Thesis for the Degree of Licentiate of Engineering

Quantifying environmental impact by log-normal regression modelling of accumulated exposure

Yulia Yurgens

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
Chalmers University of technology and Göteborg University
SE-412 96 Göteborg, Sweden, 2004

Quantifying environmental impact by log-normal regression modelling of accumulated exposure
YULIA YURGENS

ISSN 0347-2809/No. 2004:46
Department of Mathematical Statistics
Chalmers University of technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000

ABSTRACT

Statistical analysis of air-pollution levels measured by carry-on sensors is presented. The analysis aims at estimation of the average exposure to hazardous substances like toluene, benzene and xylene in different environments and at understanding the importance of general air-pollution, smoking habits, car traffic, and other potential sources of these substances for humans health.

The data on accumulated exposure are modelled as lognormal with linear regression on exposure times and other covariates.

A new element of our analysis is the use of untransformed raw data, i.e. directly measured exposures and not their logarithms as in previous studies. We argue that our approach is more correct because exposure effects of various hazardous substances accumulate, which is not possible to model in a linear model for the logarithms of data.

Comparing our approach with "conventional" models (regression of log exposure) we find the same significance for the analyzed data but parameter estimates differ both in values and interpretation, and only our modelling can be reasonably extrapolated to different exposure times.

Keywords: air pollution, carry-on sensors, asymptotic uncertainty, lognormal regression, parametric bootstrap.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. PREFACE

The word ”environment” can be used in various contexts. It can be working environment, bio-environment, outdoor environment, natural environment, etc. In the following, I will use the word environment as everything that surrounds us.

It is well known that mankind and environment are constantly and inevitably interacting entities. The man’s activity on Earth has a profound effect on environment as well as environment clearly affects our everyday’s life. Therefore any environmental problem has two sides. One is a problem arising in nature due to some man’s (industrial) activity and another is problems associated with nature hitting back. One can define environmental problem as undesired negative effects of mankind’s activity which are regarded as objective change and are experienced as a problem [1].

In general, there are four questions to be answered when dealing with environmental problems:

1) What has happened and what is happening now?

2) What will happen in the future and what happens if...?

3) What is acceptable?

4) What should or must we do?

Environmental problems are not always global and large, they can be quite small and invisible in the beginning. Nonetheless, even those small problems may have severe consequences in following years. It is utterly important to wake up public interest to these problems and show that they require serious attention and pressing actions for their solving. The existence of public interest to environmental problems usually ignites mass media which closely follow their development and supply more information to the public. It is also important that this circulation of information is fed by scientifically valid studies involving different models and hypotheses, observations and measurements, as well as scientific analysis and interpretation [1].

2. ENVIRONMENT AND HEALTH OF POPULATION

2.1 Measurement and assessment of state of health

What is health? According to WHO's definition, health is not only absence of illnesses and weakness but a complete physical, mental and social well-being. It is not easy to assess all components of health in view of many perceptions and opinions about what is the most important and what should be evaluated.

To measure and evaluate state of health of population and be able to judge on environmental impacts is one of the central tasks of environmental statistical research. Through systematic and alert examination of population's lifestyle and life quality, a reliable and timely analysis can be made to avoid rush decisions. It is therefore important that methods and standards that are used are well scrutinized and are reliable and trustworthy.

Problems with population's health are not only purely medical. They often mean a major social problem where different environmental aspects play an important role and where surveys involving studies on social structure, working life, and environment are hence necessary [2].

Health-related questions are very complex and opinions about what should be measured differ. It is therefore of no surprise that measurement methods and results are often unreliable and vary from source to source. This leads to a situation where the choice of measure (gauge) is largely dictated by a possibility of accumulating enough data and by access to developed methods of analysis. All this makes it clearly important to develop statistical research which can yield new more advanced and mathematically proven methods that could be successfully used in all scientific fields.

2.2 Air pollution and its effect on health

Clean air is of vital importance for health. During the last years, a large scale research is carried on that aims at mapping all air-born hazardous pollution that we exposure ourselves to in our everyday's life and at understanding the consequences of that exposure. Degradation of lung capacity, respiratory sickness and asthma are just a few of the maladies that are connected to air pollution in many studies.

There is a lack of knowledge about air quality with regard to health that leads to demands on further research on air pollution coming, say, from traffic and its negative impacts on health. Moreover, there are demands on measure and standard methods which can be used to follow the influence of different substances on population's health. Various hazardous substances that are measured in environmental studies should be considered as indicators for air pollution even if quantitatively the effect of that pollution is not known yet [3].

In order to estimate exposure one uses data either from stationary sensors or from measurements obtained with the aid of sensors put on individuum (person-carried) or theoretically derived data. The numerical models are good when dangerous substances just potentially can cause complex air pollution, while person-carried data are better when measured substances have direct impact on health.

In Sweden, the stationary measurements of cancerogenic substances are usually made at relatively high levels from the ground, often the sensors are placed at the roof of houses [4]. Since particular risks for humans depend on personal (local) exposures rather than on the average background concentration measured by the stationary sensors, the former is better suited for working out guiding principles in risk assessments and political decisions. Person-carried data gives much better estimate of individual accumulated exposure to a hazardous substance. It represents the basis of recent health-related statistical research.

The measurements of personal exposures are heavy on resources while it can also give results that are different from what stationary measurements show.

3. GENERAL INFORMATION ABOUT EXPOSURE DATA

Exposure data commonly have a skewed distribution with lots of low-value points and a few observations taking high values, see Fig. 1.

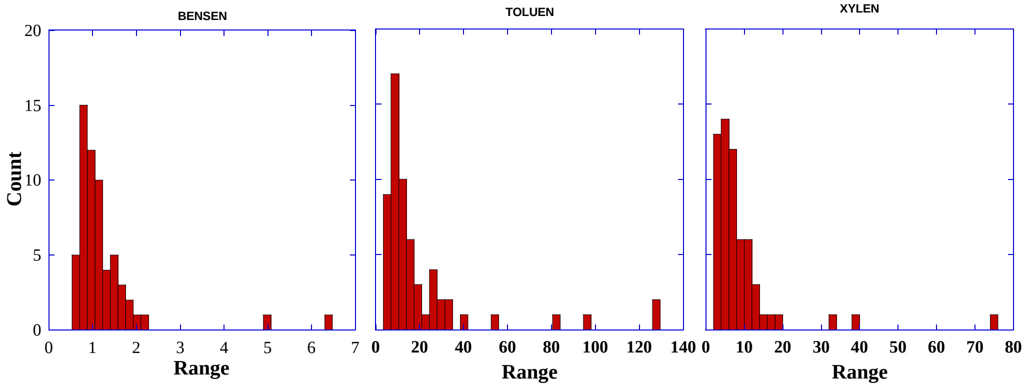

Fig. 1: Observed exposure data on benzene, toluene, and xylene.

It is convenient to use the so called log-normal distribution for analysis of such data. See for example the modelling in [5, 6, 7]. In such a distribution, it is assumed that the logarithm of the statistical variable Y, $\ln Y$, is normally distributed, i.e. Y has an exponential functional form, $Y = \exp(Z)$, where Z is normal. In this case the large and irregular variations in exposure data become smoother in the logarithmic scale and one can apply well-developed theories of normal distribution for statistical analysis of the data. In principle, a number of other known distributions (Gamma, Weibuell, etc.) can be used as well.

In the majority of studies on exposure data, one has a background information in form of covariates. Then, one traditionally can apply a linear regression analysis to the log-normally distributed exposure data [5, 6, 7, 8, 9].

Since linear regression theory is usually derived for normally distributed variables, it is natural as a first approach to apply this to the logarithms of the exposure data as it was done in the references above.

The most interesting covariates refer to as to how long an individual is exposed to different environments. These exposure times should then have linear effect on exposure measurements because two times longer exposure time should double the contribution. This implies that the exposure times cannot be additive when logarithms of statistical variables are used. This is therefore a shortcoming of the traditional approach using linear regression on the logarithmic exposure and makes the interpretation of estimated regression coefficients less obvious (as an average linear effect on the logarithms).

Other variables of different types can also affect linearly on exposure. In some cases, multiplicative effect of variables can be anticipated that would justify linear response in logarithmic scale. Under certain conditions, however, linear and multiplicative effects can approximate one another.

4. SOME RELATIONS BETWEEN LOG-NORMAL AND NORMAL DISTRIBUTIONS.

4.1 One-dimensional case

Let Y be a log-normally distributed variable with characteristic parameters (μ, σ), modelling for instance, an accumulated exposure to a certain substance. This implies that $Z \equiv \ln Y$ is $N(\mu, \sigma)$. It follows from above that $(\mu, \sigma) = (\mu_z, \sigma_z)$ are the characteristic parameters of the lognormal distribution.

In order to distinguish between expectation and variance in the original and logarithmic scales, we introduce different notations μ_z, σ_z and μ_Y, σ_Y, respectively, and use μ_z, σ_z as average and variance of normally distributed $Z = \ln Y$, while $\mu_Y = E(Y)$, $\sigma_Y^2 = Var(Y)$.

