# Chapter 9

# Introduction to the Use of Regression Models in Epidemiology

**Ralf Bender**

## Summary

Regression modeling is one of the most important statistical techniques used in analytical epidemiology. By means of regression models the effect of one or several explanatory variables (e.g., exposures, subject characteristics, risk factors) on a response variable such as mortality or cancer can be investigated. From multiple regression models, adjusted effect estimates can be obtained that take the effect of potential confounders into account. Regression methods can be applied in all epidemiologic study designs so that they represent a universal tool for data analysis in epidemiology. Different kinds of regression models have been developed in dependence on the measurement scale of the response variable and the study design. The most important methods are linear regression for continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson regression for frequencies and rates. This chapter provides a nontechnical introduction to these regression models with illustrating examples from cancer research.

**Key words:** Regression, linear regression, logistic regression, Poisson regression, Cox regression.

## 1. Introduction

In analytical epidemiology, a common research question is whether an exposure such as radiation or dust has an impact on a response such as mortality or cancer. Frequently, the exposure is measured by means of a binary indicator (exposure yes/no). However, the use of more than two exposure categories as well as the use of a continuous variable to describe the exposure is also possible. In theory, the best study design to investigate the impact of an exposure on a response is given by a randomized controlled trial. However, in epidemiologic research randomization of people to different exposure groups is impossible in almost all cases. Thus,

Mukesh Verma (ed.), Methods in Molecular Biology, Cancer Epidemiology, vol. 471

Book doi: 10.1007/978-1-59745-416-2

common study designs in epidemiology to investigate relations between exposure and response are given by cohort studies and case-control studies. In these designs without randomization, the relation between two variables can be confounded by other variables. If this is not taken into account in data analysis, the results may be seriously biased.

Principally, there are two ways to deal with confounders in epidemiologic research: via study design or via data analysis. One example of a design-related method to take confounders into account is matching. One of the most important analysis related methods to deal with confounding is multiple regression analysis. Other methods are, for example, stratification, the use of instrumental variables, or the application of propensity scores. In this chapter, an overview of the most important multiple regression models is given with a focus on applications in modern epidemiology.

The term regression is somewhat misleading, and it does not describe the main feature of the method in its current applications. Historically, the term regression goes back to Galton who described the tendency for the offspring of seeds to be closer to the average than their parent seeds *(1)*. However, modern applications of regression methods do not only analyze such "regression to the mean" effects. Nowadays, regression methods are used to describe in general any relationship between a response variable, also called dependent or outcome variable and one or multiple explanatory variables, also called independent or predictor variables, covariates, or risk factors. In the case of only one explanatory variable, the corresponding model is called simple regression, the use of multiple explanatory variables leads to the application of multiple or multifactorial regression. Frequently, multiple regression is also called multivariate or multivariable regression. However, the use of these terms may be confusing because regression models dealing with multiple response variables are also called multivariate regression models *(2–5)*. To avoid confusion, we use the term multiple regression for models with one response and multiple explanatory variables. In this chapter, we discuss only regression models with one response variable. An introduction to multivariate regression models can be found in textbooks about multivariate statistical methods *(3,4)*.

Different regression models have been developed in dependence on the measurement scale of the response variable. The basic standard regression models are linear regression for continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson regression for frequencies and rates. In the next sections, the basic features of these types of regression models are summarized, followed by some remarks about model building. Finally, important extensions of the standard regression models are summarized.

## 2. Linear Regression

### 2.1. Simple Linear Regression

The use of simple linear regression is possible if the effect of one continuous explanatory variable X, for example, body mass index measured in kilograms per square meter on one continuous response variable Y, for example, systolic blood pressure measured in millimeters of Hg is to be investigated. The fundamental model equation is given by

$$Y = \beta_0 + \beta_1 x + e \tag{1}$$

where $\beta_0$ is called intercept, $\beta_1$ is the regression coefficient for X, x is the observed value of X, and e is the random error describing individual deviations from the mean of Y given X = x, also called residual term. Through the functional form a straight line is described, where the intercept $\beta_0$ represents the mean of Y for x = 0 and the regression coefficient $\beta_1$ represents the slope of the line, that is, the average increase of Y for a one-unit increase of X. The fundamental assumption of simple linear regression is that the true association of Y and X is in fact linear. This assumption has to be fulfilled at least approximately to yield interpretable results. If the association of Y and X cannot be approximated by a straight line, more complicated models should be applied (see subheadings 7.4. and 7.5.). Another usual assumption is that the residual term is normally distributed with mean 0 and variance $\sigma^2$. However, the assumption of normally distributed residuals is negligible in practice if the sample size is large enough and the distribution of the residual term is not too skew.

In epidemiologic practice, the parameters $\beta_0$ and $\beta_1$ have to be estimated from data of n subjects that should represent a random sample of the considered population. In the simplest case, the data consist of n independent pairs $(x_i,y_i)$, i=1,...,n, obtained from n subjects, where $x_i$ and $y_i$ represent the observed values of subject i for X and Y, respectively. In the case of repeated measurements on the same subject or other reasons leading to correlated data, the use of more complicated models is required (see subheading 7.6.). In the case of n independent pairs $(x_i,y_i)$ for i=1,...,n, the parameters $\beta_0$ and $\beta_1$ of model 1 can be estimated from these data. The standard estimation method is given by the method of ordinary least squares (OLS), which results in a fitted regression line that minimizes the average of the squared deviation of the line from the observed data *(1)*. The OLS estimates of the parameters $\beta_1$ and $\beta_0$ are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the arithmetic means of $x_i$ and $y_i$, i=1,…,n, respectively.

