population at risk is the same in each unit of observation, then *ln*(*n*) will be constant and thus can be incorporated into β_0, which takes us back to the original formulation; that is, Eq. (1).

OVERDISPERSION

Besides including a separate independent variable to capture differences in the size of the population at risk, the initial model of Poisson regression also often needs to be modified for another reason. One important characteristics of the Poisson distribution, as already noted, is that the mean is equal to the variance. The default model of Poisson regression is built on this identity; computations are performed assuming the mean and the variance of the dependent variable really are the same. In many cases, however, this assumption does not hold true and *the variance exceeds the mean*. This is what is known in the statistical literature as overdispersion (Agresti 1996: 92–3; Le 1998: 226–228).

Overdispersion generally arises from one of two sources (King 1989: 766–769; Osgood 2000: 28). First, in almost all research, *some explanatory variables uncorrelated with the ones included in the analysis are left out from the model*, either because they do not come to mind, or because we cannot capture them empirically. Suppose, for example, the frequency of suicide for cities or villages is influenced by economic development and regional culture, but we only include the former in the study. What is likely to happen in this situation? Holding the level of economic development constant, areas belonging to that level will be geographically – and culturally – *heterogeneous*; they will be a mixture of sub-populations, with each sub-population having its own distinct, regionally-determined mean frequency of suicide. That is, the mean of the dependent variable for a certain value of explanatory variable will not be constant, as is assumed by the Poisson distribution, but we will instead have as many different means as there are regions. The result of this heterogeneity is that the variance of the conditional distribution of suicide – that is, of the distribution pertaining to a given level of economic development – will be greater than expected based on the Poisson distribution. The source of overdispersion, then, in this case is the *excess variation caused by explanatory variables left out from the model*; quite similarly to ordinary linear regression, where the impact of the omission of independent variables uncorrelated with those included in the analysis is to increase the error or residual variance.

Another factor that may give rise to overdispersion is *dependence among observations*. One assumption of Poisson distribution is that observations are independent of each other; that the occurrence of an event (e.g., a suicide) does not influence the occurrence of another. What happens when this requirement is not met; when, for example, the incidence of a suicide increases the chance of another? In this situation, the *frequency of small and large values – that is, those at the two extreme tails – will be greater than expected* based on the Poisson distribution and the variance of the variable will *increase* as a result. If, for instance, kids in a school imitate each other's behavior, then instead of observing a random distribution of cases, what we will likely see is that while nothing happens for months, a substantial number of suicide attempts take place within a very short time frame. The literature is replete with examples of such time-space clusterings of cases (e.g., Phillips 1974; Phillips and Carstensen 1986; Gould et al. 1990). From our point of view, the important thing about these examples is that the process of imitation and the resulting dependence of observations may lead to an increase of the variance of the dependent variable, thus violating a fundamental assumption of Poisson distribution, namely that the mean is equal to the variance.

How does overdispersion affect Poisson regression results? The most serious consequence is that although regression coefficients remain unbiased, their *standard errors will be underestimated* and thus confidence intervals will be unduly narrow and significance tests will give overly optimistic results. One way to cope with this problem is to *adjust the standard errors* using a correction factor that reflects the degree of overdispersion. This correction is performed based on an analysis of the residuals from the original Poisson regression and involves forming the ratio of the sum of squared standardized residuals to their degrees of freedom. If this ratio, also called dispersion parameter, turns out to be considerably greater than 1, then, provided the model is well-specified[6], this is a sign of overdispersion and standard errors can be adjusted by multiplying them with the square root of the dispersion parameter (Agresti 1996: 93; Allison 2001: 223; Gelman and Hill 2007: 114-116).