For $Z = \ln Y$, we have the frequency function of the normal distribution:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\ \sigma_z} \exp\left[-\frac{(z-\mu_z)^2}{2\sigma_z^2}\right]. \tag{4.1}$$

The density function $f_Y(y)$ for $y > 0$ is [10]:

$$\begin{aligned} f_Y(y) = \frac{d}{dy} P(e^Z \leqslant y) = \frac{d}{dy} F_Z(\ln y) = \frac{1}{y} f_Z(\ln y) = \\ \frac{1}{y} \frac{1}{\sqrt{2\pi}\ \sigma_z} \exp\left[-\frac{(\ln y - \mu_z)^2}{2\sigma_z^2}\right]. \end{aligned} \tag{4.2}$$

The expected value and the variance for the lognormal are most easily computed in the normal distribution using $Y = \exp Z$. This gives the well known results:

Tab. 1: Comparison between log-normal and normal distributions [11]

	Log-normal scale	Original scale
Mean	$\mu_z = \ln \mu_y - \frac{\sigma_z^2}{2}$	$\mu_y = \exp(\mu_z + \sigma_z^2/2)$
Median	$\tilde{\mu}_z = \mu_z$	$\tilde{\mu}_y = \exp(\mu_z)$
Mean/Median	$\mu_z/\tilde{\mu}_z = 1$	$\mu_y/\tilde{\mu}_y = \exp(\sigma_z^2/2)$
Variance	σ_z^2	$\sigma_y^2 = (\exp(\sigma_z^2) - 1)\exp(2\mu_z + \sigma_z^2)$

$$\mu_Y = E[Y] = \int_{-\infty}^{\infty} e^z \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left[-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2\right] dz = \exp\left(\mu_z + \frac{\sigma_z^2}{2}\right). \tag{4.3}$$

$$E[Y^k] = \int_{-\infty}^{\infty} e^{kz} \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left[-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2\right] dz = \exp\left(k\mu_z + \frac{\sigma_z^2 k^2}{2}\right). \tag{4.4}$$

and since $Var(Y) = E[Y^2] - \mu_Y^2$, we have the relations shown in Table 1.

4.2 Two-dimensional case

When more than one substance is measured on the same person or when measurements are repeated on the same individual we can expect such data to be dependent. A natural approach to dependent log-normal distribution is to start from the multivariate normal distribution.

A particular feature of the data we will study later on is that there were two observations of the same type made on some individuals.

There is therefore a possible interdependence between repeated observations on the same person. Now, the natural approach to the case of *two* measurements per person is to consider observations on different persons as independent and having a two-dimensional log-normal distribution for the two data on the same person. With a relevant covariance structure, the extension to more than two dimensions is straightforward.

For a vector $Y = \left(\begin{array}{c} Y_1 \\ Y_2 \end{array} \right)$, we define $\ln Y$ as a component-wise function

$$\ln Y = \left(\begin{array}{c} \ln Y_1 \\ \ln Y_2 \end{array} \right).$$

In the Z-scale, $Z = \ln Y$ and is modelled as $N(\mu, C)$ which means that $Z = \left(\begin{array}{c} Z_1 \\ Z_2 \end{array} \right)$ has the two-dimensional normal distribution. Since the two measurements were made in the same way both times, it is natural to assume the same standard deviations: $\sigma_1 = \sigma_2 = \sigma_z$. The expectation μ and the covariance matrix C are:

$$\mu = \left(\begin{array}{c} \mu_{z_1} \\ \mu_{z_2} \end{array} \right) \tag{4.5}$$

$$C = \left(\begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array} \right) = \left(\begin{array}{cc} \sigma_z^2 & \sigma_z^2\rho \\ \sigma_z^2\rho & \sigma_z^2 \end{array} \right) = \sigma_z^2 \left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array} \right), \tag{4.6}$$

where $\rho = \rho(Z_1, Z_2) = Cov(Z_1, Z_2)/\sigma_z^2$.

Z has the following density function:

$$f_Z(z_1, z_2) = \frac{1}{2\pi\sqrt{\det(C)}} \ \exp\left[-\frac{1}{2}(z_1 - \mu_{z_1}, z_2 - \mu_{z_2}) \ C^{-1} \left(\begin{array}{c} z_1 - \mu_{z_1} \\ z_2 - \mu_{z_2} \end{array} \right) \right] \tag{4.7}$$

which can be rewritten in the following form:

$$\begin{aligned} f_Z(z_1, z_2) = \frac{1}{2\pi\sigma_z^2\sqrt{1-\rho^2}} \ \exp\Big[- \frac{1}{2\sigma_z^2(1-\rho^2)}\Big((z_1 - \mu_{z_1})^2 + \\ (z_2 - \mu_{z_2})^2 - 2\rho(z_1 - \mu_{z_1})(z_2 - \mu_{z_2})\Big)\Big] \end{aligned} \tag{4.8}$$

In the Y-scale, the density function $f_Y(y_1, y_2)$ for $y_1, y_2 > 0$ is:

$$f_Y(y_1, y_2) = f_Z(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|. \tag{4.9}$$

Using $z_1 = \ln y_1$ and $z_2 = \ln y_2$, we get:

$$\frac{d(z_1, z_2)}{d(y_1, y_2)} = \begin{vmatrix} \frac{dz_1}{dy_1} & \frac{dz_1}{dy_2} \\ \frac{dz_2}{dy_1} & \frac{dz_2}{dy_2} \end{vmatrix} = \frac{1}{y_1} \cdot \frac{1}{y_2}$$

and

$$f_Y(y_1, y_2) = \frac{1}{2\pi\sigma_z^2\sqrt{1-\rho^2}} \exp\Big\{ - \frac{1}{2\sigma_z^2\sqrt{1-\rho^2}} \Big[(\ln y_1 - \mu_{z_1})^2 + \\ +(\ln y_2 - \mu_{z_2})^2 - 2\rho(\ln y_1 - \mu_{z_1})(\ln y_2 - \mu_{z_2}) \Big] \Big\} \cdot \frac{1}{y_1} \cdot \frac{1}{y_2} \quad (4.10)$$

5. REGRESSION IN LOG-NORMAL DISTRIBUTION.

Suppose that we have covariates $x_{i,k}$, $k = 1 \dots K$, associated to each observation Y_i. Also let the effects of the covariates be linear in the expected value of Y. This is the situation when exposure times in different environments sum up and yield the measured accumulated exposure. In Sec. 8 this model is compared to an alternative regression model.

In a one-dimensional model with fixed effects and non-stochastic covariates, we set

$$E[Y_i] = \beta_0 + \sum_k x_{ik}\beta_k, \tag{5.1}$$

where k refers to environmental group or factor.

The model is now such that Y is log-normally distributed with constant σ_z^2 for which the following relation can be written:

$$E[Y_i] = x_i'\beta = e^{(\sigma_z^2/2+\mu_{z_i})} \tag{5.2}$$

which gives, as in Table 1:

$$\mu_{z_i} = \ln(x_i'\beta) - \sigma_z^2/2, \tag{5.3}$$

where x_i' denotes the vector of covariates for observation i.

The frequency function of Z_i can now be written explicitly as

$$f_{Z_i}(z) = \frac{1}{\sqrt{2\pi}\ \sigma_z} \exp\left[-\frac{(z - (\ln(x_i'\beta) - \sigma_z^2/2))^2}{2\sigma_z^2}\right] \tag{5.4}$$

For the two-dimensional case, the corresponding repetition on each component have the form described by Eq. (4.8).

6. MAXIMUM LIKELIHOOD ESTIMATION

6.1 One-dimensional case

6.1.1 One-dimensional data without covariates

Let $(y_1, y_2 \dots y_n)$ be a random sample on Y which is lognormal(μ, σ). In the Z-scale, observations $(z_1, ...z_n)$, where $z_i = \ln(y_i)$ give the likelihood:

$$L_1(\mu_z, \sigma_z) = \frac{1}{\sigma_z^n (2\pi)^{n/2}} \ \exp\left(-\frac{1}{2\sigma_z^2}\sum_{i=1}^{n}(z_i - \mu_{z_i})^2\right). \tag{6.1}$$

We have also the following maximum likelihood estimations:

$$\widehat{\mu}_z = \frac{1}{n}\sum z_i = \frac{1}{n}\sum \ln(y_i) \tag{6.2}$$

$$\widehat{\sigma}_z = \sqrt{\frac{1}{n}\sum(z_i - \overline{z})^2} \tag{6.3}$$

The likelihood can be re-written in the Y-scale:

$$L_1(\mu_z, \sigma_z) = \frac{1}{\sigma_z^n (2\pi)^{n/2}} \ \exp\left(-\frac{1}{2\sigma_z^2}\sum_{i=1}^{n}(\ln y_i - \mu_{z_i})^2\right) \prod_{i=1}^{n}\left(\frac{1}{y_i}\right) \tag{6.4}$$

with corresponding estimations:

$$\widehat{\mu}_z = \frac{1}{n}\sum(\ln y_i) \tag{6.5}$$

$$\widehat{\sigma}_z = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\ln y_i - \frac{1}{n}\sum(\ln y_i)\right)^2} \tag{6.6}$$

Comparing equations 6.1 and 6.4 we see that the latter is multiplied by a factor $\prod 1/y_i$ which does not depend on the parameters μ_z and σ_z, but only on data y_i.

We see that one gets the equivalent formulas for likelihood estimations in both the original- and logarithmic scales, and by using the relations in Table 1 we find the corresponding estimates of $E(Y)$ and $Var(Y)$.

6.1.2 One-dimensional data with regression

In order to estimate the parameters when regression is involved, we could find a maximum of the logarithmic likelihood using of Newton-Raphson iteration scheme.

If $\theta = \binom{\beta}{\sigma_z}$ denotes the parameter vector, $\frac{\partial \ln L}{\partial \theta}$ and $\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}$, the vector and matrix of the first and second derivatives and $\theta^{(0)}$ an initial guess, the Newton-Raphson method uses

$$\theta^{(n+1)} = \theta^{(n)} - \left(\frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial \ln L}{\partial \theta}, \tag{6.7}$$

where $\theta^{(n)}$ is inserted in the derivatives. The equations converge without problems for our data example if $\theta^{(0)}$ is reasonably selected.