Before a straight line is fitted to data they should be examined graphically by means of a scatter plot of Y versus X to assess whether the linearity assumption holds at least approximately. If this is the case, the estimated effect of X on Y can be described by the estimated slope, because the larger $\beta_1$ the steeper the regression line. The estimation uncertainty should be described by means of the standard error (SE) and the 95% confidence interval (CI) for $\beta_1$. It is also possible to test the null hypothesis $H_0$: $\beta_1=0$ by means of a *t* test. Only in the case of a significant test result we can conclude that a significant effect of X on Y is found. Additionally, we can estimate the expected value of Y for given values of X. Again, the estimation uncertainty should be described by means of SEs and CIs. For the whole fitted line confidence bands are useful.

In summary, the goals of regression models are threefold:

1. To determine whether the variables Y and X are systematically related (test of $H_0$: $\beta_1=0$).
2. To estimate the effect size of X on Y by means of $\hat{\beta}_1$ (complemented by a 95% CI).
3. To predict the expected value of Y for given values of X (with 95% CI).

### 2.2. Multiple Linear Regression

In epidemiology, simple linear regression plays only a negligible role. First, binary and time to event outcomes are much more common than continuous response variables (see subheadings 3.4., and 5.). Second, in almost all cases, the effects of confounders have to be taken into account, so that multiple linear regression should be applied even if we are interested mainly in the effect of one primary explanatory variable. The multiple linear regression model is an extension of model 1 where instead of only one explanatory variable X several explanatory variables $X_1,\ldots,X_k$ with observed values $x_1,\ldots,x_k$ are considered. The fundamental model equation is given by

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e \qquad (3)$$

where $\beta_j$ represent the regression coefficients for $X_j$ for j=1,…,k. It is not required that all of the explanatory variables $X_1,\ldots,X_k$ are measured on a continuous scale. It is also possible to include binary explanatory variables (e.g., exposure yes/no). In this case, we should choose one reference group (e.g., unexposed subjects), assign the value 0 to this group, and assign the value 1 to the other group (e.g., exposed subjects). Then, the corresponding regression coefficient describes the mean response difference of exposed and unexposed subjects. It is also possible to include categorical explanatory variables with more than two groups. One

possibility is to create so-called dummy variables. For example, if the categorical variable X is given by marital status with the categories single, married, divorced, and widowed, we should choose one reference group (e.g., single) and create the new binary dummy variables $X_1$=married (yes/no), $X_2$=divorced (yes/no), and $X_3$=widowed (yes/no). In general, a categorical explanatory variable with m categories can be included in a multiple regression model by means of m–1 dummy variables. Thus, model 3 represents a general tool to assess the effects of continuous and categorical explanatory variables on a continuous response variable. As a consequence, other well-known statistical methods like unpaired *t* test and analysis of variance are embedded in the general multiple linear regression model 3. For example, the test of $H_0$: $\beta_1=0$ is equivalent to the unpaired *t* test, if the model contains only one binary explanatory variable.

As discussed above, the usual method for parameter estimation is given by OLS. The corresponding formulas are best summarized by using matrix notation, which goes beyond the scope of this chapter. The interested reader can find the formulas for example in Matthews *(6)* or standard textbooks *(7,8)*. As discussed above, we can test whether the explanatory variable $X_j$ (j=1,…,k) has a significant effect on the response Y, we can describe the estimated effect size of $X_j$ by means of $\hat{\beta}_j$ for j=1,…,k (with CIs), and we can predict the expected value of Y for given values of $X_1,\ldots,X_k$ (with CIs). The advantages of model 3 in comparison with model 1 are given by the fact that model 3 can describe the effects of several explanatory variables simultaneously and that the regression coefficient $\beta_j$ represents the effect for $X_j$ adjusted for all other explanatory variables. In general, it is misleading to assess the associations of the response Y and several explanatory variables by means of several simple linear regression models. The corresponding simple estimates may be seriously biased, because the effects of other explanatory variables are not taken into account.

### 2.3. Interactions

In addition to the assumptions of model 1, model 3 contains another important assumption. It is assumed that the effects of the explanatory variables $X_1,\ldots,X_k$ are additive, i.e., the effect of one explanatory variable is independent of all other explanatory variables. If this is not the case, the model should be extended by inclusion of appropriate interaction terms. This can be best explained by using two binary explanatory variables, say $X_1$=sex (male=1, female=0) and $X_2$=exposure (yes=1, no=0). If the effect of exposure is different for men and women, the model should contain the corresponding interaction term given by $X_3=X_1\times X_2$, that means $X_3$=1 for exposed men and $X_3$=0 otherwise. If we include the interaction term $X_3$ in the regression model, the corresponding regression coefficient describes the additional

effect of the exposure in males. In detail, $\beta_0$ represents the mean response for unexposed females, $\beta_1$ the mean response difference between unexposed males and females (sex effect in unexposed subjects), $\beta_2$ the mean response difference between exposed and unexposed females (exposure effect in females), $\beta_2+\beta_3$ the mean response difference between exposed and unexposed males (exposure effect in males), and $\beta_1+\beta_3$ the mean response difference between exposed males and females (sex effect in exposed subjects). In this model, there are no unique effects, neither for sex nor for exposure. The effect of sex is dependent on the exposure status, and the effect of exposure is dependent on sex.

