Journal of Applied Mathematics, Statistics and Informatics (JAMSI), **6** (2010), No. 1

# TICK HOST-SEEKING ACTIVITY AND TICK-BORNE ENCEPHALITIS INCIDENCE: REGRESSION AND HOMOGENEITY

VÁCLAV HÖNIG, MILAN STEHLÍK, VLASTA DANIELOVÁ, MILAN DANIEL, PAVEL ŠVEC AND LIBOR GRUBHOFFER

## Abstract

Data on the distribution of ticks and host-seeking activity of tick populations were acquired within the framework of the project „Ticks and tick-borne infectious diseases under the conditions of South Bohemia and Bavaria“ by terrain work on 30 locations in the South Bohemian region (Czech Republic). The data were statistically analyzed in order to characterize the relationship between the host-seeking activity of tick populations and the number of clinical cases of tick-borne encephalitis (TBE). In this paper we discuss the usage of regression tools for modelling the tick activity depedant TBE incidence. In particular, we use a logarithmical nonlinear regression model and, for the sake of completeness, also the linear regression. We discuss the heterogeneity of the data with respect to the logarithmic regression fit

**Mathematics Subject Classification 2000**: 62K05, 62P10
**Additional Key Words and Phrases**: regression, optimal design, model selection, homogeneity testing

## 1. INTRODUCTION

A relatively large amount of data was acquired by terrain work within the project „Ticks and tickborne infectious diseases under the conditions of South Bohemia and Bavaria“. The data describe the distribution and hoste-seeking activity of Ixodes ricinus tick populations in a network of 30 testing locations selected in the Region of South Bohemia (Czech Republic). The data were analyzed in order to characterize the relationship between the host-seeking activity and the number of clinical cases of tick-borne encephalitis (TBE), the second most widespread tick-borne disease in Europe. TBE was selected because of its relatively accurate case definition and diagnostics (when compared to Lyme borreliosis, the most widespread tick-borne disease in Europe). A clinical case of TBE can be understood as an actual realization of the risk represented by the tick activity. The prevalence of TBE virus (portion of samples positively tested for the pathogen) in ticks and the differences in the prevalence in high-risk and low-risk areas are generally low (0-5%). Therefore we may consider the tick activity the major factor influencing the local TBE risk and thus the tick host-seeking activity may be a useful predictor of TBE disease risk as shown similarly for Lyme borreliosis (see e.g. [Kitron and Kazmierczak (1997)]). Following data were used in the analysis: number of inhabitants in municipality (The Population and Housing Census 2001, CZSO), number of TBE cases in municipality (EPIDAT National Health Institute, Prague), host seeking activity of nymphal and adult ticks (estimated as number of ticks collected by one person per hour of collection).

From the biological justification we have found a reasonable model of the relationship between tick host-seeking activity and TBE incidence to be a 3 parameter logarithmic regression model $y = a + b\log(x+c)$. Such a model may be a rather flexible tool of saturation. We have observed somehow interesting unexpected clustering above and below the regression fit. Therefore, we have decided to seek for a reasonable heterogeneity measure/test. There exist a class of exact homogeneity likelihood ratio tests spanning from works [Stehlík (2003)], [Stehlík (2006)], [Stehlík and Wagner (2009)]. However, none of these procedures takes into account the observed clustering (which is somehow atypical for a classical homogeneity testing). Therefore, one aim of this paper is to adjust the existing procedures for observed clustering information. The other, and probably major aim is to provide some results obtained on density dependent incidence modelling.

## 2. ADAPTATION OF THE EPIDEMIOLOGICAL DATA

Statistical analysis were carried out using STATISTICA 8.0 (StatSoft,USA) and R version 2.8.1 (The R Foundation for Statistical Computing). At first, we worked with the data available for all of the 623 municipalities of the South Bohemian region. Strong linear relationship was found between the total number of TBE cases and number of inhabitants of particular municipality (R2= 0.647; F= 1140.239; p= 0.00001). Linear regression was carried out also for the data from our 30 locations only. Similar results were acquired with this data subset (R2= 0.629; F= 50.162; p= 0.00001). An actual TBE case requires a tick carrying TBE virus to come in contact with human, thus the number of TBE cases is influenced by the activity of human. The TBE risk is in some works derived directly from the number of TBE cases in the area. Such a risk assessment may be strongly influenced by the human activity. Areas with low activity of humans may result in virtually low TBE risk, underestimating the actual risk. We were interested in the "biological background risk" of TBE independent on the activity of humans. Since the human activity can hardly be measured and expressed, we decided to at least minimize the obvious sources of local inhomogeneities of human activity in the case of different human population size in small and large settlements. Therefore the TBE incidence was relativized to the number of inhabitants per municipality. The data were recalculated as number of cases per 1,000 inhabitants to obtain comparable entries for all locations.