For that, we would need the following expression:

$$\begin{aligned} \ln L_1(\beta, \sigma_z) &= -\frac{n}{2} \ln(2\pi) - n \ln \sigma_z - \frac{1}{2\sigma_z^2} \sum (\ln y_i - \mu_{z_i})^2 - \sum_{i=1}^{n} \ln y_i \\ &= -\frac{n}{2} \ln(2\pi) - n \ln \sigma_z - \frac{1}{2\sigma_z^2} \sum (\ln y_i)^2 - \frac{1}{2\sigma_z^2} \sum \ln(x_i'\beta)^2 + \\ &+\frac{1}{2} \sum \ln(x_i'\beta) - \frac{n\sigma_z^2}{8} + \frac{1}{\sigma_z^2} \sum (\ln y_i) \ln(x_i'\beta) - \frac{1}{2} \sum \ln y_i - \sum \ln y_i \end{aligned} \tag{6.8}$$

where μ_{z_i} is given by Eq. (5.3), and its derivatives:

$$\frac{\partial \ln L_1}{\partial \beta_i} = \frac{1}{\sigma_z^2}\sum_{i=1}^{n}\frac{x_{ij}}{\sum_j x_{ij}\beta_j}\left[\ln y_i - \ln(\sum_j x_{ij}\beta_j) + \frac{\sigma_z^2}{2}\right], \tag{6.9}$$

$$\frac{\partial^2 \ln L_1}{\partial \beta_i \beta_k} = \frac{-1}{\sigma_z^2}\sum_{i=1}^{n}\left[\frac{x_{ik}x_{im}}{(\sum_j x_{ij}\beta_j)^2}\left(1 - \ln(\sum_j x_{ij}\beta_j) + \ln y_i + \frac{\sigma_z^2}{2}\right)\right], \tag{6.10}$$

$$\frac{\partial \ln L_1}{\partial \sigma_z} = -\frac{n}{\sigma_z} + \frac{1}{\sigma_z^3}\sum_{i=1}^{n}\left(\ln y_i - \ln(x_i'\beta)\right)^2 - \frac{n\sigma_z}{4}, \tag{6.11}$$

$$\frac{\partial^2 \ln L_1}{\partial \sigma_z \partial \beta} = -\frac{2}{\sigma_z^3}\sum_{i=1}^{n}\frac{x_{ik}}{\sum_j x_{ij}\beta_j}\left(\ln y_i - \ln(x_i'\beta)\right), \tag{6.12}$$

$$\frac{\partial^2 \ln L_1}{\partial \sigma_z^2} = \frac{n}{\sigma_z^2} - \frac{3}{\sigma_z^4}\sum_{i=1}^{n}\left(\ln y_i - \ln(x_i'\beta)\right)^2 - \frac{n}{4}. \tag{6.13}$$

6.2 Two-dimensional case

We will now extend the solution to bivariate data with the same variance σ_z^2 of both components in the Z-scale. This takes only one more parameter, the correlation ρ.

In the Z-scale:

For two measurements on the same person, the likelihood function can be written, taking $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \ln y_1 \\ \ln y_2 \end{pmatrix}$ in the following form:

$$\begin{aligned} & L_2(\beta, \sigma_z, \rho) = f_z(z_1, z_2) = \\ & = \frac{1}{2\pi\sqrt{det(C)}}\exp\left\{-\frac{1}{2}(z_1 - \mu_{z_1}, z_2 - \mu_{z_2})\ C^{-1}\begin{pmatrix} z_1 - \mu_{z_1} \\ z_2 - \mu_{z_2}\end{pmatrix}\right\} = \\ & = \frac{1}{2\pi\sigma_z^2\sqrt{1-\rho^2}}\cdot\exp\Big\{ - \frac{1}{2\sigma_z^2(1-\rho^2)}\Big[(\ln y_1 - \mu_{z_1})^2 + (\ln y_2 - \mu_{z_2})^2 \\ & - 2\rho(\ln y_1 - \mu_{z_1})(\ln y_2 - \mu_{z_2})\Big]\Big\}, \end{aligned} \tag{6.14}$$

where $\mu_{z_1} = \ln\left(\sum_j x_{1,j}\beta_j\right) - \frac{\sigma_z^2}{2}$
and $\mu_{z_2} = \ln\left(\sum_j x_{2,j}\beta_j\right) - \frac{\sigma_z^2}{2}$.

Taking the logarithm of Eq. (6.14) gives:

$$\ln L_2(\beta, \sigma_z, \rho) = \ln\left(\frac{1}{2\pi\sigma_z^2\sqrt{1-\rho^2}}\right) - \frac{1}{2\sigma_z^2(1-\rho^2)}\Big[(\ln y_1 - \mu_{z_1})^2 + (\ln y_2 - \mu_{z_2})^2 - 2\rho(\ln y_1 - \mu_{z_1})(\ln y_2 - \mu_{z_2})\Big]. \tag{6.15}$$

In case of m independent individuals with repeated measurements on each, the likelihood function will be a product of $L_2(\beta, \sigma_z, \rho)$ for each individual and we get the corresponding sum of terms, Eq. (6.15) in the log likelihood.

In the Y-scale, the likelihood function is according to (4.10)

$$L_2(\beta, \sigma_z, \rho) = f_z(\ln y_1, \ln y_2) \cdot \frac{1}{y_1 \cdot y_2}. \tag{6.16}$$

$$\begin{aligned}\ln L_2(\beta, \sigma_z, \rho) &= \ln\left[f_z(\ln y_1, \ln y_2) \cdot \frac{1}{y_1 \cdot y_2}\right] = \\ &= \ln[f_z(\ln y_1, \ln y_2)] - \ln y_1 - \ln y_2\end{aligned} \tag{6.17}$$

It is seen that the likelihood function in the Y scale is different from the one in the Z scale by the factor $1/(y_1 \cdot y_2)$ that only depends on data and does not depend on β, σ_z, or ρ. With subsequently taking logarithm and derivative over β, σ_z, and ρ this factor disappears.

As in the one-dimensional case, we use the Newton-Raphson iteration scheme for estimating parameters. The corresponding formulas for the first- and second derivatives, $\frac{\partial \ln L_2}{\partial \sigma_z}$, $\frac{\partial \ln L_2}{\partial \beta_j}$, $\frac{\partial \ln L_2}{\partial \rho}$, $\frac{\partial^2 \ln L_2}{\partial \beta_i \partial \beta_j}$, $\frac{\partial^2 \ln L_2}{\partial \sigma_z \partial \beta_j}$, $\frac{\partial^2 \ln L_2}{\partial \rho^2}$, $\frac{\partial^2 \ln L_2}{\partial \sigma_z^2}$, $\frac{\partial^2 \ln L_2}{\partial \sigma_z \partial \rho}$, $\frac{\partial^2 \ln L_2}{\partial \beta_j \partial \rho}$, can be found in the Appendix 13.1.

The first and second derivatives are needed for both estimation of parameters and for estimating their variances.

6.3 Combination of one- and two-dimensional cases

In our case we have a mixture of individuals with single and with repeated measurements and we can describe the data in the following manner.

Y data:

$$\underbrace{y_1, y_2, y_3 \cdots y_{n_1}}_{\text{one-dimensional}}, \underbrace{y_{n_1+1}, y_{n_1+2}, y_{n_1+3}, \cdots y_{n_1+n_2}}_{\text{two-dimensional}}$$

In the Z scale the same data would look like:

$$\underbrace{z_1 = \ln y_1, \dots z_n = \ln y_{n_1};}_{\text{one}-\text{dimensional}} \underbrace{z_{n_1+1} = \begin{pmatrix} \ln y_{n_1+1,1} \\ \ln y_{n_1+1,2} \end{pmatrix}, \dots z_{n_1+n_2} = \begin{pmatrix} \ln y_{n_1+n_2,1} \\ \ln y_{n_1+n_2,2} \end{pmatrix}}_{\text{two}-\text{dimensional}}$$

In both cases, the covariances can be denoted as X-data and take the form:

$$\underbrace{x'_1, \dots x'_{n_1}}_{\text{one}-\text{dimensional;}} \underbrace{x_{n_1+1} = \begin{pmatrix} x'_{n_1+1,1} \\ x'_{n_1+1,2} \end{pmatrix}, \dots x_{n_1+n_2} = \begin{pmatrix} x'_{n_1+n_2,1} \\ x'_{n_1+n_2,2} \end{pmatrix}}_{\text{two}-\text{dimensional}},$$

where x' denotes covariance vectors.

Individuals on which measurements were performed only once are independent from the individuals on which repeated measurements were done. One therefore can consider these two cases as independent and can merge the corresponding likelihood functions:

$$L(\beta, \sigma_z, \rho) = \prod_{i=1}^{n_1} L_{1i}(\beta, \sigma_z) \prod_{i=n_1+1}^{n_1+n_2} L_{2i}(\beta, \sigma_z, \rho) \tag{6.18}$$

when L_{1i}, L_{2i} are the one-dimensional respective two-dimensional functions of data $Y_i(Z_i)$.

7. ESTIMATION RELIABILITY

For a large random sample, there is strong argument in favor of the ML method. The following properties are valid for the ML-estimations $\widehat{\Theta}_n$ of Θ based on the complete sample $x_1, ...x_n$ [12]:
1) Under some conditions on the distribution, we have that $\widehat{\Theta}_n$ is consistent;
2) $\sqrt{n}(\widehat{\Theta}_n - \Theta)$ is asymptotically normally distributed $N(0, V)$;
3) The variance or covariance matrix V in the asymptotical distribution is optimal, i.e. equivalent to Cramer-Raos lower bound for (asymptotically) unbiased estimation variance [16].