Another way, besides correcting standard errors, to cure the problem of overdispersion is to switch to a model that *incorporates the excess variability* not captured by Poisson regression. In this new model, the conditional mean of the dependent variable is *no longer a constant*, as in the original formulation, but a *variable* that scatters randomly around some central value, due to the effect of omitted explanatory variables. What we do, basically, is to *add a random component* to Eq. (2):

$$\widetilde{\lambda} = \exp(\beta_0 + \beta_1 X + e) \qquad (6)$$

where $\widetilde{\lambda}$ is a random variable that replaces λ from Eq. (2) and e is the newly added random error term that represents the additional heterogeneity introduced by causal factors not explicitly considered in the analysis. The main difference between Eqs. (2) and (6) is that while in Eq. (2) for each level of the independent variable we had a *single* mean value of the dependent variable (λ), in Eq. (6), we have a whole *distribution* of means ($\widetilde{\lambda}$), corresponding to the fact that now the omitted variables subsumed in the error term make the individual means randomly deviate from the level determined

[6] This qualification is important because a high value for the ratio of the sum of squared standardized residuals to their degrees of freedom may also indicate lack of fit or misspecification, rather than overdispersion, such as when a linear model is used to describe a curvilinear relationship (Bair, 2013).

by the systematic part of the model (Long 1997: 230–231; Gardner et al. 1995: 399–400).

This modification of Poisson regression, then, goes beyond the initial model by changing the mean of the distribution *from a constant to a variable.* In the original formulation, the dependent variable (*Y*) was a random variable that scattered around the mean (*λ)* following a Poisson distribution. The mean itself, however, was a constant – its value was unequivocally determined by the value of the explanatory variable. In the new model, *Y* continues to be a random variable scattering around a mean level following a Poisson distribution; but now, *this mean level is also a random variable* that fluctuates, following a Gamma-distribution, around some central value, due to the random error, *e.* Since the expected value of the error is zero by assumption, the central value that the individual means ($\tilde{\lambda}$) scatter around is identical to the mean from the original model given by Eq. (2); that is,

$$E(\tilde{\lambda}) = \lambda .$$

In final analysis, then, in the modified form of Poisson regression we have a *mixture of two different distributions:* first, we have the dependent variable (*Y*) that scatters around the mean ($\tilde{\lambda}$) following a Poisson distribution; second, we have the mean that is itself a random variable that fluctuates around a central value following a Gamma-distribution. The result of the combination of these two distribution is the negative binomial distribution, which is why the modified form of Poisson regression described here is called *negative binomial regression.*

The advantage of negative binomial regression over conventional Poisson regression is *that it does not require the variance to be equal to the mean* and allows the former to exceed the latter. As against the original form of Poisson regression, where

$$\mathrm{E}(Y) = \mathrm{var}(Y) = \lambda,$$

in negative binomial regression,

$$\mathrm{E}(Y) = \lambda \quad \text{and} \quad \mathrm{var}(Y) = \lambda(1+\alpha\lambda) = \lambda + \alpha\lambda^2,$$

and

where α is the dispersion parameter that indicates the degree of overdispersion. In the special case when there is no overdispersion, $\alpha = 0$ and *var (Y)* $= \lambda + \alpha \lambda^2 = \lambda$ and negative binomial regression simplifies to Poisson regression.

AN EXAMPLE: LOCAL AREA DEPRIVATION AND SUICIDE IN RURAL HUNGARY

To illustrate the use of the methods discussed in the previous sections of this paper, we now present results from one of our studies that looked at the impact of deprivation on suicide in rural Hungary. Given the main character of our article, in what follows, we focus on methodological issues at the expense of substantive details. The analysis spanned the years from 1990 to 1995 and covered areas officially qualified as villages throughout the whole period. The number of villages meeting this criterion was 2869 and the number of suicides committed in those villages was 9237. Statistical data analysis proceeded in two steps. We first employed principal component analysis to create a composite measure of deprivation, then we used that measure as the main explanatory variable in a Poisson regression in which the number of suicides was the dependent variable and the size of the population at risk was entered as an offset.

The following indicators were included in the principal component analysis:

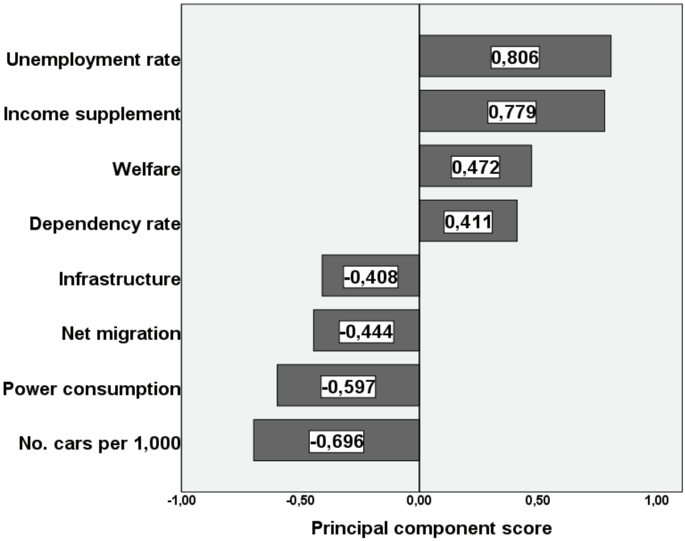
– Number of cars per 1,000 inhabitants, 1992-1995
– Power consumption per household, 1993-1995
– Quality of educational infrastructure and health services, 1993[7]
– Unemployment rate, 1993-1994
– Percentage of the population living on welfare, 1993-1995
– Percentage of the population receiving income supplement, 1993-1995
– Net migration, 1990-1995
– Dependency rate[8]

Figure 2 shows the principal component loadings for these indicators. Variables capturing economic development, such as power consumption and the number of cars, all have negative loadings, while those reflecting the lack thereof, such as unemployment and percentage living on welfare, all have positive loadings, lending some face validity to the principal component as a summary measure of deprivation.

[7] This variable measures the number of educational and health institutions such as schools and hospitals.

[8] This is the ratio of individuals below 18 and above 59 to those between 18 and 59.

Figure 2. *Graphic display of principal component loadings*



Having constructed our composite index of deprivation, in the second phase of the analysis, we ran Poisson regression, using the following model:

$$\ln(\lambda_i) = \ln(n_i) + \beta_0 + \beta_1 X_i,$$

where λ_t is the average number of suicide in village *i* during 1990 to 1995; n_i is the size of the population for the same village in the same period; and X_i is the principal component score for village *i.* We employed the statistical software Stata 8.0 to estimate the regression coefficients, β_0 and β_1.

Table 2 displays the results obtained from Poisson regression. The coefficient for the principal component score is, as can be seen, positive and statistically significant, with a one unit increase in our summary measure of deprivation being associated with an increase of about .115 in the natural logarithm of the number of suicides. To get rid of logarithms, we exponentiate the coefficient and get exp(.1148) = 1.122, which means that a one unit rise in the principal component score raises the number of suicides by 12.2 percent. All in all, these findings suggest that local deprivation aggravates the risk of self-destruction.

Table 2. *Effect of deprivation on suicide. Poisson regression results*

Variable	Coefficient	Std. error	Z-value	Antilog of coefficient
Principal component score	.1148*	.0114	10.11	1.122
Constant	-7.7462	.0110	-706.17	

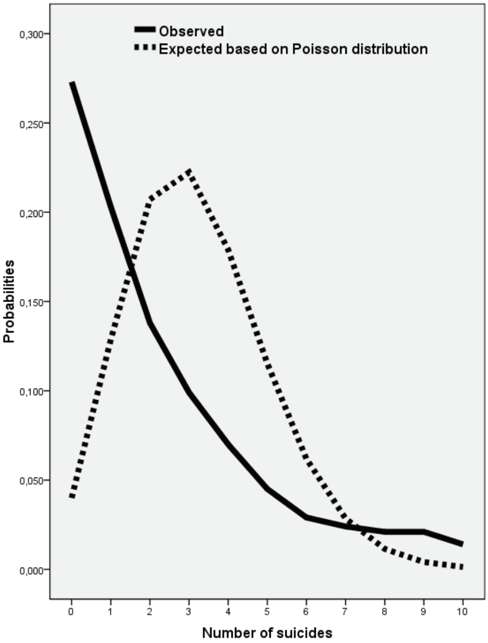
n = 2869 * p < .001

Checking for overdispersion

As already noted, overdispersion, while not introducing bias into the regression coefficients, leads to an underestimation of standard errors, thereby potentially undermining the validity of confidence intervals and significance tests. It is, therefore, important, before going too far with our conclusions, to check for the presence of overdispersion and adjust the analysis accordingly, if necessary.

