Langner, Ingo; Bender, Ralf; Lenz-Tönjes, Rebecca; Küchenhoff, Helmut; Blettner, Maria

Working Paper

Bias of maximum-likelihood estimates in logistic and Cox regression models: a comparative simulation study

Discussion Paper, No. 362

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Langner, Ingo; Bender, Ralf; Lenz-Tönjes, Rebecca; Küchenhoff, Helmut; Blettner, Maria (2003) : Bias of maximum-likelihood estimates in logistic and Cox regression models: a comparative simulation study, Discussion Paper, No. 362, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1737

This Version is available at:
https://hdl.handle.net/10419/31093

BIAS OF MAXIMUM-LIKELIHOOD ESTIMATES IN LOGISTIC AND COX REGRESSION MODELS: A COMPARATIVE SIMULATION STUDY

Ingo Langner[1], Ralf Bender[2], Rebecca Lenz-Tönjes[1], Helmut Küchenhoff[2], Maria Blettner[2]

[1]Department of Epidemiology and Medical Statistics,
School of Public health University of Bielefeld
[2]Institute of Medical Biometry, Epidemiology & Informatics,
University of Mainz
[3]Department of Statistics, Ludwig Maximilians Universität München

SFB 386 Discussion Paper 362

ABSTRACT

Parameter estimates of logistic and Cox regression models are biased for finite samples. In a simulation study we investigated for both models the behaviour of the bias in relation to sample size and further parameters. In the case of a dichotomous explanatory variable *x* the magnitude of the bias is strongly influenced by the baseline risk defined by the constants of the models and the risk resulting for the high risk group. To conduct a direct comparison of the bias of the two models analyses were based on the same simulated data. Overall, the bias of the two models appear to be similar, however, the Cox model has less bias in situations where the baseline risk is high.

1. INTRODUCTION

Logistic regression and Cox regression are frequently used for analyzing study data in medical research. Both methods analyze the relation of time dependent events to other influencing variables. The logistic model describes the event probability in a distinct observation time window and the Cox model the instantaneous event probability at a given time point both in dependence on explanatory variables. For the logistic regression the individual binary information "event occurred: yes or no" is used whereas the Cox regression considers the individual time till the event occurs. In general the method for estimating the parameters is based on the likelihood or partial likelihood, respectively. It is well known that

the estimates of both Maximum-Likelihood (ML) methods are only asymptotically unbiased which results in a bias for finite samples (1, 2, 3). A more or less relevant bias is present especially for small samples (4, 5), which is the case for many epidemiological and clinical studies, the bias could be relevant for parameter estimation, but a clear rule for a threshold for sample size cannot be found in the literature, as the size and direction of the bias does not rely on this issue alone. The bias of ML-estimates is directed away from zero which means that the expectation of the estimate is always larger in absolute value than the true parameter (2, 3). However, results of a systematic observation of this relations of the single regression methods have not been published.

Comparisons between logistic regression and Cox regression models describing the occurrence of a dichotomous event in a distinct observation time interval were mainly related to the similarity of the parameter estimates for the same explanatory variable deriving from the different methods (6, 7, 8, 10). The bias of the parameter estimates was not investigated in these papers. The most consistent finding was that the estimates are nearly identical in the case of a rare event and a short observation time interval. Green and Symons (8) gave a short overview for the theoretical reasons of this. Increasing event probability or/and increasing observation time interval led to increasing difference of the estimates. Peduzzi et al. (7) examined a binary explanatory variable x in the models and suggested that the agreement between the estimates depends on the baseline event probability and to some extend on the probability ratio of x=1 and x=0. Ingram and Kleinman (6) discussed the influence of non-exponential survival times and non-proportional hazards as strengthening effects for the difference. They also found that sample size had no or only little effect on the difference of the estimates independent of the event probability (6). No considerations concerning a possible difference of the bias of the two methods have been made. Annesi et al. (9) found that the asymptotic relative efficiency of logistic regression model and Cox regression model is very close to 1 unless the event probability is increasing. They concluded that the Cox model is superior to the logistic model mainly when analyzing longitudinal data.

In this paper, the amount of small sample bias in dependence on the data situation is investigated for both the logistic and the Cox regression model. The influence of different parameters on the bias is investigated for the single regression methods and the two approaches are compared with regard to bias and variance of the parameter estimates by means of simulation.