The inverse of V, the information matrix $D = V^{-1}$, has elements:

$$d_{i,j} = \int -\frac{\partial^2 \ln f(x, \Theta)}{\partial \Theta_i \partial \Theta_j} f(x, \Theta) \mathrm{d}x \tag{7.1}$$

Confidence intervals with approximate confidence degree $(1 - \alpha)$ we take from the approximate normal distribution for the parameter estimations where the asymptotic relation gives:

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) \approx N(0, V) \tag{7.2}$$

$$(\widehat{\Theta}_n - \Theta) \approx N(0, \frac{1}{n} V), \tag{7.3}$$

where $\frac{1}{n}V$ will be estimated as the inverse of the matrix $-\frac{\partial^2 \ln L}{\partial \theta\, \partial \theta'}$. Thus for component i

$$\Theta_i = \widehat{\Theta}_{ni} \pm a_{1-\alpha/2} \sqrt{\frac{1}{n}\, \hat{v_{ii}}}, \tag{7.4}$$

where $\frac{1}{n}\, \hat{v_{ii}}$ is the i-th diagonal term of the estimate

$$\frac{1}{n} \hat{V} = \left(-\frac{\partial^2 \ln L}{\partial \theta\, \partial \theta'} \right)^{-1} \tag{7.5}$$

taken in the estimated parameter point, and $a_{1-\alpha/2}$ is the percentile in the $N(0, 1)$-distribution.

In Sec. 10.1 we compute these intervals for our data case. With 60 data from 40 individuals, we can not assume that the asymptotic analysis assures good accuracy. We will therefore check their validity by simulations in Sec. 10.2

8. COMPARISON BETWEEN REGRESSION BEFORE AND AFTER TAKING LOGARITHMS OF DATA

To compare our modelling with the "traditional" one, we look at regression in the original scale (Model A) and in the logarithm scale (Model B).

1) **A simple regression in the one-dimensional case:**

Model A: The model in Eq. (5.1) gives

$$Z = \ln Y \sim N\left[\ln(x'\beta) - \frac{\sigma_z^2}{2}, \sigma_z^2\right]$$

with $E(Y) = x'\beta$

Assuming that $x'\beta = \beta_0 + \beta_1(x_i - \bar{x})$ we can write

$$Z_i \sim N\left[\ln\left(\beta_0 + \beta_1(x_i - \bar{x})\right) - \frac{\sigma_z^2}{2}, \sigma_z^2\right].$$

Estimates of β_0, β_1, and σ_z are found iteratively.

Model B: The "traditional" approach is instead

$$Z = \ln Y \sim N\left[x'b, \sigma_z^2\right],$$

with the same data size as for model A. This means that

$$Z_i = b_0 + b_1(x_i - \bar{x}) + \epsilon,$$

where $\epsilon \sim N(0, \sigma_z^2)$.

For $z_i = \ln y_i$ we have the following two estimates of regression coefficients in model B:

$$\hat{b}_0 = \bar{z} \tag{8.1}$$

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})z_i}{\sum(x_i - \bar{x})^2} \tag{8.2}$$

and if we assume that model A is true, then

$$E(\hat{b}_0) = \frac{1}{n}\sum\left[\ln\left(\beta_0+\beta_1(x_i-\bar{x})\right)-\frac{\sigma_z^2}{2}\right] \tag{8.3}$$

$$\begin{aligned} E(\hat{b}_1) = & \frac{\sum\left[(x_i-\bar{x})\left(\ln\left(\beta_0+\beta_1(x_i-\bar{x})\right)-\sigma_z^2/2\right)\right]}{\sum(x_i-\bar{x})^2} = \\ & \frac{\sum\left[(x_i-\bar{x})\ln\left(\beta_0+\beta_1(x_i-\bar{x})\right)\right]}{\sum(x_i-\bar{x})^2}. \end{aligned} \tag{8.4}$$

The expression for $E[\widehat{b_1}]$ is the same as if one made a linear least-squares fit (regression) to $\ln(\beta_0+\beta_1(x_i-\bar{x}))$ for the real data points. Since the logarithm function is non-linear and its derivative is unconfined, the value of $E[b_1]$ depends on the size of the expression $\beta_0+\beta_1(x_i-\bar{x})$ and especially on the mean level β_0. This means that there is no general relation between the sizes of $E[\widehat{b_1}]$ and β_1.

We can compare both the models in another way. If variations in x_i are small, one can make a series expansion of the logarithm and use the linear term only:

$$\ln\left(\beta_0+\beta_1(x_i-\bar{x})\right) = \ln(\beta_0)+\ln\left(1+\frac{\beta_1}{\beta_0}(x_i-\bar{x})\right) \approx \ln(\beta_0)+\frac{\beta_1}{\beta_0}(x_i-\bar{x})$$

If we now assume that we have a situation where model B is true but model A is fitted then the true structure $Z_i \sim N(b_0+b_1(x_i-\bar{x}), \sigma_z^2$ is analyzed as $Z_i \sim N(\ln(b_0+b_1(x_i-\bar{x}))-\sigma_z^2/2, \sigma_z^2$.

Since the likelihood function is a smooth function (has continuous derivatives and apparently has only one maximum), maximum-likelihood estimates for small variations should be very close to estimates using the above written series with only the linear term retained. This gives us a model for usual linear regression (which is model B) but with

$$\begin{aligned} b_0 &= \ln\beta_0-\sigma_z^2/2, \text{ and} \\ b_1 &= \beta_1/\beta_0 \end{aligned}$$

and shows how the models can approximate each other for small x-variations.

2) **The multiple regression.**

We continue to discuss the two estimates in the multiple-regression case with univariate Y-data, but this time under the assumption that model A is true. Here we use vector notations for Z, Y, β and let X be a matrix.

Model A:

$$Z = \ln Y \sim N(\ln X\beta - \sigma_z^2/2, \sigma_z^2) \tag{8.5}$$

$$\begin{gathered} \text{with } E[Y] = X\beta; \\ n \text{ observations;} \\ m \text{ covariates;} \\ \beta = (\beta_0 \ \ \tilde{\beta}\ ')', \text{ with } \tilde{\beta} = (\beta_1, \ldots \beta_m)'. \end{gathered}$$

Again, estimates are found iteratively and closed expressions do not exist.

Model B:

$$Z = \ln Y \sim N(Xb, \sigma_z^2)$$

$$\text{with } b = (b_0 \ \ \tilde{b}')'.$$

The relation between estimates in models A and B applied to the same data can be most easily discussed if the covariates are centered so that the mean is subtracted out of every column. Let therefore

$$X = \begin{pmatrix} \mathbf{1} & \widetilde{X} \end{pmatrix}$$

where $\sum_{i=1}^{n} \tilde{x}_{i,j} = 0, \ j = 1, 2, \ldots m$ and where the first column is a vector of ones.

In model B we have

$$Z = \ln Y = Xb + \nu, \quad \nu \sim N(0, \sigma_z^2 I),$$

where I is the unity matrix.

The estimate of b is

$$\hat{b} = (X'X)^{-1}X'Z.$$

Now, if model A is true, the expected value of Z is $\ln(X\beta) - \sigma_z^2/2$ and

$$E[\hat{b}] = (X'X)^{-1}X'\left(\ln(X\beta) - \frac{\sigma_z^2}{2}\right).$$

If the variation in $X\beta$ is relatively small compared to the constant β_0 (in the structure with centered covariates) we have

$$\ln(X\beta) = \ln(\beta_0\ \mathbf{1} + \widetilde{X}\tilde{\beta}) = \ln\beta_0 + \ln(\mathbf{1} + \widetilde{X}\tilde{\beta}/\beta_0)$$

And since $\ln(1+x) \approx x$ for small x, this gives the approximation

$$E[\hat{b}] \approx (X'X)^{-1}X' \left[(\ln\beta_0 - \frac{\sigma_z^2}{2})\mathbf{1} + \widetilde{X}\frac{\tilde{\beta}}{\beta_0} \right].$$

Since

$$X'X = \begin{pmatrix} \mathbf{1}' \\ \widetilde{X}' \end{pmatrix} \begin{pmatrix} \mathbf{1} \; \widetilde{X} \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \widetilde{X}'\widetilde{X} \end{pmatrix}$$

has the inverse

$$(X'X)^{-1} = \begin{pmatrix} 1/n & 0 \\ 0 & (\widetilde{X}'\widetilde{X})^{-1} \end{pmatrix}$$

and writing a for $(\ln\beta_0 - \sigma_z^2/2)$

$$X'\left(a\mathbf{1} + \widetilde{X}\tilde{\beta}/\beta_0\right) = \begin{pmatrix} \mathbf{1}' \\ \widetilde{X}' \end{pmatrix} \left(a\mathbf{1} + \widetilde{X}\tilde{\beta}/\beta_0\right) = \begin{pmatrix} na \\ \widetilde{X}'\widetilde{X}\tilde{\beta}/\beta_0 \end{pmatrix}.$$

We get

$$E[\hat{b}_0] \approx \ln\beta_0 - \frac{\sigma_z^2}{2} \tag{8.6}$$

$$E[\hat{\tilde{b}}] \approx \tilde{\beta}/\beta_0 \tag{8.7}$$

The centering of X is important for the correct size of β_0 (b_0) and for giving variations around zero. If non-centered covariates are used, the estimates of $\tilde{\beta}$ and $\tilde{b}$ will not change , only β_0 and b_0 are transformed and the corresponding relation becomes

$$E[\tilde{b}_j] \approx \beta_j/(\beta_0 + \sum \beta_k \bar{x}_{.k}), \;\; j = 1, 2, \ldots m \tag{8.8}$$

Using this relation we must remember that it is approximate and only concerns the expected values. The actual estimates will have random errors added to their expected values.