For this purpose, we compared the observed distribution of suicides with the one predicted from the Poisson model. The latter gives, for different frequencies of occurrence, the probability that we would expect if the mean number of suicides was the same as the one actually obtained (3.22) and the assumptions underlying the Poisson model were fully met. *Figure 3* displays the results. As can be seen, *at the two tails* of the distribution, the solid line runs *above* the dotted one, indicating that at the lowest and highest number of suicides, observed proportions *exceed* Poisson probabilities. In the *middle* portion of graph, in contrast, the solid line consistently stays *below* the dotted one, implying that over this range, observed proportions are *lower* than those predicted from the model. Observed values, then, appear to be spread out more widely than those calculated based on the Poisson distribution. Corresponding to this finding, we found the variance of the observed distribution to be much larger (22.72) than its mean (3.22) and the ratio of the sum of squared standardized residuals to their degrees of freedom also proved to be higher than 1 (1.42), which, as already noted, is a sign of overdispersion, provided the model is well-specified. All in all, these results speak for the fact that overdispersion presents a real threat to the validity of conclusions drawn from our study.

Figure 3. *Observed and expected probabilities* (based on Poisson distribution, mean = 3.22)



Correcting for overdispersion

Earlier in this paper, we described two ways to handle the problem of overdispersion: adjusting standard errors and switching from Poisson to negative binomial regression. As for the former, original standard errors are, as previously explained, multiplied by the square root of the dispersion parameter, which is the ratio of the sum of squared standardized residuals to their degrees of freedom. *Table 4* reports these corrected standard errors, along with the original ones, so we can better judge the change that results from the adjustment.[9]

[9] Stata reports two different corrected standard errors: one is based on the traditional Pearson Chi-square, which is equivalent to the ratio of the sum of squared standardized residuals to their degrees of freedom, while the other is calculated using the Likelihood-ratio Chi-square. Although the two usually give similar results and thus the choice between them does not generally make much difference, statistical theory suggests that we prefer Pearson Chi-square to Likelihood-ratio Chi-square (Allison 2001: 223).

Table 3. *Observed and expected distribution of suicide*

Number of suicide	Observed relative frequency	Expected probability (based on Poisson distribution)
0	.273	.040
1	.203	.129
2	.138	.207
3	.099	.222
4	.070	.179
5	.045	.115
6	.029	.062
7	.024	.028
8	.021	.011
9	.021	.004
10	.014	.001

Table 4. *Poisson regression: corrected standard errors*

Variable	Coefficient	Original standard error	Corrected standard error based on	
			Pearson Chi-square	Likelihood-ratio Chi-square
Principal component score	.1148	.0114 (10.11)	.0140 (8.50)	.0135 (8.20)
Constant	-7.7462	.0110 (-706.17)	.0130 (-593.65)	.0140 (-573.22)

Note: Numbers in parentheses below standard errors are z-values

As can be seen, although the standard errors have increased somewhat after correcting them for overdispersion and the z-values have correspondingly declined slightly (since the coefficients themselves have remained unchanged), the effect of deprivation, as captured by the principal component score, continues to be highly statistically significant: the z-value is 6.74 and the associated p-value is well below the usual 5% threshold.

Besides adjusting the standard errors, we also ran negative binomial regression, the results of which are displayed in *Table 5.* The coefficient for the principal component score is similar to that obtained from Poisson regression, both testifying to the harmful effect of deprivation on suicide. The antilogarithm of the coefficient equals 1.109, meaning that a one unit increase in the principal component score increases the mean number of suicides by about 11 percent. This effect of deprivation is statistically significant, as is the estimate of the dispersion parameter, alpha, which provides further evidence that overdispersion presents a problem in our study.

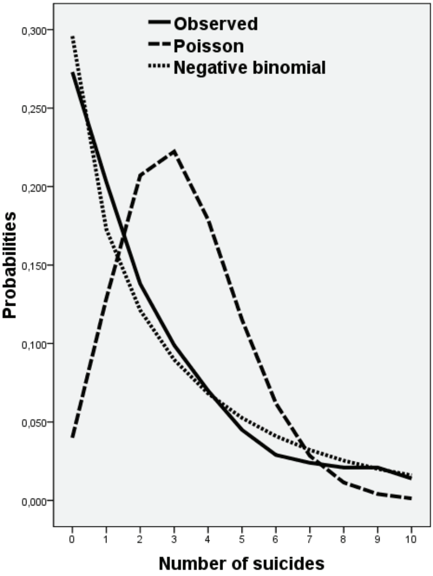
***Table 5.** The impact of deprivation on suicide. Results from negative binomial regression*