### 2.4. Coefficient of Determination

A frequently used measure to describe the predictive ability of a linear regression model is the coefficient of determination, $R^2$, defined by

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (4)$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_1 x_{ki} \quad (5)$$

is the estimated response value for subject i with values $x_{1i},\ldots,x_{ki}$ for the explanatory variables. $R^2$ is the proportion of the model sums of squares to the total sums of squares and represents the proportion of variability of the response explained by the explanatory variables all together. Frequently, in practice $R^2$ is low although the model contains one or more significant explanatory variables with very low *P* values, especially if the sample size is large. A number of other measures and graphical tools to describe the goodness-of-fit of linear regression models are available *(7,9)*.

### 2.5. Remarks

Multiple linear regression can be applied in cohort studies and in cross-sectional studies. In all cases, one should be careful in interpreting the results, because significant regression coefficients demonstrate statistical associations but not necessarily causal effects. An additional major limitation of cross-sectional studies is that, in general, it cannot be clearly determined whether an explanatory variable has an effect on the response or vice versa.

### 2.6. Example

Chan et al. *(10)* applied multiple linear regression to estimate standard liver weight for assessing adequacies of graft size in live donor liver transplantation and remnant liver in major hepatectomy for cancer. Standard liver weight (SLW) in grams, body weight (BW) in kilograms, gender (male=1, female=0), and other anthropometric data of 159 Chinese liver donors who underwent donor right hepatectomy were analyzed. The formula

$$SLW = 218 + 12.3 \times BW + 51 \times gender \quad (6)$$

was developed and a coefficient of determination of $R^2$=0.48 was reported *(10)*. These results mean that in Chinese people, on average, for each 1-kg increase of BW, SLW increases about 12.3 g, and, on average, men have a 51-g higher SLW than women. Unfortunately, SEs and CIs for the estimated regression coefficients were not reported. By means of formula 6 the SLW for Chinese liver donors can be estimated if BW and gender are known. About 50% of the variance of SLW is explained by BW and gender.

## 3. Logistic Regression

### 3.1. Importance of Logistic Regression

In cancer epidemiology, frequently the associations of risk factors with development of cancer or cancer mortality are investigated. This leads to binary response variables (e.g., cancer yes/no). The usual regression model for binary responses is logistic regression. It has been shown that this model can not only be applied in cohort studies but also in case-control studies *(11,12)*. This makes logistic regression one of the most important statistical methods in epidemiologic research. Levy and Stolte found that in about one of three articles published in the *American Journal of Public Health* and the *American Journal of Epidemiology* between 1990 and 1998 logistic regression was used *(13)*.

### 3.2. Explanation of Logistic Regression

The basic question leading to logistic regression is the same as in subheading 2. about linear regression. Are there associations between the explanatory variables $X_1,\ldots,X_k$ and the response variable Y? The only difference to the data situation described in subheading 2. is that the response Y is a binary variable (event yes/no). The naive use of linear regression would mean that we fit a continuous relation between the explanatory variables and the response, although there are only two possible values for Y, namely, 0 and 1. However, if we look at the event probability $\pi$=P(Y=1), we have a continuous term with possible values between 0 and 1, which can be modeled in a continuous manner, but it has the inconvenient restriction that values lower than 0 and higher than 1 are impossible. If we look at the odds $\pi/(1-\pi)$, we have a continuous term with possible values between 0 and infinity and if we take the natural logarithm of the odds $\log(\pi/(1-\pi))$, called logit, we have a mathematically convenient term with no restrictions of the domain. The logistic regression model is given by the linear relation between the logit and the values of the explanatory variables,

$$\log(\pi/(1-\pi)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \quad (7)$$

Equation 7 is mathematically equivalent to

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)} \quad (8)$$

The mathematical function described in formula 8 is called logistic function explaining the term logistic regression. Fortunately, the basic features of multiple regression as described in subheading 2. (need to check model assumptions, use of dummy variables, use of interaction terms) also apply to logistic regression. Thus, we can concentrate on the differences between logistic and linear regression.

Because the logits are not available for each individual, we cannot simply create scatter plots between the response and the explanatory variables as in linear regression. We have to use percentage plots based on grouped data as described by Hosmer and Lemeshow *(14)* to assess the relation between the event probability and the explanatory variables graphically.

For the same reason, it is not possible to use OLS for parameter estimation in logistic regression. However, parameter estimation can be performed by means of maximum likelihood, which is available in all statistical program packages containing logistic regression.

Several special tools are available to assess the goodness-of-fit and predictive ability of multiple logistic regression models *(14)*. One frequently used method is the Hosmer–Lemeshow test *(15)*.

### 3.3. Adjusted Odds Ratios (ORs)

The effect size of $X_j$ on the response in logistic regression is best described by means of $\exp(\beta_j)$, because in model 7 it can be shown that

$$\exp(\beta_j) = OR_j \quad (9)$$

where $OR_j$ represents the OR for $X_j$ adjusted for the other explanatory variables. However, the simple equation 9 is only valid in models without interactions. If the model contains interaction terms the OR of one variable depends on the values of other explanatory variables, that is, it is impossible to describe the effect of a variable by means of only one unique OR. In the case of a continuous explanatory variable $X_j$ (and a model without interactions), $OR_j$ according to equation 9 describes the factor by which the odds of an event changes for each one-unit increase of $X_j$. Thus, the presentation of estimated ORs based upon logistic regression with continuous explanatory variables is only meaningful if the unit of the explanatory variables is known.