2. RELATIONS BETWEEN THE LOGISTIC AND THE COX REGRESSION MODEL

The model for the logistic regression describes the dependence of the odds of an event on explanatory variables X with parameters β. Let Y the binary response variable indicating the occurrence of an event in a distinct time interval ranging from 0 to t_z and let α the constant defining the risk in the case of all x_i=0 than the relation is given by

$$\frac{P(Y=1 \mid X, t_z)}{1-P(Y=1 \mid X, t_z)} = e^{(\alpha+\beta X)} \qquad (1)$$

whereas the Cox regression models the hazard of an event depending on explanatory variables X:

$$h(t) = h_0(t)\, e^{(\theta X)} \qquad (2)$$

Here, the response variable is given as the time till the event occurs, which is commonly called 'survival time'. The baseline hazard $h_0(t)$, which is characterized by the underlying baseline survival time distribution defining the risk in the case of all x_i=0, is unspecified in the Cox regression model. Although we were only interested in the bias of the parameter estimate of the explanatory variable X the parameters specifying the survival time distribution are needed to describe the behaviour of this bias as we will show below.

The bias of the estimate of the parameters β and θ will be investigated, respectively. Is x a dichotomous covariable and is t_z the length of the observation time than β is given by

$$\beta = \log\left[\frac{P(Y=1 \mid x=1, t_z) \,/\, P(Y=0 \mid x=1, t_z)}{P(Y=1 \mid x=0, t_z) \,/\, P(Y=0 \mid x=0, t_z)}\right] = \log(OR) \qquad (3)$$

in the logistic regression model, which is equal to the log odds ratio (OR) and in the Cox regression model θ is

$$\theta = \log\left[\frac{h(t, x=1)}{h(t, x=0)}\right] = \log(HR) \qquad (4),$$

which is equal to the log of the hazard ratio (HR).

To conduct a direct comparison of the bias of the two regression methods both analyses were done for the same sample data. Data simulated by survival time simulation models can be transformed into a dataset suitable for logistic regression if as a restriction censoring only at the end of the observation time window is allowed. For this, the binary censoring variable indicating whether a observed survival time ended with an event or not is used as the binary response variable for the logistic regression model. The true parameters of the logistic model can be calculated using the parameters of the survival time simulation model and the equations for risk for a distinct observation time or distinct time point, respectively, of the logistic and the Cox model (see appendix I). Equating the risk formulae and solving for the parameters of the logistic model gives exact relations between the logistic and the Cox parameters. However, this relations hold only true if the observation time length was the same for all individuals.

3. SIMULATION STUDY

We wanted to compare the performance of the covariable parameters estimates when the regression methods are applied to sample datasets. For this, the true underlying parameter values determining the covariable values in the dataset have to be known. Therefore we used simulation models to produce the datasets. For each constellation of true parameters and sample size 10,000 datasets were simulated and analyzed. To keep models simple we included only one binary covariable $x \sim B(1, 0.5)$ for which the bias was investigated.

Data for the logistic regression were simulated by using the logistic model as a parameter for a Bernoulli distribution resulting in a dichotomous response variable y indicating whether an event occurred or not.

$$y \sim B(1,p) \qquad \text{with} \qquad p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{5}$$

The intercept α determines the baseline risk $P_0(Y=1|X=0)$ and β is the parameter for the covariable x.

For the Cox regression the survival time data t were simulated by the inverse of the survival time distribution function referring to the Cox model applied in the regression analysis. The

event probability is replaced by a uniformly distributed random number z with a value range from 0 to 1 (11). The distribution function also determines the baseline risk and baseline hazard of the Cox model, respectively. To evaluate a possible influence of the chosen survival time distribution we used three different of them: the exponential (6), the Weibull (7), and the Gompertz (8) distribution.

The simulation models for the survival times are, respectively:

(Exponential) $$t = \frac{-\log(1-z)}{\lambda \ \exp(\theta\, x)} \tag{6}$$

(Weibull) $$t = \left[\frac{-\log(1-z)}{\kappa \ \exp(\eta\, x)}\right]^{\frac{1}{\gamma}} \tag{7}$$

(Gompertz) $$t = \frac{1}{\delta}\log\left(1 - \frac{\delta \ \log(1-z)}{\tau \ \exp(\nu\, x)}\right) \tag{8}$$

Here λ, κ and γ, τ and δ determine the baseline risk and θ, η, ν are the parameters for the covariable, respectively. A distinct observation time window was defined for censoring of the survival time data. No further censoring was simulated to allow also the application of logistic regression.