## 3. RELATIONSHIP BETWEEN THE DISEASE INCIDENCE AND TICK HOST-SEEKING ACTIVITY

The data from the tick collections were employed. The host-seeking activity of ticks was expressed as the number of ticks (nymphal and adult) collected by one person per hour. The data were acquired in three collection periods (May, June, September 2008) with regard to the seasonal dynamics of the tick activity. For this set of tests, the results from the three collections were expressed as an annual mean for each location. The relationship between the number of TBE cases and the mean annual host-seeking activity of ticks was examined in further detail in order to find out whether the tick activity, as estimated in our study, was correlated with the number of TBE cases.

Interestingly, no significant relationship between the mean annual activity of ticks and annual number of TBE cases per 1,000 inhabitants was found using simple linear regression. Only negatively sloped regression curves were obtained using the log regression. Only if the regression through zero was included, biologically interpretable and statistically significant results were acquired. The error was probably due to inaccurate design of the selection of the sampling sites - the selection was not random, areas with high TBE incidence and thus probably high tick activity were preferred (see [Švec et al. (2009)]). Nevertheless, the existence of a point of zero tick activity and thus zero TBE incidence can be assumed. Therefore, we propose a regression curve with start at zero. The character of the curve in the area of low TBE incidence is affected by big error since no data are available for this region. Linear regression (F=10.1244; p=0.0035; R2=0.258) did not represent the saturation character of the data, therefore log regression (F=13.799; p=0.0009; R2=0.322) was employed (see Figure 1).

## 4. TESTING FOR HETEROGENEITY

The regression curve still did not seem to represent the trend in the data well. The individual observations seemed to form two compact clusters, one above and one under the curve. Therefore we have decided to test the dependent variable for homogeneity. Since the data had no normal distribution, but rather exponential or gamma and the number of observation was relatively small we were forced to use an exact likelihood ratio test developed in [Stehlík (2006)]. Critical value was calculated according to the ELRH test and we got $-\ln \Lambda = 71.15786$. On the basis of the test results we were able to reject the hypothesis of homogeneity of variances on each level of significance, since p-value was 0 (10.000 iterations).

## 5. TESTING FOR THE STRUCTURE OF THE HETEROGENEITY

The regression picture gave us strong suspicion on inhomogeneity of the data. The heterogeneity was recognized by the ELRH test (see [Stehlík (2006)]) and we have decided to study its structure. First, note that conducting the complete test against 2 component mixture (ELR2, see [Stehlík and Ososkov (2003)] and [Stehlík and Wagner (2009)]) is a complex tasks. Even on the assumption that size of one cluster is minimal 3 we have great number of possible latent cases, since heterogeneity is unobserved. Therefore we have decided to consider just ELR2 version of test against the fixed number of observations in both clusters (let us call them K and N-K). The evidence for that can be taken from the log-regression picture since we can see an upper sector with higher variance and lower sector having a smaller variance. Therefore, we derive now the exact likelihood ratio test for the homogeneity versus two cluster observed heterogeneity. The LR of the test of the hypothesis

$$H_0 : homogeneity\ versus\ H_1 :\ K\ and\ N-K\ heterogeneity$$

has the form

$$\lambda_N = \frac{N^N (y_1 + ... + y_K)^K (y_{K+1} + ... + y_N)^{N-K}}{K^K (N-K)^{N-K} (y_1 + ... + y_N)^N}. \qquad (1)$$

THEOREM 1. *Le $y_1, ..., y_N$ be i.i.d. according to the Exponential distribution with the unknown scale parameter $\sigma$, then the LR test statistics $-\ln \lambda_N$ where $\lambda_N$ is given by the formula (1) has the form*

$$-N \ln N + K \ln K + (N-K)\ln(N-K) - K \ln(\sum_{i=1}^{K} y_i) -$$

$$-(N-K)\ln(\sum_{i=K+1}^{N} y_i) + N \ln(\sum_{n=1}^{N} y_n)$$

*and it has the same distribution as the random variable $u_N$ given by*

$$-N \ln N + K \ln K + (N-K)\ln(N-K) - K \ln(\sum_{i=1}^{K} u_i) -$$

$$-(N-K)\ln(\sum_{i=K+1}^{N} u_i) + N \ln(\sum_{n=1}^{N} u_n)$$

*where $u_1, ..., u_N$ are iid according to exponential distribution with scale parameter $\sigma = 1$.*

For proof of Theorem 1 see Appendix. Application to the observed clustering: we have $K = 19, N = 27, \sum_{i=1}^{K} y_i = 3.434889, \sum_{i=K+1}^{N} y_i = 9.181511$ and thus getting $-\ln \Lambda = 10.85396959$. The corresponding p-value is 0, both for simulation based on 1000 and 10000 replications. Therefore we reject the hypothesis of homogeneity of the observations, which supports also formally our observation of heterogeneity.

