

A statistical assessment of silica levels and lung cancer¹

Una valutazione statistica dei livelli di silicio e cancro al polmone

Ambrogio Fassina and Matteo Corradin

Dipartimento di Scienze Medico-Diagnostiche e Terapie Speciali,
Università degli Studi di Padova

Annamaria Guolo and Laura Ventura ²

Dipartimento di Scienze Statistiche,
Università degli Studi di Padova, e-mail: guolo@stat.unipd.it

Riassunto: Si presenta una analisi statistica su pazienti affetti e non affetti da cancro al polmone, allo scopo di confrontare i livelli di silicio presenti nel tessuto polmonare. Si utilizzano modelli di regressione di Poisson e la discussione si focalizza su aspetti metodologici legati a tali modelli, quali eteroschedasticità, sovradisersione, tecniche di diagnostica e predizione.

Keywords: Diagnostics, Heteroskedasticity, Overdispersion, Poisson count data.

1. Introduction

In lung pathology the role of silica is still debated at various levels of experimental, epidemiological and forensic evidence. To physicians and other health professionals it could be of relevance to know silica levels in lung tissue in order to better define cut-off limits of exposure. In this study, the focus is on comparing silica levels between subjects with or without lung cancer. The analysis takes into account a set of covariates related to risk factors. We focus on Poisson regression models, which allow to evaluate the difference in silica levels between lung cancer patients and non-cancer controls. Much of the statistical discussion is concerned with technical aspects of applying Poisson models (see Cameron and Trivedi, 1998) to this kind of data, focusing on issues such as heteroskedasticity, overdispersion, outliers, robustness, diagnostic and prediction.

2. Data used in this study

The data are provided by the Pathology Department of the University of Padova. They are divided into three groups. The first one (group A) is constituted by 25 patients with non-small cell lung cancer, who had no known history of workplace silica exposure. The second group (group B) is constituted by 20 patients without lung cancer, who deceased for extra-pulmonary diseases. Finally, a third group (group C) is constituted by only 2 patients with suspected workplace silica exposure, who were operated for lung adenocarcinoma and are inconsistent with the pattern of the rest of the observed data. For each subject the silica levels in lung tissues are quantified by considering the number of spots

¹Partially supported by a grant from MIUR, Italy.

²Address of correspondence: Via C. Battisti 241, 35121 Padova, Italy.

(NS), that is the number of silica particles, and the number of positive zones (NPZ), in which there is at least one spot. Measures are provided by using the Environmental Scanning Electron Microscopy technology. Moreover, for each subject additional information is collected about risk factors, such as age, gender and smoking habits.

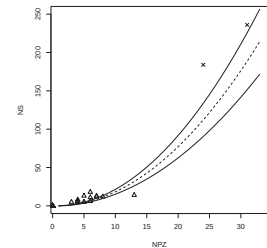
3. The regression analysis

The relationship between NPZ and NS were studied using Poisson-type regression models (McCullagh and Nelder, 1989; Davison, 2003). This class of models is needed for dealing with highly asymmetric and possible overdispersed count data and it allows to determine which risk factors are statistically significant. Moreover, the model is able to describe the values of the two patients belonging to group C, who stand as outliers.

A first observation in fitting a Poisson regression to the data is the large variation in NPZ and NS values, especially due to many zeros values. One standard way to deal with this issue is to use the square-root link function, instead of the canonical link function. Indeed, the square-root transformation is approximately variance-stabilizing for a Poisson random variable. Since the values in the exposed patients are clearly outliers, we omit them from the analysis and we predict their values based on the regression analysis fitted on the remaining patients.

Regression analyses were performed to determine which covariates were statistically significant. The only variable selected is NPZ, while all the other variables are not statistically significant. This is true in particular for the variables Group (p-value = 0.28), which identifies the normal and control groups, and Smoke (p-value = 0.89).

The estimated regression model is $NS^{1/2} = 0.24 + 0.45NPZ$, with residual deviance equal to 45.58 (df 43, p-value= 0.36). We have also considered other diagnostics for the fit of the model, in particular plotting techniques for the analysis of the residuals deviance and of measures of influence, such as Cook's and DFFITS statistics (Cook and Weisberg, 1982). All the diagnostics support the estimated regression model. Based on the fitted model, 95% prediction intervals were calculated for the values of NS.



The result is shown in the Figure on the right, where symbol Δ denotes normal and control patients, while symbol $*$ indicates exposed patients, predicted from the estimated model. Although the resulting model is essentially considered with descriptive aims, it can be employed to define threshold values in silica levels which discriminate noncancer controls, cancer cases and patients with suspected working exposition to particles.

References

- Cameron A. and Trivedi P. (1998) *Regression Analysis of Count Data*, Cambridge University Press.
- Cook R. and Weisberg S. (1982) *Residuals and Influence in Regression*, Chapman and Hall, New York.
- Davison A. (2003) *Statistical Models*, Cambridge University Press.
- McCullagh P. and Nelder J. (1989) *Generalized Linear Models*, Chapman and Hall, London.