The intercept and the survival time distribution parameters were selected that the baseline risk $P_0(Y=1|X=0)$ was approximately between 0.2 and 0.8 for the logistic simulation models and 0.07 and 0.88 for the Cox simulation models. Additionally the covariable parameter was varied resulting in a range for risk $P_1(Y=1|X=1)$ of 0.08 to 0.92 for logistic models and of 0.03 to 0.97 for the Cox models. Further variations of the parameters for the logistic simulation model to extend the range for risk P_1 led to an increase of simulated data sets where the regression analysis did not converge. For the direct comparison of the bias of the logistic and the Cox regression the variation of the parameters was restricted to the risk range that was suitable for the logistic simulation models. To get an impression of the relative effect of these parameters compared to the variation of sample size, this parameter was varied in an

interval between 100 and 500, for which a relevant sample size dependent bias has been reported (5).

The bias is defined as the mean deviation of the estimate from the true parameter. For a specific parameter constellation k=10,000 simulated datasets were analyzed. So, the bias of a parameter β of a covariable x in the logistic model is estimated as

$$bias_{Log} = \frac{1}{k} \cdot \sum_{i=1}^{k} (\hat{\beta}_i - \beta) \qquad (9)$$

and correspondingly for θ, η, and ν in the Cox model.

Bias estimates of logistic regression and Cox regression, both based on the same 10,000 datasets, were compared by calculating the difference of the percentage values (PBD: percentage bias difference):

$$PBD = \left(\frac{bias_{Log}}{\beta} - \frac{bias_{Cox-\exp}}{\theta} \right) \cdot 100 \qquad (10)$$

and correspondingly for the estimates of the Weibull ($bias_{Cox\text{-}wei}$, η) and Gompertz ($bias_{Cox\text{-}gom}$, ν) distributed data.

To get an impression whether the estimates from logistic regression or those from Cox regression are closer to there true parameter values for a distinct set of true parameters, the mean difference of the absolute deviation found in the analyses of a single dataset was calculated (APED: absolute percentage error difference):

$$APED = \frac{1}{k} \sum_{i=1}^{k} \left(\left| \frac{\hat{\beta}_i - \beta}{\beta} \right| - \left| \frac{\hat{\theta}_i - \theta}{\theta} \right| \right) \cdot 100 \qquad (11)$$

where k denotes the number of datasets (10,000). Correspondingly this was done for the parameter estimates of the Cox-regression for the Weibull (η) and Gompertz (ν) distributed data. Saying it simple, APED measures which of the two estimates on an average lies closer to its true parameter value

4. RESULTS

We plotted the bias in dependence of $P_1(Y=1|X=1)$ because o plot using the true parameter values on the horizontal axis was not suitable to show the relations and symmetry of the bias behaviour of the different models.

Fixing all parameters except of the explanatory variable parameter, the resulting bias of the logistic regression models, when plotted versus the risk $P_1(Y=1|X=1)$, produces a curve shown in Fig. 1a. The bias is monotone increasing with increasing risk P_1. Around $P_1=0.5$ the increase is small but towards values closer to 0 or 1 the bias becomes very large, respectively. The point of intersection of the bias graph with the horizontal axis bias=0 is determined by the intercept parameter, e.g. the baseline risk.

A point symmetry regarding the bias estimate for $P_1=0.5$ is found for the bias of the logistic regression.

Variation of the main model parameters leaves the shape of the bias graph unchanged, but the whole curve is shifted vertically downwards with increasing baseline risk (Fig. 1a). This also results in a horizontal shift of the point of intersection of the bias curve with the axis bias=0.

In an interval of 'moderate' P_1 (0.15 to 0.85) the percentage bias of logistic regression varies only little but towards 'extremer' values (<0.15 and >0.85) a strong increase of the bias is present (Fig. 1b). The boundaries of this different behaviour move towards 'extremer' risks with increasing sample size. The point symmetry of the graph of the absolute bias results in a further symmetry when looking at the percentage bias: the combined graphs of complementary baseline risks (Fig. 1b: 0.27 and 0.73) are symmetrical to the vertical axis at $P_1=0.5$. Another interesting fact is that if sample size and baseline risk are constant the smallest percentage bias is found near an risk P_1 that is complementary to the true baseline risk.