If the considered event is rare (risk lower than 10%), ORs can be interpreted as risk ratios because the corresponding numbers are approximately identical. This is usually valid in case-control studies. However, in cohort studies investigating common diseases, ORs cannot be interpreted as risk ratios, because the absolute value of ORs is larger than that of risk ratios. In this case the reporting of the results in terms of risk ratios requires additional calculations *(16)*.

### 3.4. Example

Gorini et al. *(17)* investigated the association between alcohol intake and Hodkin's lymphoma by means of logistic regression. A population-based case-control study of 363 Hodkin's lymphoma cases and 1,771 controls was conducted yielding information about sociodemographic characteristics, tobacco, and alcohol consumption. Because a significant interaction between smoking and alcohol consumption was found separate analyses for nonsmokers and ever-smokers were performed taking the potential confounders gender, age, area of residence, education level, and type of interview into account. We consider only the results for nonsmokers. For the alcohol intake categories 0.1–9.0, 9.1–17.9, 18.0–31.7, and >31.7 g/day, the adjusted ORs (95% CIs) 0.45 (0.27–0.74), 0.52 (0.30–0.90), 0.27 (0.13–0.57), and 0.35 (0.15–0.79) compared with never-drinkers were reported *(17)*. Overall, the adjusted OR for ever-drinkers compared with never-drinkers was OR=0.46. This result means that in nonsmokers alcohol consumption has a protective effect because the risk of Hodkin's lymphoma is approximately halved in ever-drinkers compared with never-drinkers. However, a clear decreasing trend of the risk with increasing alcohol consumption could not be found.

## 4. Cox Regression

### 4.1. Time to Event Data

Logistic regression can be used when time is not relevant for the considered event or the risk of the considered event refers to a fixed time interval, for example 5 years. In cohort studies and clinical trials, frequently the participants are observed for different time periods due to staggered entry or censoring. Those designs lead to time to event data, also called survival or failure time data, where the time T until an event occurs is used as response rather than a binary variable Y. The distribution of time to event data is mostly described by the survival distribution function $S(t)=P(T \geq t)$ estimated by the method of Kaplan and Meier *(18)*. The impact of a categorical explanatory variable can be assessed by estimating Kaplan–Meier curves for each group. However, to assess the association of several possibly continuous explanatory

variables $X_1,\ldots,X_k$ with the response variable T a multiple regression model is required. For analyzing time to event data, frequently the Cox regression model is the first considered approach, and this method is probably the most widely used statistical model in general medical research *(19)*.

### 4.2. Modeling of Hazard Function

The essential idea of Cox *(20)* was to model the hazard function rather than the mean of T in dependence on the values of the explanatory variables. The hazard function of a survival time variable T is defined as

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(T \le t + \delta t \mid T \ge t)}{\delta t} \qquad (10)$$

Other names for the hazard function are hazard rate, intensity function, instantaneous death rate, and force of mortality. Roughly speaking, the hazard function is the probability that a person who is alive at time t will die in the next moment after t per unit of time. Cox proposed to model the hazard function as product of an arbitrary unspecified baseline hazard $\lambda_0(t)$ and an exponential term that is linear in the values of the explanatory variables $X_1,\ldots,X_k$

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 x_1 + \ldots + \beta_k x_k) \qquad (11)$$

The baseline hazard $\lambda_0(t)$ represents the hazard function for an individual with $x_1=\ldots=x_k=0$. Model 11 forces the hazard ratio (HR) of two persons to be constant over time. Therefore, model 11 is also called proportional hazards model.

Due to the unspecified baseline hazard $\lambda_0(t)$ the Cox model is a semiparametric model in which full maximum likelihood estimation is not possible. However, parameter estimation can be performed by means of the so-called maximum partial likelihood estimation *(21)*.

Several graphical tools and formal tests have been proposed to investigate the goodness-of-fit of Cox models, especially to assess the adequacy of the proportional hazards assumption. A comprehensive overview is presented by Sasieni *(19)*.

### 4.3. Adjusted HRs

An important feature of the Cox model (without interactions) is that

$$\exp(\beta_j) = HR_j \qquad (12)$$

where $HR_j$ represents the hazard ratio for $X_j$ adjusted for the other explanatory variables. For a binary explanatory variable HR is the ratio of the hazards between the two groups and for a continuous explanatory variable HR represents the factor by which the hazard changes for each one-unit increase of the explanatory variable. As discussed before, in models containing interactions it

is not possible to describe the effect of an explanatory variable by means of one unique effect measure.

### 4.4. Example

The Cox regression model was used in the Seven Countries Study to analyze the association between cigarette smoking and total and cause-specific mortality *(22)*. In short, the Seven Countries Study is a longitudinal observational study of risk factors for coronary heart disease (CHD) in 16 cohorts situated in seven countries (Europe, United States, and Japan). The baseline examination of 12,763 men aged 40 to 59 years took place between 1957 and 1964. In the analysis of the 25-year follow-up data for the outcomes total mortality, CHD mortality, and lung cancer mortality the adjusted hazard ratios per 20 cigarettes/day 1.7 ($p$<0.001), 1.7 ($p$<0.001), and 4.2 ($p$<0.001) were reported *(22)*. In the corresponding Cox models baseline cohort residence, body mass index, serum cholesterol, systolic blood pressure, and presence of clinical cardiovascular disease were included as covariates. These findings confirm that smoking is a risk factor for total, CHD, and lung cancer mortality.