9. ANALYSIS OF DATA

9.1 Introduction

Today, large quantities of potentially cancer-promoting substances are used by the society. It is highly important to know if these substances adversely affect our health. It is also important to know particular doses of those substances that a person exposes him/herself to. Measurements of their levels made on stationary stations are not sufficiently informative since they give average values at the stations while local (time and place) peak values of the concentrations of these substances can be much higher. We have therefore adopted data from carry-on instruments for accumulated exposure.

In Gothenburg, 40 persons aging from 20 to 50 years drawn at random from a given population agreed to take part in measurements where they would carry small sensors sensitive to benzene, xylene, and toluene during a week. The individuals were supposed to regularly fill special diaries and questionnaires regarding their activity or unhealthy habits (smoking) during the week. This information supplemented the measurements.

One half of those persons were involved in the measurements twice. We have applied our modelling to a data set collected by the department of Occupational and Environmental Medicine, see Ref. [13].

We first show results for an univariate analysis where we only use the first measurement for individuals with repeated data. The analysis is based on the univariate likelihood L_1 (see Eq. (6.4)) and uncertainties of estimates are computed by asymptotic theory as in Sec. 7. Then we go on to a mixture of univariate and bivariate data and show the corresponding analysis based on the likelihood L_2 (see Eq. (6.14))

The contributions to measured amounts of benzene, toluene, and xylene due to the variables in the list below (potential dangerous environments) were estimated with help of linear and lognormal regressions.

9.2 List of covariate information

1 - house-heating method ((0) for electricity and central heating, (1) for oil)
2 - whether car is parked in a garage inside house (1-yes, 0-no)
3 - smoking (1-yes, 0-no)
4 - passive smoking, total hours of exposure
5 - filling up gasoline tank of a car
6 - hours spent in intensive traffic
7 - total time of going by car or a bus
8 - exposure to benzin vapors/car-exhaust gases professionally (1-yes, 0-no)
9 - the same during free time (1-yes, 0-no)
10 - total time being indoor (not home)
11 - total time being outdoor
12 - estimated home exposure.

Here the estimated home exposure is $U\ T_1/T_2$, where
U is the accumulated exposure in a stationary measurement device in the bedroom;
T_1 is the time at home with carry-on device during the measurement period;
T_2 is the measurement time of the stationary device.

For repeated date the stationary device was only used once and U/T_2 is the same both times, only T_1 different.

The set of covariates are well behaved. Correlations between the columns never exceed 0.65 in absolute value and during the estimation they were all centered by subtraction of their mean values.

9.3 Results

The analysis of data was made in three steps.

- In step one we used measurements on 40 individuals that were considered as independent measurements. For individuals with repeated observations we only used the first measurement. The theory for the one-dimensional case was used to estimate parameters, see Sec. 6.1.2

- In step two we analyzed repeated measurements on 20 persons. The data pairs were considered as dependent (y_{i_1} and y_{i_2}, $i = 1 \dots 20$) and a two-dimensional theory was used, see Sec. 6.2. For comparison with the traditional approach, with linear regression on the logarithms of data, we analyze this model in Sec. 8.

Tab. 2: Benzene data. Table showing resulting estimated parameters for model A. The parameters which have been found to be significant at 5% level ($|\hat{\beta}| > 1.96|\hat{\sigma}|$) are shown in bold.

parameter	40 persons		20 persons		40 persons	
description	independ. meas.		depend. meas.		whole data	
	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\sigma}$
house-heating	0.0022	0.1432	**0.3078**	**0.1529**	0.1787	0.1437
car parking	0.3615	0.2310	0.2703	0.1434	0.1821	0.1524
smoking habits	**0.3250**	**0.1398**	**0.2760**	**0.1152**	**0.4046**	**0.1149**
passive smoking	0.0015	0.0178	-0.0158	0.0103	-0.0232	0.0121
tanking car	-0.0032	0.0023	-0.0023	0.0016	**-0.0037**	**0.0018**
traffic	0.0175	0.0168	-0.0258	0.0188	0.0002	0.0141
car or bus	**0.0684**	**0.0148**	**0.0595**	**0.0147**	**0.0730**	**0.0132**
benzin work	-0.1881	0.2420	0.0219	0.2291	-0.2965	0.1952
-”- free time	-0.0625	0.1080	-0.0306	0.0697	-0.0154	0.0788
indoor not home	0.0023	0.0026	**0.0045**	**0.0016**	**0.0033**	**0.0016**
outdoor	0.0070	0.0088	**0.0175**	**0.0075**	0.0128	0.0081
home exposure	**1.1563**	**0.2123**	**1.1744**	**0.2781**	**1.0524**	**0.1721**
σ	**0.2278**	**0.0255**	**0.1821**	**0.0222**	**0.2443**	**0.0271**
ρ			-0.3711	0.2236	-0.4879	0.2957

- The final analysis was made using all the data, corresponding to both single (on 20 persons) and repeated measurements (on another 20 persons), and a theory which is described in Sec. 6.3.

Estimated values of the parameters for the case of benzene are shown in Table 2.

The most interesting result of our analysis was obtained on benzene data. Analysis that was made using all the data, corresponding to both single (on 20 persons) and repeated measurements (on another 20 persons) showed that smoking, rides by bus and car, estimated home exposure, staying indoor but not at home, and tanking cars have a pronounced effect with regard to the exposure to benzene (numbers in bold in Table 2). Exposure to benzene during tanking of a car is questionable because of negative sign of the estimate for that parameter. Here the significance is judged from asymptotic approximation. See Sec. 10 for more details on this point.

Analogous analysis was done for toluene and xylene as well. However, it was found that only estimated home exposure is a significant parameter for these two substances. Furthermore, the validity of asymptotic uncertainty estimates is examined in Sec. 10.1 and gives motivations for making confidence intervals broader, see Sec. 10.2.

9.4 Comparing the models A and B

We have fitted model B to the same data as a comparison. In order to discuss this in relation to Sec. 8, we select the case with 40 independent univariate data. However, the variation in estimated $x_i'\hat{\beta}/\hat{\beta}_0$ is moderate but not small as needed for Sec. 8 to be accurate.

Analysis of the full model with 12 covariates (plus β_0 and σ_z) is given for model B (40 person independent measurements) in Table 3.

Although $\hat{b}$ and $\hat{\beta}/\hat{\beta}_0$ differ, we can see that for significant parameters the sign and order of magnitude are reasonably close.

Reducing the models so that only significant covariates are retained only give marginal changes of the estimates as shown in Table 4.

We can see from Tables 3 and 4 that both models give the same information about which covariates are significant and also that $\hat{b}$ and $\hat{\beta}/\hat{\beta}_0$ are reasonably close for significant variables, possibly with the exception of the last covariate (estimated home exposure). Here the model A:s parameter estimate is close to 1 and appears much more reasonable than model B:s since it would have the value 1 if no estimation errors were involved.

Another point of interest for the comparison is extrapolation to situations with different exposure times. In model A, different exposure times will give different values of $x_{i,j}$ but we can use the same estimated β-values for example for the time in traffic etc. In model B, such extrapolation is not possible since the logarithm of accumulated exposure does not react linearly on exposure times, only accumulated exposure itself does!

Tab. 3: Benzene data. Table showing resulting estimated parameters for model B. The parameters which have been found to be significant at 5% level $(|\hat{\beta}| > 1.96|\hat{\sigma}|)$ are shown in bold. 40 persons, independent measurements. $\hat{\beta}/\hat{\beta}_0$ are taken from model A

parameter	$\hat{b}$	$\hat{\sigma_b}$	$\hat{\beta}/\hat{\beta}_0$
β_0	0.1140	0.0412	1.0000
house-heating	-0.0415	0.1497	0.0017
car parking	0.2982	0.1928	0.0279
smoking habits	**0.4366**	**0.1309**	0.2516
passive smoking	-0.0078	0.0161	0.0011
tanking car	-0.0031	0.0027	-0.0025
traffic	0.0317	0.0161	0.0135
car or bus	**0.0583**	**0.0127**	0.0530
benzin work	-0.0889	0.2055	-0.1456
-”- free time	-0.2000	0.1194	-0.0484
indoor not home	-0.0001	0.0024	0.0018
outdoor	0.0039	0.0103	0.0054
home exposure	**0.4522**	**0.0727**	0.8952

Tab. 4: Benzene data. Table showing resulting estimated parameters for models A and B with significant covariates only.

parameter	Model B		Model A		
description	$\hat{b}$	$\hat{\sigma_b}$	$\hat{\beta}/\hat{\beta}_0$	$\hat{\beta}$	$\hat{\sigma}$
β_0	0.1140	0.0412	1.0000	1.2842	0.0609
smoking habits	0.3706	0.0869	0.2534	0.3254	0.1160
car or bus	0.0491	0.0088	0.0459	0.0589	0.0137
home exposure	0.4883	0.0684	0.8316	1.0680	0.2076