As the bias is decreasing with increasing sample size, the bias curve is shifted closer towards the axis bias=0 for higher sample sizes (Fig. 2).

The main facts mentioned so far are also true for the bias of parameter estimates in the Cox regression except of the point symmetry which is not found for the latter (Fig. 3a). The strong decrease of bias for small P_1 is similar to that for logistic regression parameters but the increase at high P_1 starts not till much extremer values. As a consequence, when comparing two complementary baseline risks, the bias estimates of the different risks P_1 for the higher baseline risk are all smaller than for the corresponding baseline risk less than 0.5 (Fig. 3b).

Different survival time distributions have no effect on the bias: the percentage values are identical if the distribution parameters resulted in a corresponding baseline risk in the different models.

The percentage bias of logistic regression is always higher than the corresponding bias of the Cox regression, however, for small P_1 the difference is almost zero (Fig. 4). The difference of the percentage bias's is increasing with increasing risk P_1 and is decreasing with increasing sample size. However, even for a sample size of n=100 the percentage bias's of both regression methods as well as the absolute difference of the percentage bias's do not exceed 6% when looking at moderate risks P_1 only.

The increase of the mean difference of the percentage absolute deviations (APED) of the single estimates with increasing risk P_1 depends on the baseline risk: higher baseline risk results in a steeper slope (Fig. 5). For small baseline risks the increase is almost zero. At small P_1 the mean absolute deviations for logistic regression estimates seem to be a little bit smaller than those for Cox regression estimates.

Again, these findings are not changed when different survival time distributions are used for the simulation of the data.

5. DISCUSSION

We conducted a simulation study to observe the behaviour of the bias of ML parameter estimates in logistic regression and Cox regression in relation to sample size and the true values for baseline risk and the explanatory variable parameter. The baseline risk for the logistic model is characterized by the intercept and for the Cox model by the parameters of the survival time distribution responsible for the baseline hazard.

To our knowledge, no results concerning a systematic comparison of the bias of the two regression methods has been presented so far. Callas et al. (10) used a measure named "percentage relative bias in point estimates", but they observed the difference between the point estimates of two logistic regression methods and the estimate of the Cox regression, respectively.

Results presented here give rise to the presumption that the published differences found for comparisons of the crude estimates of the both methods when analyzing the same explanatory variables (6, 7, 8) are at least partly due to the different bias behaviour of the parameter estimates in the different models.

From our simulations the following conclusions can be drawn.
In the case of a binary explanatory variable *x* the bias of the parameter estimates obtained from the ML regression methods depends not only on the sample size but also on the baseline risk and the risk $P_1(Y=1|X=1)$. The intensity of the influence of a single parameter on the bias is not independent from the other parameters. This relation is different for logistic regression and Cox regression. In general a strong bias is present for extreme baseline risks and for extreme risks $P_1(Y=1|X=1)$. With both regression methods a high bias of the covariable parameter has to be expected if the number of events in the group only affected by the baseline risk is small. For the logistic regression this is also true for the number of non-events which results in the point symmetry of the bias curve.

The following limitations of our study should be considered. Of course, much higher values for the percentage bias in logistic regression than presented here would be expected if small sample sizes and extreme baseline risks would be combined as it was possible for the simulated datasets for Cox regression. As we simulated the sample data by chance the case with no events or no non-events for x=0 occurred more often for extremer baseline risks. In this cases the logistic regression analysis did not converge and no estimate could be obtained, e.g. this cases could not be included when estimating the bias. As an exclusion of these cases would shift the bias estimate, we decided here to present no results for parameter constellations where the converge criteria of the regression were not reached for at least 99.9% of the regarding 10,000 simulations.
Further research is required to investigate the association between the risk of continuous covariables and the parameter bias as well as the influence of additional covariables on the bias.