## 5. Poisson Regression

### 5.1. Modeling of Counts

For a response variable Y that represents counts or frequencies it is frequently reasonable to assume that the logarithm of the mean $\mu$ of Y is linearly related to the values of the explanatory variables $X_1,\ldots,X_k$, i.e.,

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad (13)$$

which is equivalent to an exponential relationship between the mean and the explanatory variables

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = \exp(\beta_0) \times \exp(\beta_1) x_1 \times \ldots \times \exp(\beta_k) X_k \qquad (14)$$

Thus, a one-unit increase of $X_j$ has a multiplicative effect given by $\exp(\beta_j)$ on the expected frequency $\mu$.

Model 13 is called Poisson regression because maximum likelihood estimation of the regression coefficients can be performed based upon the assumption that the response variable Y is Poisson distributed *(23)*. The so-called log-linear models for contingency tables represent a special case of model 13 where only categorical explanatory variables are considered.

### 5.2. Modeling of Rates

Poisson regression models play an important role in epidemiologic research, because they also can be used in cohort studies to

analyze time to event data. Here, one makes use of the so-called Poisson process, in which the waiting times between successive events are independent and exponentially distributed with mean $1/\lambda$. Then, the number $Y(t)$ of events that occur up to time $t$ follows a Poisson distribution with mean $\mu=\lambda t$. Note that the mean of $Y(t)/t$ equals $\lambda$. Applying model 13 for $\lambda$

$$\log(\lambda) = \log(\mu / t) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad (15)$$

is equivalent to

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \log(t) \qquad (16)$$

where $\log(t)$ is called offset, because it can be considered as explanatory variable with regression coefficient set equal to 1.

Model 15 also can be written as

$$\lambda = \exp(\beta_0) \times \exp(\beta_1 x_1 + \ldots + \beta_k x_k) \qquad (17)$$

By defining $\lambda_0=\exp(\beta_0)$ as baseline rate we should notice the similarity between formulas 17 and 11. For a reasonable application of Poisson regression to time to event data, it is required to organize the data in an event-time table defined by a cross-classification over a set of time intervals and categories of the explanatory variables. A relatively fine stratification should be used because it is assumed that the hazard in each cell is constant. This could mean that the table is based on individual subjects and the only grouping refers to small time units in which approximately an exponential distribution can be assumed *(24)*.

The Poisson regression approach represents an important alternative to the Cox model because it is an efficient and intuitive method for handling of cumulative exposures, for allowing dependence of risk on multiple time scales, for direct estimation of the survival distribution function, and for direct estimation of relative risks and risk differences *(24,25)*.

An important goodness-of-fit measure suitable for Poisson regression models is given by the deviance, which compares the likelihood for the model that perfectly fits the data with the likelihood of the model under consideration *(26)*. Other methods are summarized by Seeber *(26)*. However, the usefulness of common methods to assess goodness-of-fit of Poisson regression models is limited in the case of large detailed event-time tables *(24)*.

### 5.3. Example

Romundstad et al. *(27)* used Poisson regression to analyze cancer incidence in a cohort of 2,620 male workers in the Norwegian silicon carbide industry. Cumulative exposures to total dust, silicon carbide fibers, and others were assessed by means of a job-exposure matrix. In Poisson regression models age and calendar period of diagnosis were taken into account. Concerning the outcome lung cancer and the exposure groups 0.1–0.9, 1.0–4.9, and ≥5

silicon carbide fibers per ml/year the adjusted risk ratios (95% CI) 3.0 (1.6–5.9), 2.9 (1.4–6.0), and 4.4 (2.1–9.0) compared with unexposed workers were reported *(27)*. These findings suggested an increased risk of lung cancer in workers of the silicon carbide industry due to work related agents such as silicon carbide fibers.

## 6. Model Building

In subheadings 2. to 5. the basic features of regression models are summarized focusing on the most important application areas in epidemiology and the interpretation of the main results. To apply these models adequately in practice, fundamental statistical knowledge about model building is required. In short, model building consists of the following interactive steps: choice of dependent and explanatory variables, model fitting, and model checking. Important issues in model building are given by the functional form of the considered variables (*see* also **Subheading 7.**), interactions, correlation among explanatory variables, outliers, sample size, and number of events for binary and time to event data. In general, a useful model can be obtained if an appropriate balance between goodness-of-fit, simplicity, and consistency with subject-matter knowledge is possible *(28)*. The choice of a model building strategy depends on the aim of the analysis and the specific data situation. More detailed guidelines how to use regression methods in practice are given in other textbooks and articles *(2,7,8,14,29–32)*.

## 7. Model Extensions

Only the basic standard formulations of the most important types of regression models are described so far. Several generalizations and extensions have been developed allowing more flexible applications. In the following sections, a short overview with main references is given.

### *7.1. Generalized Linear Models*

At first, it should be noted that linear regression, logistic regression, and Poisson regression are special types of the class of generalized linear models (GLMs) *(2)*. Let $\mu$ be the mean of the response variable Y and $x_1,\ldots,x_k$ the observed values of the explanatory variables $X_1,\ldots,X_k$. Then, the GLM is defined by

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \qquad (18)$$

where $g(\cdot)$ is called link function. For continuous response variables the use of the identity link leads to linear regression. For categorical responses, logistic regression and Poisson regression are obtained by using the logit link and the log link, respectively. The use of other link functions leads to other models. For example, in dose-response studies the so-called probit link, which represents the inverse of the cumulative distribution function of the standard normal distribution, is frequently used. The corresponding model is called probit regression. A comprehensive overview of GLMs is given by McCullagh and Nelder *(2)*.