## 6. RELATIONSHIP BETWEEN THE LONG-TERM DISEASE INCIDENCE AND THE TICK ACTIVITY IN 2008

Furthermore, we were interested if the data on tick host-seeking activity from one particular season may reflect the difference in long-term spatially specific TBE incidence. The epidemiological data (number of TBE cases per municipality and year) were available for a 8-year period 2001-2008. We have assumed a closer relationship between the tick activities in 2008 to epidemiological data from recent years, therefore we used weighted annual means with geometric progression to emphasize higher importance of the closer entries over the remote ones. The question we asked was whether there is a relation between a long-term risk, as a characteristic of a locality, and the actual host-seeking activity of ticks. The linear regression although significant (SSR/SST= 0.349344 (corrected), 0.213003 (uncorrected); p=0.003466) again did not correspond with the saturation character of the data. Therefore log regression was employed which seemed to be more appropriate (SSR/SST=0.345892 (corrected), 0.567293 (uncorrected), p= 0.000344) (see Figure 2). The regression curve still did not seem to represent the trend in the data well. The individual observations seemed to form two compact clusters, one above and one under the curve, therefore we decided to test the dependent variable for homogenity. Since the data had no normal distribution (p=0.01, Kolmogorov-Smirnov test; P=0.0001, ?2) but rather exponential (p=0.2, Kolmogorov-Smirnov test; P=0.136, ?2) or gamma (Kolmogorov-Smirnov d = 0.13532, p = n.s.; ?2, p = 0.24422) and the number of observation was relatively small we were forced to use an exact likelihood ratio test developed in [Stehlík (2006)]. Critical value was calculated according to the ELRH test and we got $-\ln \Lambda = 11.87688$. On the basis of the test results we were able to reject the hypothesis of homogeneity of variances on 0.005% level of significance (p=0.0017494; 100 000 iterations).

## 7. CONCLUSION AND DISCUSSION

The relationship between the actual host-seeking activity of tick populations and disease incidence seemed to be best represented by a log regression curve. Similar curve was observed in the regression of actual tick host-seeking activity and annual weighted mean incidence (2001-2008). Concerning the heterogeneity of the data, two clusters were identified by the means of ELR2 test modified for observed two cluster heterogeneity.

The important issue for further investigation shall be the construction of confidence bounds, based on result [Potocký and Van Ban (1992)]. Note, that by the means of optimal design theory, the lower bound information, $y(0) = 0$ should be used to get a better fits. The following issues may be of interest for further investigation: D-optimal designs (for estimation of parameters) and prediction designs, model selection designs and outlier detection rules.

**Acknowledgement**

Research was supported by Aktion Austria - Czech Republic, Stipendium and project 53p19. The data collections were supported from the European Regional Development Fund of the European Union.

## 8. APPENDIX

**Proof of Theorem 1** *Under null hypothesis is the sample $x_i/\sigma$ iid from Exponential distribution with scale parameter 1. The form of the likelihood ratio statistic can be obtained by the direct algebra.*

## REFERENCES

Kitron U. and Kazmierczak J.J. (1997). Spatial analysis of the distribution of Lyme disease in Wisconsin. Am. J. Epidemiol. 145 (6): 558-566.

Potocký, R.-Van Ban, T. Confidence regions in nonlinear regression models.Appl.Math. 37 (1992),29-39.

Stehlík, M. (2003) Distributions of exact tests in the exponential family. *Metrika* **57**, 145–164.

Stehlík M. (2006). Exact likelihood ratio scale and homogeneity testing of some loss processes, *Statistics & Probability Letters* **76**, 2006, 19-26.

Stehlík M. and Ososkov G.A. (2003). Efficient testing of the homogeneity, scale parameters and number of components in the Rayleigh mixture. JINR Rapid Communications E-11-2003-116.

Stehlík M. and Wagner H. (2009). Exact likelihood ratio testing for homogeneity of the exponential distribution, IFAS Res. Rep. 39.

Švec P., Hönig V., Daniel M., Danielová V., Grubhoffer L. (2009). Use of GIS for mapping of ticks and tick-borne pathogens in South Bohemia. Geografie-Sborník České geografické společnosti. 114(3): 157-68.

The Population and Housing Census 2001, Czech Statistical Office.

Václav Hönig, Libor Grubhoffer
Institute of Parasitology,
BC ASCR and Faculty of Science,
University of South Bohemia

Milan Stehlík
Department of Applied Statistics,
Johannes Kepler University in Linz,
e-mail: Milan.Stehlik@jku.at

Vlasta Danielová, Milan Daniel
National Institute of Public Health,
Prague

Pavel Švec
Institute of Geoinformatics VŠB,
Technical University of Ostrava

Received October 2009