Despite these limitations the following practical implications can be made.
Although the higher power of the Cox regression allows to extend the interval of possible risks that can be analyzed with small sample sizes when compared to logistic regression this extent is combined with a strong increase in bias, especially for low baseline risks and low $P_1(Y=1|X=1)$.
To define situations when bias correction is required three factors have to be considered: baseline risk, risk of the exposed (here risk $P_1(Y=1|X=1)$), and sample size. Schaefer (5) presented a bias correction method for logistic regression. He recommended to apply this

method to regression analyses if sample size is 200 or less and mentioned no further criteria. However, as shown here, for small sample sizes the bias is rather small when the baseline risk and the risk $P_1(Y=1|X=1)$ are moderate and bias correction methods might not be necessary. However, also if the total number of events is high but the baseline risk is rather small this would lead to strong bias. This means for logistic regression that, if the numbers of events or non-events for the group characterized only by the intercept of the logistic model tend towards zero, the estimates for all covariable parameters would be highly biased. In general an extreme baseline risk and/or an extreme $P_1(Y=1|X=1)$ leads to a strong bias. For logistic regression the boundaries between moderate and extreme risks are symmetrical shifted towards zero and one, respectively, by increasing the sample size. This shift of the boundaries is not symmetrical for Cox regression parameters and is more pronounced towards zero as even for small sample sizes a strong bias increase occurs only at very high risks.

Most critical are the cases where a low baseline risk is present and parameters for protective covariables are estimated. The case vice versa is only relevant for logistic regression.
However, the decision, whether a bias correction is suitable or not, should depend on the ability of the considered correction method to give sufficient results especially in the cases of extreme baseline risk or extreme $P_1(Y=1|X=1)$.

If there is no necessity for doing Cox regression because of other than type I censoring then the additional effort for collecting survival times instead of the binary information ‘event: yes or no’ might be not justified when working on low baseline risks or low risks P_1 for the investigated covariables. Concerning the bias the advantages of Cox regression are most pronounced when analyzing high baseline risks. Another advantage is that at high extreme risks P_1 parameters can be estimated with a rather small bias.

In summary, not only sample size is the main factor for the decision, whether a bias correction is suitable in an analysis based on a logistic or Cox regression model or not, but also the baseline risk and the risk for the exposed (here risk $P_1(Y=1|X=1)$). Concerning the bias the advantages of Cox regression compared to logistic regression are mainly in the case of high baseline risks and/or extreme risks for the exposed.

APPENDIX I

We used the exponential, the Weibull, and the Gompertz distribution for the simulation of survival times. As a restriction censoring is only allowed at the end of the simulated observation time window. The true parameters of the logistic model were calculated using the parameters of the survival time simulation model and the equations for risk *R* for a distinct observation time t_z or distinct time point t_z, respectively, of the logistic and the Cox model.

(Logistic) $$R(t_z) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} \quad (12)$$

(Cox: exponential) $$R(t_z) = 1 - e^{\left(-\lambda t_z \cdot e^{(\theta x)}\right)} \quad (13)$$

(Cox: Weibull) $$R(t_z) = 1 - e^{\left(-\kappa t_z^{\gamma} \cdot e^{(\eta x)}\right)} \quad (14)$$

(Cox: Gompertz) $$R(t_z) = 1 - e^{\left[\frac{\tau}{\delta}\left(1-e^{(\delta t_z)}\right)e^{(\nu x)}\right]} \quad (15)$$

Equating (12) and (13) (or (12) and (14), or (12) and (15), respectively) for x=0 and is λ (or κ, γ, or τ, δ, respectively) the parameter of the survival time model based on an exponential (or Weibull, or Gompertz, respectively) distribution then the intercept of the logistic model, α, is given by:

(Exponential) $$\alpha = \log\left(1-e^{(-\lambda t_z)}\right) + \lambda t_z \quad (16)$$

(Weibull) $$\alpha = \log\left(1-e^{\left(-\kappa t_z^{\gamma}\right)}\right) + \kappa t_z^{\gamma} \quad (17)$$

(Gompertz) $$\alpha = \log\left(1-e^{\left[\frac{\tau}{\delta}\left(1-e^{(\delta t_z)}\right)\right]}\right) - \frac{\tau}{\delta}\left(1-e^{(\delta t_z)}\right) \quad (18)$$

Analogous the parameter β of the binary covariable of the logistic model is calculated for the case x=1 and substituting α by (16) (or (17) or (18), respectively). Here, the parameter θ (or η or ν, respectively) of the survival time model has to be considered:

(Exponential) $$\beta = \log\left(\frac{1-e^{\left(-\lambda t_z \cdot e^{(\theta x)}\right)}}{1-e^{\left(-\lambda t_z\right)}}\right) + \lambda t_z \cdot \left(e^{(\theta x)} - 1\right) \qquad (19)$$