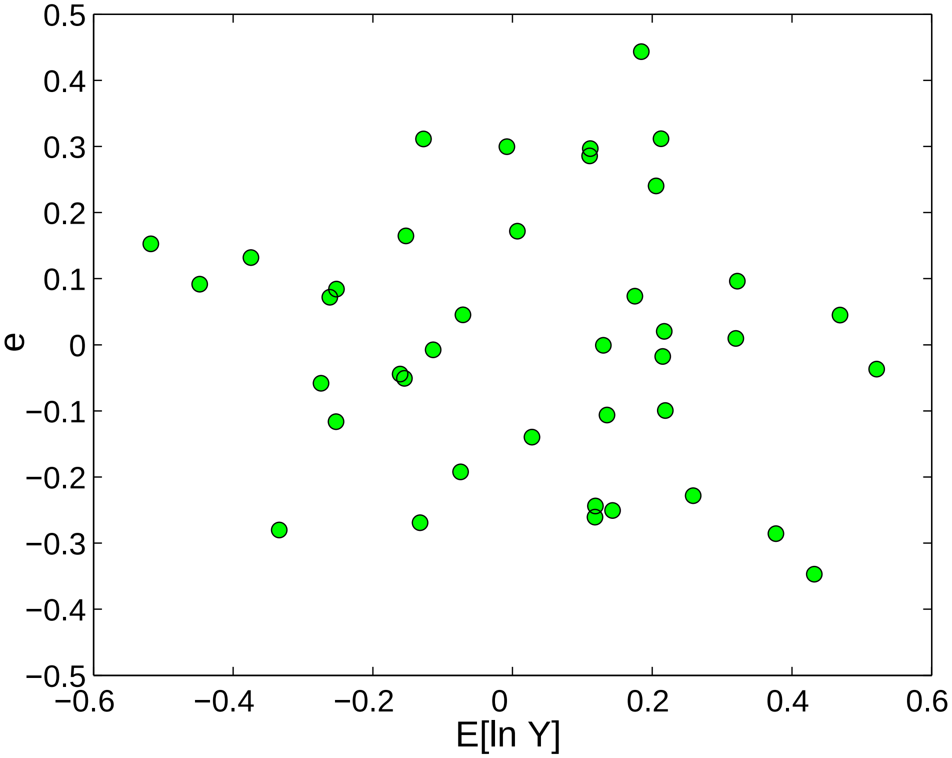

Fig. 2: Residuals for benzene data.

9.5 Residuals

Graphical representation of data is very useful in a case of just few variables. In the opposite case of many variables, such graphs are difficult to interpret. Then, as a rule, it is better to first try to fit the data by a model, and then only plot the residuals in order to reveal flaws of the model (non-linear effects of predictors or non-constant variance). In our case, we plot the values of y_i or residuals e_{y_i} as functions of regression equation $\hat{E}[\ln Y]$, see Fig. 2,

$$\hat{E}[\ln Y] = \mu_z = \ln x'\hat{\beta} - \sigma_z^2/2 \tag{9.1}$$

$$Var(\ln Y) = \sigma_z^2 \tag{9.2}$$

$$e_y = \ln Y - (\ln x'\hat{\beta} - \sigma_z^2/2) \tag{9.3}$$

The plot of residuals shows random variations without any clear pattern.

10. CONFIDENCE INTERVAL FOR ESTIMATED PARAMETERS

10.1 The validity of asymptotic confidence interval

In order to investigate the true confidence level of the asymptotic confidence interval (95%) we generate simulated data $Y_i^*, i = 1 \ldots n$ from the normal distribution as follows.

20 independent data points are simulated in accordance with $z_i \sim N(\ln x'\beta - \sigma_z^2/2, \sigma_z^2)$ and 20 independent *pairs* of data points are simulated in accordance with the following equations:

$$\begin{aligned}
z_{i1} &= \mu_{i1} + \sigma_z U, \text{ where} \\
\mu_{i1} &= \ln x'_{i1}\beta - \sigma_z^2/2 \text{ and} \\
z_{i2} &= \mu_{i2} + \rho(z_{i1} - \mu_{i1}) + \sqrt{1-\rho^2}\sigma_z U, \text{ where} \\
\mu_{i2} &= \ln x'_{i2}\beta - \sigma_z^2/2, \text{ and } U \sim N(0,1)
\end{aligned}$$

Here U denotes new independent random numbers in every generation.

We use estimates for β_i, σ, and ρ while generating the data (see Table 2, 40 persons, whole data). The parameters β_i^* are estimated in accordance with Sec. 6.3 and the confidence intervals are then calculated for the estimated values following Sec. 7, i.e. in our case:

$$\beta_i^* = \hat{\beta_i^*} \pm 1.96\hat{\sigma_\beta^*}, \tag{10.1}$$

where $\hat{\sigma_\beta^*}$ is asymptotic standard deviation for the corresponding β^*. Finally, we count how often the estimates of the real data happen to be within the confidence intervals of the equation above.

Confidence levels for the parameters corresponding to exposure to benzene for the number of simulations $n = 7500$ are as presented in Table 5 from which we see that the confidence level is somewhat low. Therefore we use a parametric-bootstrap method below.

Tab. 5: Observed covering of asymptotic 95%-confidence intervals in 7500 simulations of benzene data.

Param.	Description	95% conf. interval	Simul. hit
β_1	House-heating method (el./oil)	(-0.1029 ; 0.4603)	81.11
β_2	Garage inside house	(-0.1166 ; 0.4808)	80.46
β_3	Smoking	(0.1794 ; 0.6298)	79.84
β_4	Passive smoking	(-0.0469 ; 0.0005)	78.14
β_5	Tanking a car	(-0.0072 ; -0.0002)	80.46
β_6	Hours in intense traffic	(-0.0274 ; 0.0278)	83.37
β_7	Total time in a car or bus	(0.0471 ; 0.0989)	81.28
β_8	Benzin vapors professionally	(-0.6791 ; 0.0861)	81.95
β_9	Benzin vapors, free time	(-0.1698 ; 0.1390)	80.03
β_{10}	Total indoor time	(0.0002 ; 0.0064)	80.42
β_{11}	Total outdoor time	(-0.0031 ; 0.0287)	81.96
β_{12}	Estimated home exposure	(0.7157 ; 1.3897)	84.29
σ		(0.1912 ; 0.2974)	83.57
ρ		(-1.0675 ; 0.0917)	44.76

10.2 Bootstrap simulation in finite data sets

The asymptote is not sufficiently exact for 40 persons, but simulations can give better accuracy for confidence intervals, see Ref. [12]. Parametric bootstrap can be used instead, a more general method which is based on numerical computer simulations.

In this method, one uses an estimated parametric distribution with the introduced Maximum Likelihood (ML)-estimations of the parameters as a replica of the initial source distribution (i.e. the probability distribution which the data originate from).

Estimations uncertainty can be obtained by generating new data sets of the same size as the initial one, using this distribution [12]. First, we find the parameter $\tilde{\theta}$ in the parametric distribution $F(x, \hat{\theta})$ that corresponds to the estimation $\hat{\theta} = \hat{\theta}(y_1, y_2 \ldots y_n)$ of θ based on the real data $y_1, y_2 \ldots y_n$. As usual, $\tilde{\theta} = \hat{\theta}$.

Next, we generate new data $y_1^*, y_2^* \ldots y_n^*$ from $F(y, \hat{\theta})$ with each y_i^* being independent of others and use the same method as described in Sec. 10.1.

Since y_i^* are log-normal, we generate first the normal variables $z_i^* \sim N(\ln x_i'\tilde{\beta} - \tilde{\sigma_z}^2/2, \tilde{\sigma_z}^2)$ and take $y_i^* = \exp(z_i^*)$. The estimation $\hat{\theta}^* = \tilde{\theta}(y_1^*, y_2^* \ldots y_n^*)$ is then calculated exactly as for the real data. This procedure is repeated 7500 times. The distribution of $(\hat{\theta}^* - \tilde{\theta})$ gives an estimate of the distribution of $(\hat{\theta} - \theta)$ [14] and the distribution of $(\hat{\theta}^* - \tilde{\theta})/\hat{\sigma}^*_{\hat{\theta}^*}$ gives an estimate of the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$ which is the studentised expression often used for confidence intervals. It turns out that for the regression coefficients $\hat{\beta}^*$ the distribution looks very normal (Fig. 3 left panel). For $\hat{\sigma^*}$, and $\hat{\rho}^*$ the distribution is non-symmetric, see Fig. 3. Here $\hat{\beta}$ is an estimate of the real β, and $\hat{\sigma}$ is an estimate of σ.

We calculate then the "studentized" quantity,

$$u_i = \left(\frac{\hat{\beta}^* - \tilde{\beta}}{\hat{\sigma}^*} \right)_i, i = 1 \ldots 7500$$

where $\tilde{\beta}$ is the true estimate of β, $\hat{\beta}^*$ is an estimate of β that is given by our simulation, and $\hat{\sigma}^*$ is an asymptotic standard deviation of each $\hat{\beta}^*$ given by our modelling, and obtain the distributions of, for instance, β_{12}, σ, and ρ, see Fig. 3.