### 7.2. Extensions of Binary Logistic Regression

As mentioned, logistic regression can be applied in cohort studies as well as in case-control studies. For matched case-control studies, however, a special procedure is required, the so-called conditional logistic regression *(14,33)*. Further extensions of the standard binary logistic regression have been developed to analyze nominal and ordinal response variables. The most important models are given by polytomous logistic regression *(34)* and the proportional odds model *(35)*. A nontechnical introduction to regression models for ordinal data is given by Bender and Grouven *(36)*, more comprehensive overviews are provided by Bender and Benner *(37)* and in several textbooks *(2,8,14)*.

### 7.3. Extensions of Cox Model

The most important generalizations of the Cox model are given by stratified Cox models and Cox models with time-dependent covariates *(30)*. Stratified Cox models allow for separate baseline hazards for different groups and can be used to take categorical covariates into account for which the proportional hazards assumption is not fulfilled. Cox models including time-dependent covariates can be used in situations in which the values of explanatory variables change over time. However, caution is advised for model fitting and interpretation of the results *(38,39)*. Especially in radiation research, besides the Cox model 11, which represents a multiplicative model for the hazard function, also additive hazard models are frequently used *(24,40)*. It is also common to incorporate external standard rates in the analysis leading to standardized mortality ratio regression *(40)*. Recent research focuses on developing additive hazard models including external standard rates to estimate relative survival in population-based cancer studies *(41)*.

### 7.4. Fractional Polyomials

One basic assumption of the presented regression models is linearity, that is, the assumption that the outcome is linearly related to all explanatory variables $X_1,\ldots,X_k$. This strong assumption is frequently violated in practice so that a more flexible approach is required. One possibility is to use transformations of the

explanatory variables. A useful approach that is developed to deal simultaneously with several continuous explanatory variables is given by fractional polynomials *(42,43)*. Useful introductions to this approach for epidemiologic applications are available *(44,45)* as well as a description of software tools *(46)*.

### 7.5. Nonlinear Models

Although possibly nonlinear transformations of explanatory variables such as fractional polynomials are used, all regression models discussed so far belong to the class of linear models in the sense that they are linear in the regression coefficients after appropriate transformation of the response. Sometimes relations between explanatory variables and response variables cannot be adequately described by linear models, but models that are nonlinear in the regression coefficients are suitable. These methods are called nonlinear regression models, and they are frequently used for example in pharmacokinetics. An overview of nonlinear regression models is given in several textbooks *(47–49)*.

### 7.6. Clustered and Correlated Data

Another basic assumption of the standard regression models is that the individual observations used to estimate the regression coefficients are independent. However, many designs and data situations lead to correlated response data, for example, cross-over trials, cluster randomized trials, or repeated measurements over time. In case of correlated or clustered data, the application of standard regression models is in general invalid and leads to incorrect *P* values and CIs. In short, there are two common ways to deal adequately with correlated data, both represent extensions of the class of GLMs. In the generalized estimating equations (GEE) approach *(50)*, the usual formulation of regression models is used. However, the estimation procedure accounts for correlation and a robust estimation of SEs and CIs is used *(51,52)*. An alternative to GEE modeling is given by generalized linear mixed models (GLMMs) *(53)*, also called random effects models, random coefficient models, multilevel models, or hierarchical models. In the context of survival time data the term frailty model is used *(54)*. Mixed models extend standard regression models by adding random effects. A random effect is permitted to vary from subject to subject or from cluster to cluster leading to a specific correlation structure of the considered data. A comprehensive overview of GLMMs is given by Brown and Prescott *(53)*, a nontechnical introduction in the context of public health research is provided by Diez-Roux *(55)*.

### 7.7. Missing Data and Measurement Error

Finally, two frequent complications in practical applications of regression models in epidemiology should be mentioned: missing values and measurement error. In standard applications of regression analyses it is assumed that all data are available and all data are measured without error. Frequently, however, more or less data

values are missing or some explanatory variables can be observed only with substantial measurement error. Ignoring missing values and measurement error could lead to substantial bias. Thus, it is important to investigate the reasons leading to missing data or measurement error. An overview of techniques to deal with missing values are given by Little and Rubin *(56)*, regression methods accounting for measurement error are described by Carroll et al. *(57)*.