(Weibull) $$\beta = \log\left(\frac{1-e^{\left(-\kappa\, t_z^{\gamma} \cdot e^{(\eta x)}\right)}}{1-e^{\left(-\kappa\, t_z^{\gamma}\right)}}\right) + \kappa t_z^{\gamma} \cdot \left(e^{(\eta x)} - 1\right) \qquad (20)$$

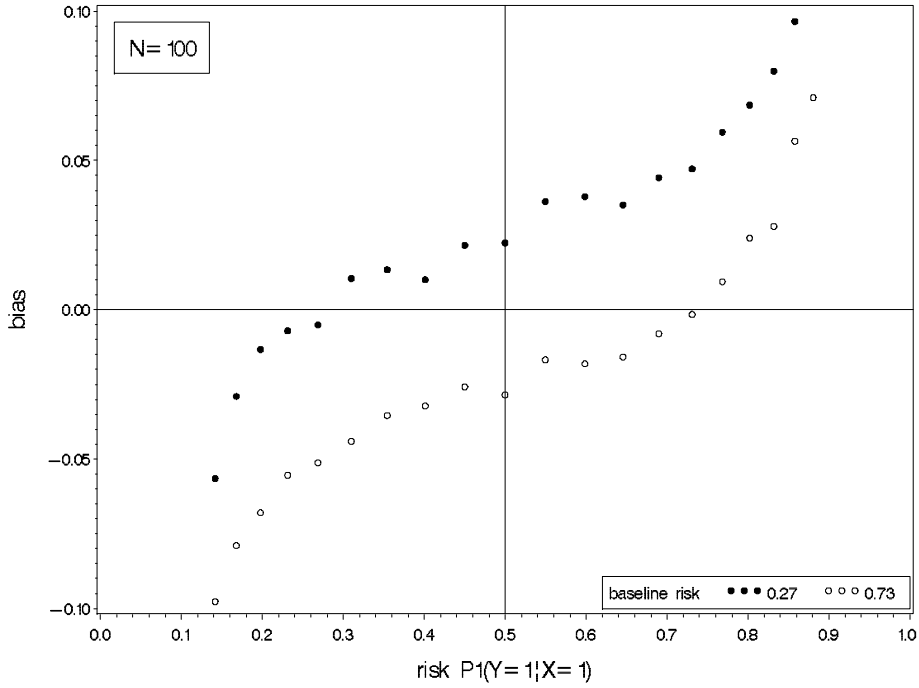
(Gompertz) $$\beta = \log\left(\frac{1-e^{\left[\frac{\tau}{\delta} e^{(\nu x)}\left(1-e^{(\delta t_z)}\right)\right]}}{1-e^{\left[\frac{\tau}{\delta}\left(1-e^{(\delta t_z)}\right)\right]}}\right) + \frac{\tau}{\delta}\left(1-e^{(\delta t_z)}\right)\cdot\left(1-e^{(\nu x)}\right) \qquad (21).$$

REFERENCES

1. Cordeiro, G.M.; McCullagh, P. Bias Correction in Generalized Linear Models. R Statist Soc B **1991**, 3, 629-643.
2. Firth, D. Bias reduction of maximum likelihood estimates. Biometrika **1993**, 89 (1), 27-38.
3. McCullagh, P.; Nelder, J.A. Generalized Linear Models. Chapman and Hall: London, **1989**.
4. Whaley, F.S. Comparison of different Maximum Likelihood Estimators in a small sample logistic regression with two independent binary variables. Statistics in Medicine **1991**,10, 723-731.
5. Schaefer, R.L. Bias correction in maximum likelihood logistic regression. Statistics in Medicine **1983**, 2, 71-78.
6. Ingram, D.D.; Kleinman, J.C. Empirical comparison of proportional hazards and logistic regression models. Statistics in Medicine **1989**, 8, 525-538.
7. Peduzzi, P.; Holford, T.; Detre, K.; Chan, Y.-K. Comparison of the logistic and Cox regression models when outcome is determined in all patients after a fixed period of time. Journal of Chronic Diseases **1987**, 40(8), 761-767.
8. Green, M.S.; Symons, M.J. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. Journal of Chronic Diseases **1983**, 36(10), 715-724.
9. Annesi, I.; Moreau, T.; Lellouch, J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Statistics in Medicine **1989**, 8, 1515-1521.
10. Callas, P.W.; Pastides, H.; Hosmer, D.W. Empirical comparison of proportional hazards, poisson, and logistic regression modeling of occupationl cohort data. American Journal of Industrial Medicine **1998**, 33, 33-47.
11. Bender, R.; Augustin, T.; Blettner, M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine **2003** (submitted for publication)