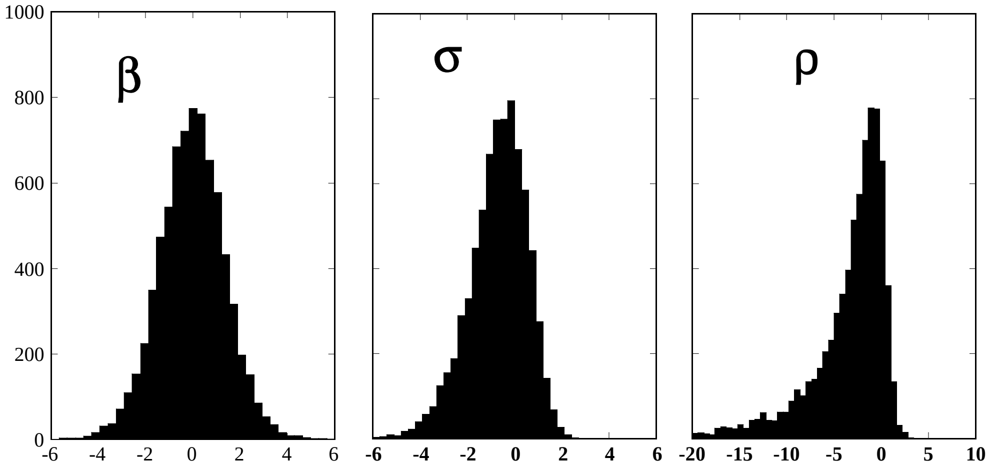

Fig. 3: Simulated histograms of u_i for estimated home exposure parameter and for σ, and ρ (from left to right).

One would expect that in accordance with the asymptotic theory (see Sec. 7), the simulated values u_i for each β were normally distributed, i.e.

$\sim N(0,1)$. However, from the percentiles for different confidence intervals (Table 6), one can see that the median value for $\lambda_{\alpha/2}$ and $\lambda_{1-\alpha/2}$ are -3.28 and 2.99, respectively, instead of 1.96 (for $\alpha = 0.05$) as for the standard normal distribution.

All distributions for studentised β-variables are wider than the $N(0,1)$-distribution, probably due to uncertainty in $\hat{\sigma}$-estimations. For the σ-parameter, the distribution is strongly distorted while for the ρ-parameter the asymptotic approximation is totally unsuitable. Here a different bootstrap approach may have an advantage, but this has not been at focus in this work.

Tab. 6: Percentiles.

	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
β_1	-4.1505	-3.3346	-2.6660	-1.9728	1.7927	2.3601	2.9889	3.6953
β_2	-4.4497	-3.4142	-2.7822	-2.0673	1.8167	2.3729	2.8998	3.5944
β_3	-4.4537	-3.3646	-2.7544	-2.0332	1.8456	2.5187	3.1446	3.9441
β_4	-4.2728	-3.3049	-2.6473	-2.0014	2.0442	2.7405	3.4140	4.4140
β_5	-4.1968	-3.3166	-2.5960	-1.9507	1.8625	2.5260	3.0766	3.8671
β_6	-3.6098	-2.9400	-2.3796	-1.7717	1.7746	2.3072	2.8574	3.4115
β_7	-4.0329	-3.1680	-2.5595	-1.9315	1.8163	2.4011	2.9611	3.5943
β_8	-4.0324	-3.1432	-2.5028	-1.8948	1.8219	2.4262	2.9867	3.9250
β_9	-4.3429	-3.4181	-2.6300	-1.9483	1.9226	2.5514	3.1913	4.2415
β_{10}	-4.1174	-3.1816	-2.5127	-1.9231	1.9159	2.5498	3.1722	3.9961
β_{11}	-4.1980	-3.2602	-2.6160	-1.9539	1.7454	2.3887	2.9440	3.7356
β_{12}	-3.5457	-2.8677	-2.3637	-1.7709	1.6652	2.2251	2.6668	3.3120
σ	-4.2432	-3.5337	-3.0008	-2.3519	0.7145	1.0377	1.2922	1.5891
ρ	-29.025	-20.125	-14.735	-9.9412	0.1752	0.6234	0.9908	1.4165

Looking back at our estimates in Table 2 (40 persons, whole data), we now see that ”smoking habits” (t-value = 3.52), ”car or bus” (t-value = 5.53), ”estimated home exposure” (t-value = 6.115) are still significant ($\alpha = 0.05$) while ”tanking cars” and ”indoor not home” fail to be significant.

11. CONCLUSIONS

Environment-oriented statistical research is important for quantification of pollution and its adverse effects on health. It is now well known that air pollution contribute to development of astma and chronic irritation of air pathways, as well as allergy in both children and adults.

During last years, a great emphasis was put to studies devoted to identification of hazards that humans exposure themselves to and to revealing relations between the exposures and their impacts on health. In this situation, it is highly desirable to further improve statistical methods for data treatment and their validation in all environmental research programs.

We present statistical analysis of accumulated air-pollution levels measured with help of carry-on sensors. In contrast to previous studies, we use the untransformed raw data, i.e. directly measured exposures and not their logarithms. We believe that our approach is more correct because it is the raw accumulated exposures to various hazardous substances rather than their logarithms that should be summed up in accumulated data. However, our analysis incurs an extra "expense" of more complex estimates and computations.

In particular, we demonstrated the following.

- Accumulated exposure is affected by exposure times in different environments. It is possible to combine the linear regression in the original scale (the untransformed raw data) with lognormal distribution.

- It is shown how a single and repeated observations can be used in the same analysis.

- Asymptotic methods for uncertainty analysis appeared to be unreliable for data volume of actual size (40 persons, 60 observations). Alternative simulations-based methods show how large the uncertainties are in reality.

- In accordance to the asymptotic uncertainty analysis (individual 95% confidence level) made on the whole data, the following parameters were found to be of significance: ”smoking”, ”tanking cars”, ”time of going by car or bus”, ”time being indoor, not home”, and ”estimated home exposure”.

  Simulation analysis for the same data shows that ”smoking”, ”time of going by car or bus”, and ”estimated home exposure” are still significant, while ”tanking cars” and ”time being indoor, not home” can no longer be distinguished from zero for a 95% confidence level.

- Comparison of a traditional modelling (linear regression for the logarithms of accumulated data) with our model (linear regression for the untransformed data) shows that both the models can approximate each other for small relative variations in regression equation. For larger variations, the difference between the models becomes more important. In particular, our suggested model can easily account for extrapolation to a new case with other exposure times in different environments, while the traditional method looses its applicability.

- For the actual data example, the same parameters turn out to be significant in model A and B , but the values of the parameters and their interpretations are different. The differences between the parameter estimates are also somewhat larger than explained by the special case with small x-variations. In both models, however, qualitative effects of the significant parameters are yet the same.

12. ACKNOWLEDGEMENTS

I thank my supervisor Urban Hjorth for his support and his truly angelic patience when quiding me through this work, for thoroughly reading the manuscript and for many useful comments.

I am grateful to Jacques de Marè for being my second supervisor through this work and giving valuable comments after reading my thesis.

Thanks to all members of the Mathematical Statistics group for everyday assistance and nice and supportive atmosphere.

I am also grateful to Peter Jagers for always leaving the door open for coming and expressing my needs and problems.

The discussions with Gerd Sällsten and Lars Barregård, as well as their kind permission to use their exposure data is greatly appreciated.

Finally, I wish to cordially thank my husband for his help, support, and his always believing in me.

13. APPENDICES

13.1 Derivatives used in two-dimensional case

In order to simplify the resulting appearance of equations, we will use the following auxiliary notations:

$$A_1 \equiv \sum_j x_{1j} \cdot \beta_j \tag{13.1}$$

$$A_2 \equiv \sum_j x_{2j} \cdot \beta_j \tag{13.2}$$

$$\mu_1 \equiv z_1 - \ln(A_1) + \frac{\sigma^2}{2} \tag{13.3}$$

$$\mu_2 \equiv z_2 - \ln(A_2) + \frac{\sigma^2}{2} \tag{13.4}$$

$$Q \equiv 1 - \rho^2 \tag{13.5}$$

The derivatives now are as follows.

First derivatives:

$$\frac{\partial L}{\partial \beta_j} = \frac{(1-\rho)}{Q\sigma^2} \left(\mu_1 \frac{x_{1j}}{A_1} + \mu_2 \frac{x_{2j}}{A_2} \right) \tag{13.6}$$

$$\frac{\partial L}{\partial \sigma} = \frac{1}{\sigma} \left[-2 + \frac{(\mu_1 + \mu_2)(1-\rho)}{Q} \right] + \frac{(\mu_1^2 + \mu_2^2 - 2\rho\mu_1\mu_2)}{Q\sigma^3} \tag{13.7}$$

$$\frac{\partial L}{\partial \rho} = \frac{(\rho + \mu_1\mu_2/\sigma^2)}{Q} - \frac{\rho(\mu_1^2 + \mu_2^2 - 2\rho\mu_1\mu_2)}{\sigma^2 Q^2} \tag{13.8}$$