## References


1. Matthews DE. (2005). Linear regression, simple. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 4. Chichester, UK: Wiley, pp. 2812–2816.
2. McCullagh P, Nelder JA. (1989). Generalized Linear Models, 2nd ed. New York: Chapman & Hall.
3. Srivastava MS. (2002). Methods of Multivariate Statistics. New York: Wiley.
4. Anderson TW. (2003). An Introduction to Multivariate Statistical Analysis, 3rd ed. New York: Wiley.
5. Krzanowski WJ. (2005). Multivariate multiple regression. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 5. Chichester, UK: Wiley, pp. 3552–3553.
6. Matthews DE. (2005). Multiple linear regression. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 5. Chichester, UK: Wiley, pp. 3428–3441.
7. Draper NR, Smith H. (1998). Applied Regression Analysis, 3rd ed. New York: Wiley.
8. Harrell FE Jr. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer.
9. Cook DR, Weisberg S. (1997). Graphics for assessing the adequacy of regression models. J Am Stat Assoc **92,** 490–499.
10. Chan SC, Liu CL, Lo CM, et al. (2006). Estimating liver weight of adults by body weight and gender. World J Gastroenterol **12,** 2217–2222.
11. Anderson JA. (1972). Separate sample logistic discrimination. Biometrika **59,** 19–35.
12. Mantel N. (1973). Synthetic retrospective studies and related topics. Biometrics **29,** 479–486.
13. Levy PS, Stolte K. (2000). Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. Stat Methods Med Res **9,** 41–55.
14. Hosmer DW Jr, Lemeshow S. (2000). Applied Logistic Regression, 2nd ed. New York: Wiley.
15. Hosmer DW, Lemeshow S. (1980). Goodness-of-fit tests for the multiple logistic regression model. Commun Stat Theory Methods **9,** 1043–1069.
16. Davies HTO, Crombie IK, Tavakoli M. (1998). When can odds ratios mislead? BMJ **316,** 989–991.
17. Gorini G, Stagnaro E, Fontana V, et al. (2007). Alcohol consumption and risk of Hodgkin's lymphoma and multiple myeloma: a multicentre case-control study. Ann Oncol **18,** 143–148.
18. Kaplan EL, Meier P. (1958). Nonparametric estimator from incomplete observations. J Am Stat Assoc **53,** 457–481.
19. Sasieni P. (2005). Cox regression model. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 2. Chichester, UK: Wiley, pp. 1280–1294.
20. Cox DR. (1972). Regression models and life tables (with discussion). J R Stat Soc B **34,** 187–220.
21. Cox DR. (1975). Partial likelihood. Biometrika **62,** 269–276.
22. Jacobs DR Jr, Adachi H, Mulder I, et al. (1999). Cigarette smoking and mortality risk: twenty-five-year follow-up of the Seven Countries Study. Arch Intern Med **159,** 733–740.
23. Frome EL, Kutner MH, Beauchamp JJ. (1973). Regression analysis of Poisson-distributed data. J Am Stat Assoc **68,** 935–940.
24. Preston DL. (2005). Poisson regression in epidemiology. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 6. Chichester, UK: Wiley, pp. 4124–4127.
25. Spiegelman D, Hertzmark E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol **162,** 199–200.
26. Seeber GUH. (2005). Poisson regression. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics, 2nd ed., vol. 6. Chichester, UK: Wiley, pp. 4115–4124.

27. Romundstad P, Andersen A, Haldorsen T. (2001). Cancer incidence among workers in the Norwegian silicon carbide industry. Am J Epidemiol **153,** 978–986.
28. Royston P. (2000). A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med **19,** 1831–1847.
29. Harrell FE Jr, Lee KL, Mark DB. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med **15,** 361–387.
30. Hosmer DW Jr, Lemeshow S. (1999). Applied Survival Analysis: Regression Modelling of Time to Event Data. New York: Wiley.
31. Bagley SC, White H, Golomb BA. (2001). Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol **54,** 979–985.
32. Katz MH. (2003). Multivariable analysis: A primer for readers of medical research. N Engl J Med **138,** 644–650.
33. Breslow NE, Day NE. (1980). Statistical Methods in Cancer Research Vol. I: The Analysis of Case-Control Studies. Lyon, France: International Agency for Research on Cancer.
34. Engel J. (1988). Polytomous logistic regression. Stat Neerl 42: 233–252.
35. McCullagh P. (1980). Regression models for ordinal data (with discussion). J R Stat Soc B **42,** 109–142.
36. Bender R, Grouven U. (1997). Ordinal logistic regression in medical research. J R Coll Physicians Lond **31,** 546–551.
37. Bender R, Benner A. (2000). Calculating ordinal regression models in SAS and S-Plus. Biom J **42,** 677–699.
38. Andersen PK. (1992). Repeated assessment of risk factors in survival analysis. Stat Methods Med Res **1,** 297–315.
39. Altman DG, De Stavola BL. (1994). Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. Stat Med **13,** 301–341.
40. Breslow NE, Day NE. (1987). Statistical Methods in Cancer Research Vol. II: The Design and Analysis of Cohort Studies. Lyon, France: International Agency for Research on Cancer.
41. Dickman PW, Sloggett A, Hills M, Hakulinen T. (2004). Regression models for relative survival. Stat Med **23,** 51–64.
42. Royston P, Altman DG. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat **43,** 429–467.
43. Sauerbrei W, Royston P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Society **162,** 71–94.
44. Royston P, Ambler G, Sauerbrei W. (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol **28,** 964–974.
45. Royston P, Sauerbrei W. (2005). Building multivariable regression models with continuous covariates in clinical epidemiology–with an emphasis on fractional polynomials. Methods Inf Med **44,** 561–571.
46. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. (2006). Multivariable regression building by using fractional polynomials: description of SAS, STATA and R programs. Comput Stat Data Anal **50,** 3646–3485.
47. Bates DM, Watts DG. (1988). Nonlinear Regression Analysis and its Applications. New York: Wiley.
48. Seber GAF, Wild CJ. (1989). Nonlinear Regression. New York: Wiley.
49. Ratkowsky DA. (1990). Handbook of Nonlinear Regression Models. New York: Marcel Dekker.
50. Liang K-Y, Zeger SL. (1986) Longitudinal data analysis using generalized linear models. Biometrika **73,** 13–22.
51. Burton P, Gurrin L, Sly P. (1998). Tutorial in biostatistics: extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med **17,** 1261–1291.
52. Hanley JA, Negassa A, Edwardes MD, Forrester JE. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. Am J Epidemiol **157,** 364–375.
53. Brown H. (2006). Applied Mixed Models in Medicine, 2nd ed. Chichester, UK: Wiley.
54. McGilchrist CA. (1993). REML estimation for survival models with frailty. Biometrics **49,** 221–225.
55. Diez-Roux AV. (2000). Multilevel analysis in public health research. Annu Rev Public Health **21,** 171–192.
56. Little RJA, Rubin DB. (2002). Statistical Analysis with Missing Data, 2nd ed. Hoboken, NJ: Wiley.
57. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. (2006). Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed. London, UK: Chapman & Hall.