Figures

1a)



1b)

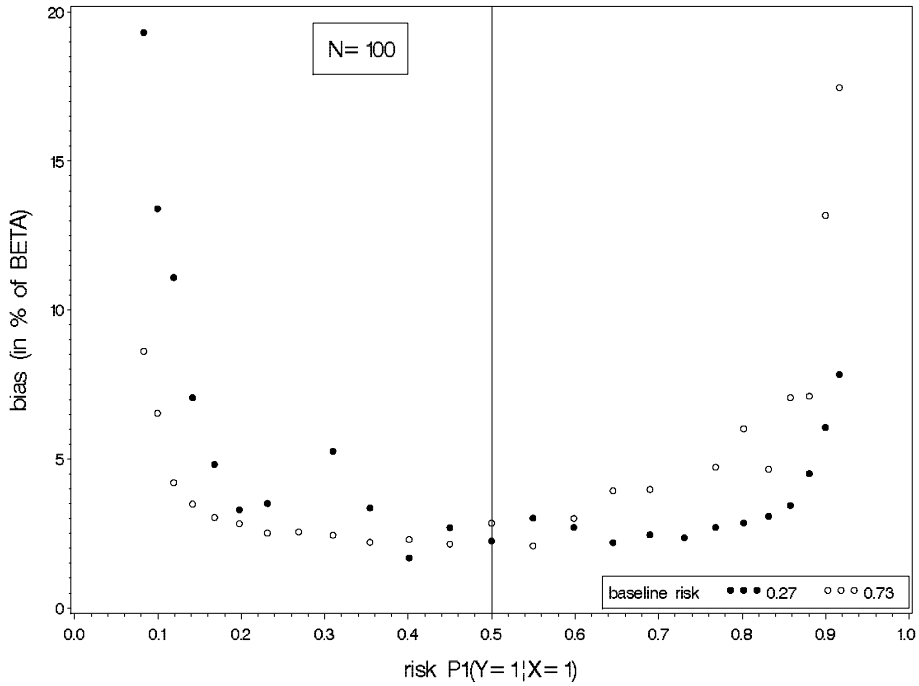


Figure 1: Bias estimates of logistic regression parameter estimates for different true values of the parameter of a binary explanatory variable *x* and different sample sizes plotted versus true risk for *x*=1; graphs shown for two different baseline risks a) as absolute values and b) in percent of the true parameter.