Second derivatives:

$$\begin{aligned}\frac{\partial^2 L}{\partial\beta_i\partial\beta_j} &= \frac{1}{Q\sigma^2}\Big[\rho\left(\frac{\mu_2 x_{1i}x_{1j}}{A_1^2} + \frac{\mu_1 x_{2i}x_{2j}}{A_2^2} + \frac{x_{1i}x_{2j} + x_{1j}x_{2i}}{A_1 A_2}\right) \\ &\quad - \left(\frac{x_{1i}x_{1j}}{A_1^2} + \frac{\mu_1 x_{1i}x_{1j}}{A_1^2} + \frac{\mu_2 x_{2i}x_{2j}}{A_2^2} + \frac{x_{2i}x_{2j}}{A_2^2}\right)\Big]\end{aligned} \tag{13.9}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial\beta_j\partial\sigma} &= \frac{1-\rho}{Q\sigma}\left(\frac{x_{1j}}{A_1} + \frac{x_{2j}}{A_2}\right) \\ &\quad + \frac{2}{\sigma^3 Q}\left[\rho\left(\frac{\mu_2 x_{1j}}{A_1} + \frac{\mu_1 x_{2j}}{A_2}\right) - \left(\frac{\mu_1 x_{1j}}{A_1} + \frac{\mu_2 x_{2j}}{A_2}\right)\right]\end{aligned} \tag{13.10}$$

$$\frac{\partial^2 L}{\partial\beta\partial\rho} = -\frac{2\rho}{Q^2\sigma^2}\left[\rho\left(\frac{\mu_2 x_{1j}}{A_1} + \frac{\mu_1 x_{2j}}{A_2}\right) - \left(\frac{\mu_1 x_{1j}}{A_1} + \frac{\mu_2 x_{2j}}{A_2}\right)\right] \tag{13.11}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial\sigma^2} &= \left[\frac{2}{Q} - \frac{3(\mu_1+\mu_2)}{Q\sigma^2} + \frac{6\mu_1\mu_2}{Q\sigma^4}\right]\rho - \\ &\quad - \frac{2}{Q} + \frac{1}{\sigma^2}\left[2 + \frac{3(\mu_1+\mu_2)}{Q}\right] - \frac{3(\mu_1^2+\mu_2^2)}{Q\sigma^4}\end{aligned} \tag{13.12}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial\rho^2} &= \frac{8\mu_1\mu_2}{Q^3\sigma^2}\rho^3 + \left[\frac{2}{Q^2} - \frac{4(\mu_1^2+\mu_2^2)}{Q^3\sigma^2}\right]\rho^2 \\ &\quad + \frac{6\mu_1\mu_2}{Q^2\sigma^2}\rho + \frac{1}{Q} - \frac{\mu_1^2+\mu_2^2}{Q^2\sigma^2}\end{aligned} \tag{13.13}$$

$$\begin{aligned}\frac{\partial^2 L}{\partial\rho^2} &= \left[\frac{2(\mu_1+\mu_2)}{Q^2\sigma} - \frac{4\mu_1\mu_2}{Q^2\sigma^3}\right]\rho^2 \\ &\quad + \left[\mu_1^2+\mu_2^2 - \sigma^2(\mu_1+\mu_2)\right]\rho + \frac{1}{Q^2\sigma^3}\left[\sigma^2(\mu_1+\mu_2) - 2\mu_1\mu_2\right]\end{aligned} \tag{13.14}$$

13.2 Characteristics of hazardous chemicals [15]

13.2.1 Toluene

Toluene is a clear, colorless liquid with an aromatic odor. It is a natural constituent of crude oil. Toluene is both volatile and flammable at room temperature. Toluene can be smelled in air at a level of about 80 parts per billion (ppb). In water, it can be tasted at a level of 40 ppb. These levels are well below the dangerous concentrations for short exposure times. Toluene has a moderate tendency to accumulate in the food chain.

Gasoline (which contains from 5% to 7% of toluene) is the largest source of toluene air pollution. Toluene is released to the atmosphere during the production, transport, and combustion of gasoline. Toluene exposures are highest in areas of intense traffic and near gasoline stations. Toluene is however short-living in air because of its high chemical reactivity.

Common household products and cigarette smoke are the principal sources of toluene indoors. Indoor toluene concentration is often several times higher than outside. Cigarette smokers inhale about 80 to 100 micrograms of toluene per cigarette. Toluene-containing consumer products include various aerosols, paints, paint thinners, varnishes, rust inhibitors, adhesives, and solvent-based cleaning agents. Toluene is used as a solvent in cosmetic nail polishes at concentrations of up to 50%.

Although most environmental toluene is released directly to the atmosphere, it is occasionally detected in drinking water supplies. Nonetheless, drinking water levels of toluene are usually low.

13.2.2 Benzene

Benzene is a volatile, colorless, highly flammable liquid. Today, most (98%) benzene is commercially derived from petrochemical and petroleum refining industries. Benzene is a by-product of various combustion processes, such as forest fires and the burning of wood, garbage, organic wastes, and cigarettes; it is also released to the air from crude oil seeps and volatilizes from plants.

Benzene is one of the world's major commodity chemicals. Benzene is an important raw material for the manufacture of synthetic rubbers, gums, lubricants, dyes, and pharmaceutical and agricultural chemicals; it is also found in consumer products such as glues, paints, and marking pens.

Benzene is also a natural component of crude and refined petroleum. The mandatory decrease of lead alkyls in gasoline has led to an increase in the aromatic hydrocarbon content of gasoline to maintain high octane levels.

Benzene is widespread in the environment, and is commonly found in air, water, and humans. The major environmental sources is car exhaust, car tanking, hazardous waste sites, chemical spills and manufacturing sites, and petrochemical industries.

As can be seen from the environmental sources, inhalation accounts for up to 99% of the total daily intake of benzene. Smoking is the largest source of benzene exposure for the general public. The estimates of daily intake of benzene from a single cigarette vary: from 5.9 to 90 μg. Passive smoking is also a source of exposure.

13.2.3 Xylene

Xylene is a colorless, sweet-smelling liquid that catches on fire easily. It occurs naturally in petroleum and coal tar and is formed during forest fires. You can smell xylene in air at $0.1 - 4$ parts of xylene per million parts of air (ppm) and begin to taste it in water at $0.53 - 2$ ppm.

Xylene is one of the top 30 chemicals produced in the world in terms of volume. It is used as a solvent and in the printing, rubber, and leather industries. It is also used as a cleaning agent, a thinner for paint, and in paints and varnishes. It can be found in small amounts in gasoline.

Xylene has been found in waste sites when discarded as used solvent, or in varnish, paint, or paint thinners. It evaporates quickly from the soil and water into the air. In the air, it is broken down by sunlight into other less harmful chemicals. It can be broken down by microorganisms in soil and water as well. Only a small amount of xylene is accumulated in fish, plants, and animals living in xylene-contaminated water.

Breathing xylene in workplace air or in automobile exhaust is the main source of xylene in human body. Next is breathing cigarette smoke that has small amounts of xylene in it or drinking contaminated water or breathing air near waste sites and landfills that contain xylene. The amount of xylene in food is much likely to be quite low.

Xylene affects the brain. High levels from long-term exposure to xylene can cause headaches, lack of coordination or sense of balance and dizziness. Exposure at high levels of xylene for short periods can also cause irritation of the skin, eyes, nose, and throat; difficulty in breathing; problems with the lungs; delayed reaction time; memory difficulties; stomach discomfort; and possibly changes in the liver and kidneys. It can cause unconsciousness and even death at very high levels.

It is not known if xylene harms the unborn child if the mother is exposed to low levels of xylene during pregnancy. Xylene has not been proven to be carcinogenic although the corresponding studies are not conclusive and cannot exclude the opposite.

BIBLIOGRAPHY

[1] S. Molander. Föreläsning om Miljöproblem och miljövetenskap. Institutionen för miljösystemanalys, Chalmers tekniska hgskola.

[2] L. Köhler. Sammanfattning av ”Indikatorer fr barns hälsa i Sverige. Bidrag till ett kommunalt barnindex”, see http://www.rb.se/bokhandel.

[3] S. Jungnelius, M. Svartengren, ”Hälsoeffekter av trafikavgaser”, Rapport från Yrkesmedicinska enheten 2000:3, see http://www.sll.se/w_amm/14569.cs?dirid=97738.

[4] M. Krusȧ, T. Bellander, and M. Nilsson, Cancerframkallande ämnen i tätortsluft. Stockholm 2002/2003. Rapport till Naturvrdsverket. Programomrde: Hälsorelaterad miljöövervakning.

[5] E.Symanski, G.Sällsten, W.Chan, and L.Barregård, Heterogeneity in Sources of Exposure Variability Among Groups of Workers Exposed to Inorganic Mercury, Ann. occup. Hyg., Vol. 45, No. 8, pp. 677-687 (2001).

[6] E.Symanski, G.Sällsten, and L.Barregård, Vatiability in Airborne and Biological Measures of Exposure to Mercury in the Chloralkali Industry: Implications for Epidemiologic Studies, Environ. Health Perspec., Vol. 108, No. 6, pp. 569-573.

[7] C.Peretz, A.Goren, T.Smid, and H.Kromhout, Application of Mixed-effects Models for Exposure Assessment, Ann. occup. Hyg., Vol. 46, No. 1, pp. 69-77 (2002).

[8] S. Gauvin, P. Reungoat, S. Cassadou, J. Déchenaux, I. Momas, J. Just, and D. Zmirou, Contribution of indoor and outdoor environments to PM2.5 personal exposure of children - VESTA study., Sci. Total Environ. 297, pp.175-181 (2002).

[9] K. Sexton, J.L. Adgate, G. Ramachandran, G.C. Pratt, S.J. Mongin, T.H. Stock, and M.T. Morandi, Comparison of personal, indoor, and outdoor exposures to hazardous air pollutants in three urban communities, Environ. Sci. Technol. Vol.38, No.2, pp.423-430, (2004).

[10] Bernard W. Lindgren, *Statistical Physics*, (Academic Press, New-York, 1999).

[11] G. van Belle, *Statistical Rules of Thumbs*, (2002).

[12] U. Hjort, *Statistisk slutledning* (Ekonomi och teknik, 1998).

[13] G. Sälsten, J. Björklund, O. Johansson, J. Melin, R. Lindahl, C. Loh, C. Östman, L. Barregård, Cancerframkallande ämnen i tätortsluft - personlig exponering, individrelaterande stationära mätningar och bakgrundsmätningar i Göteborg 2000. A report.

[14] U. Hjort, Computer intensive statistical methods - Validation, model selection, bootstrap, Chapman and Hall, 1994.

[15] Adapted from http://www.atsdr.cdc.gov/

[16] E.L. Lehmann, Theory of point estimation. Wiley 1983.