Variable	Coefficient	Std. error	Z-value	Antilog of coefficient
Principal component score	.1034*	.0153	6.74	1.109
Constant	–7.7816	.0144		
Alpha	.1462*	.0130		

n = 2869 * p < .001

Although the two regressions did not produce markedly different results, the negative binomial regression appears to fit the data better than the Poisson regression. This is evident from *Figure 4,* where the actual distribution of suicides is shown along with the distributions predicted from Poisson and negative binomial regressions. As can be seen, while Poisson regression systematically underestimates very low and very high frequencies, the negative binomial regression gives predicted values that are fairly close to the observed ones even at the two tails.

***Figure 4.** Observed and expected number of suicides, based on Poisson and negative binomial regression (mean = 3.22, overdispersion = 1.402)*



SUMMARY AND CONCLUSIONS

Our aim with this paper was to give an introduction to Poisson regression, which represents an important class of generalized linear models and can profitably be used in studies in which the dependent variable describes the *number of occurrences of some rare event* such as suicide. Although researchers sometimes apply ordinary linear regression in these situations, this method, as we have shown, is not generally appropriate for variables of this sort. And while transformation of the dependent variable may help alleviate part of the difficulties associated with conventional regression techniques, Poisson regression approaches the problem more fittingly by *tailoring the model to the distribution of the dependent variable,* rather than the other way around.

REFERENCES

Agresti, Alan (1996), *An introduction to categorical data analysis.* New York, Wiley.

Agresti, Alan (2002), *Categorical data analysis.* 2nd edition. New York, Wiley.

Allison, Paul (2001), *Logistic regression using the SAS system. Theory and application.* Cary, NC, SAS Institute

Bair, H. (2013), "Poisson Regression: Lack of Fit ≠ Overdispersion", *StatNews #86,* Cornell University, http,//www.cscu.cornell.edu/news/statnews/stnews86.pdf. Accessed July 22, 2014

Chatterjee, Samprit – Bertram Price (1977), *Regression analysis by example.* New York, Wiley.

Cohen, Jacob (1968), "Multiple regression as a general data-analytic system". *Psychological Bulletin,* 70, 426–443.

Fennessey, James (1968), "The general linear model, a new perspective on some familiar topics". *American Journal of Sociology,* 74, 1–27.

Gardner, William et al. (1995), "Regression analyses of counts and rates, Poisson, overdispersed Poisson and Negative Binomial models", *Psychological Bulletin,* 118, 392–404.

Gelman, Andrew – Jennifer Hill (2007), *Data analysis using regression and multilevel/hierarchical models,* Cambridge University Press, New York

Gould, Madelyn et al. (1990), "Time-space clustering of teenage suicide", *American Journal of Epidemiology,* 131, 71–78.

Hoffman, John (2004), *Generalized linear models,* Boston, Pearson Education Inc.

King, Gary (1989), "Variance specification in event count models, from restrictive assumptions to a generalized estimator", *American Journal of Political Science,* 33, 762–784.

Land, Kenneth – Patricia McCall (1996), "A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models", *Sociological Methods & Research,* 24, 387–443.

Le, Chap (1998), *Applied categorical data analysis,* New York, Wiley.

Long, J. Scott (1997), *Regression models for categorical and limited dependent variables,* Thousand Oaks, Sage.

Moksony, Ferenc (2001), “Victims of change or victims of backwardness? Suicide in rural Hungary”, in: Lengyel, Gy. - Rostoványi, Zs., ed., *The small transformation. Society, economy and politics in Hungary and the new European architecture,* Budapest, Akadémiai Kiadó, 366-376.

Osgood, D. Wayne (2000), “Poisson-Based Regression Analysis of Aggregate Crime Rates”, *Journal of Quantitative Criminology,* 16, 21–43.

Phillips, David (1974), “The Influence of Suggestion on Suicide, Substantive and Theoretical Implications of the Werther Effect”, *American Sociological Review,* 39, 340–354.

Phillips, David – Lundie Carstensen (1986), Clustering of teenage suicides after television news stories about suicide, *New England Journal of Medicine,* 315, 685–89.

Sváb János (1981), *Biometric methods in research [Biometriai módszerek a kutatásban],* Budapest, Mezőgazdasági Kiadó.