http://www.springer.com/978-1-58829-987-1

Cancer Epidemiology
Volume 1, Host Susceptibility Factors
(Ed.)M. Verma
2009, X, 704 p. 27 illus., Hardcover
ISBN: 978-1-58829-987-1

# Contents

*Preface* . . . . . *v*
*Contributors* . . . . . *ix*
*Contents of Volume II.* . . . . . *xi*

PART I: CANCER INCIDENCE, PREVALENCE, MORTALITY AND SURVEILLANCE

1. Cancer Occurrence . . . . . 3
***Ahmedin Jemal, Melissa M. Center, Elizabeth Ward, and Michael J. Thun***
2. Cancer Registry Databases: *An Overview of Techniques of Statistical Analysis and Impact on Cancer Epidemiology* . . . . . 31
***Ananya Das***
3. Breast Cancer in Asia . . . . . 51
***Cheng-Har Yip***
4. Cancer Epidemiology in the United States: Racial, Social, and Economic Factors . . . . . 65
***Dana Sloane***
5. Epidemiology of Multiple Primary Cancers . . . . . 85
***Isabelle Soerjomataram and Jan Willem Coebergh***
6. Cancer Screenings, Diagnostic Technology Evolution, and Cancer Control . . . . . 107
***Fabrizio Stracci***
7. Thriving for Clues in Variations seen in Mortality and Incidence of Cancer: *Geographic Patterns, Temporal Trends, and Human Population Diversities in Cancer Incidence and Mortality* . . . . . 137
***Alireza Mosavi-Jarrahi and Mohammad Ali Mohagheghi***

PART II: METHODS, TECHNOLOGIES AND STUDY DESIGN IN CANCER EPIDEMIOLOGY

8. Evaluation of Environmental and Personal Susceptibility Characteristics That Modify Genetic Risks . . . . . 163
***Jing Shen***
9. Introduction to the Use of Regression Models in Epidemiology . . . . . 179
***Ralf Bender***
10. Proteomics and Cancer Epidemiology . . . . . 197
***Mukesh Verma***
11. Different Study Designs in the Epidemiology of Cancer: *Case-Control vs. Cohort Studies* . . . . . 217
***Harminder Singh and Salaheddin M. Mahmud***
12. Methods and Approaches in Using Secondary Data Sources to Study Race and Ethnicity Factors . . . . . 227
***Sujha Subramanian***
13. Statistical Methods in Cancer Epidemiologic Studies . . . . . 239
***Xiaonan Xue and Donald R. Hoover***

14. Methods in Cancer Epigenetics and Epidemiology . . . . . 273
***Deepak Kumar and Mukesh Verma***

PART III: HOST SUSCEPTIBILITY FACTORS IN CANCER EPIDEMIOLOGY

15. Mitochondrial DNA Polymorphism and Risk of Cancer . . . . . 291
***Keshav K. Singh and Mariola Kulawiec***

16. Polymorphisms of DNA Repair Genes: *ADPRT, XRCC1 and XPD and Cancer Risk in Genetic Epidemiology* . . . . . 305
***Jun Jiang, Xiuqing Zhang, Huanming Yang and Wendy Wang***

17. Risk Factors and Gene Expression in Esophageal Cancer . . . . . 335
***Xiao-chun Xu***

18. Single Nucleotide Polymorphisms in DNA Repair Genes and Prostate Cancer Risk . . . . . 361
***Jong Y. Park, Yifan Huang and Thomas A. Sellers***

19. Linking the Kaposi's Sarcoma-Associated Herpesvirus (KSHV/HHV-8) to Human Malignancies . . . . . 387
***Inna Kalt, Shiri-Rivka Masa and Ronit Sarid***

20. Cancer Cohort Consortium Approach: *Cancer Epidemiology in Immunosuppressed Groups* . . . . . 409
***Diego Serraino, Pierluca Piselli for the Study Group***

21. Do Viruses Cause Breast Cancer? . . . . . 421
***James S. Lawson***

22. Epidemiology of Human Papilloma Virus (HPV) in Cervical Mucosa . . . . . 439
***Subhash C. Chauhan, Meena Jaggi, Maria C. Bell, Mukesh Verma and Deepak Kumar***

23. Epigenetic Targets in Cancer Epidemiology . . . . . 457
***Ramona G. Dumitrescu***

24. Epidemiology of Lung Cancer Prognosis: *Quantity and Quality of Life* . . . . . 469
***Ping Yang***

25. Hereditary Breast and Ovarian Cancer Syndrome: *The Impact of Race on Uptake of Genetic Counseling and Testing* . . . . . 487
***Michael S. Simon and Nancie Petrucelli***

*Index* . . . . . *501*