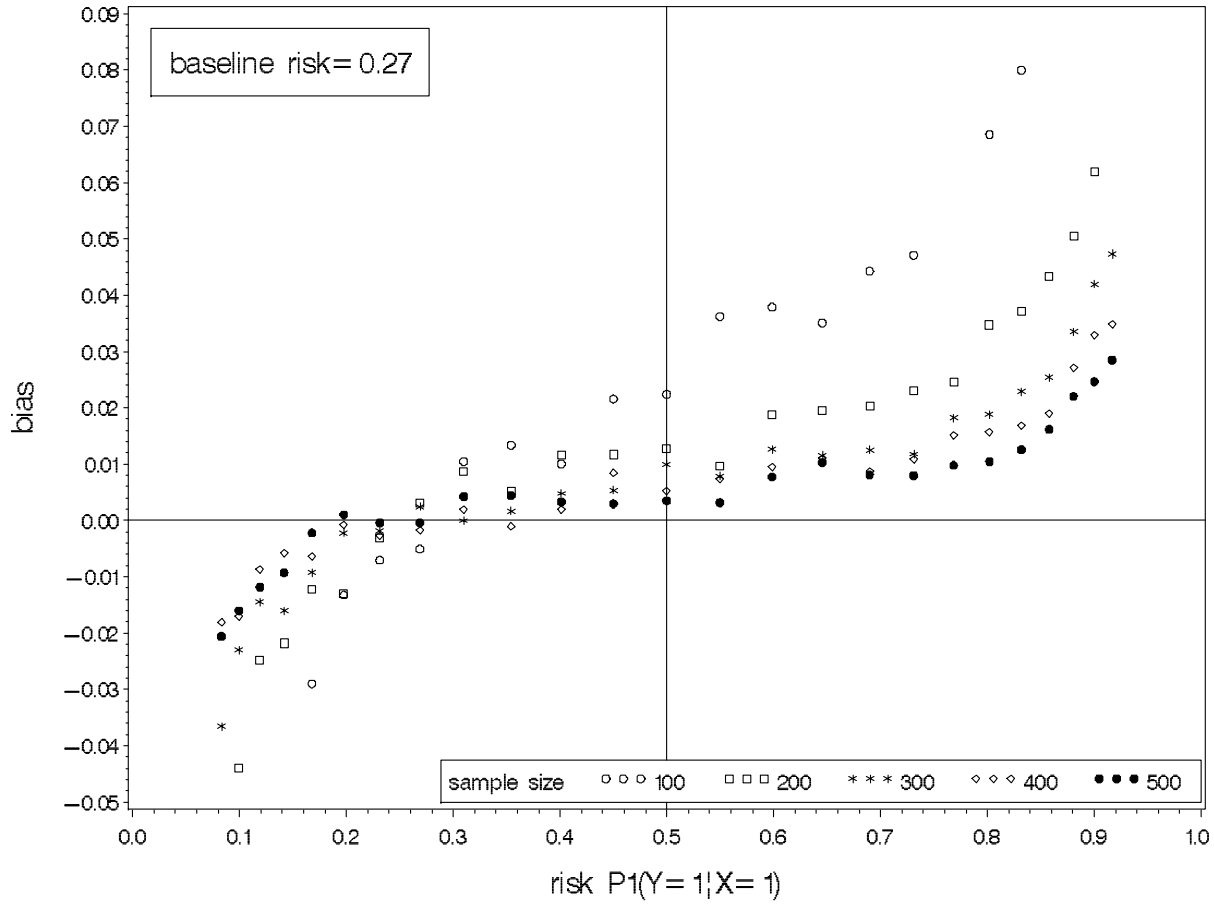
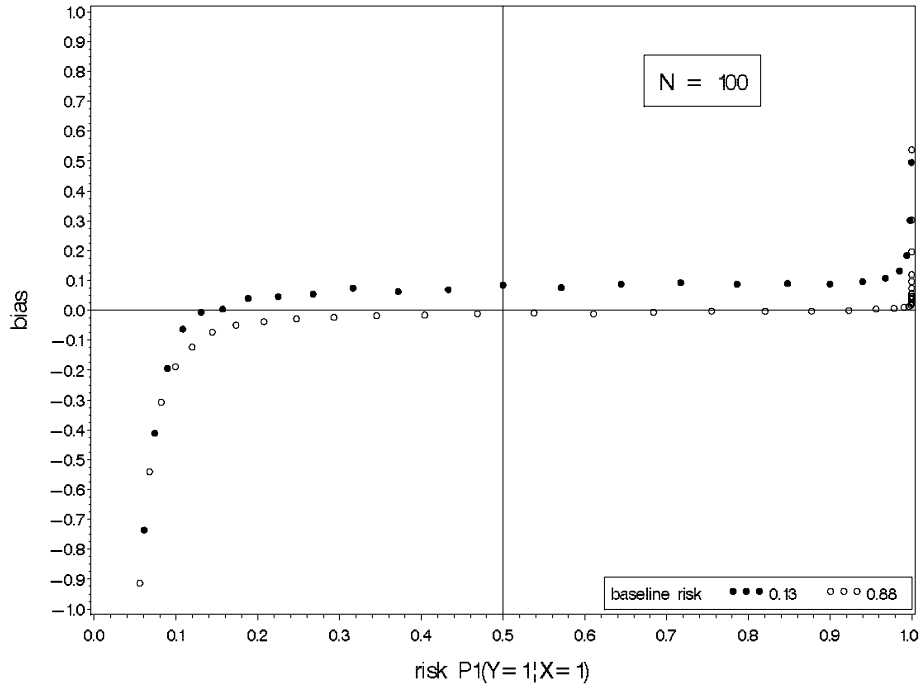


Figure 2: Bias estimates of logistic regression parameter estimates for different true values of the parameter of a binary explanatory variable *x* and sample size *n*=100 plotted versus true risk for *x*=1; graphs shown are all for the same baseline risk.

3a)



3b)

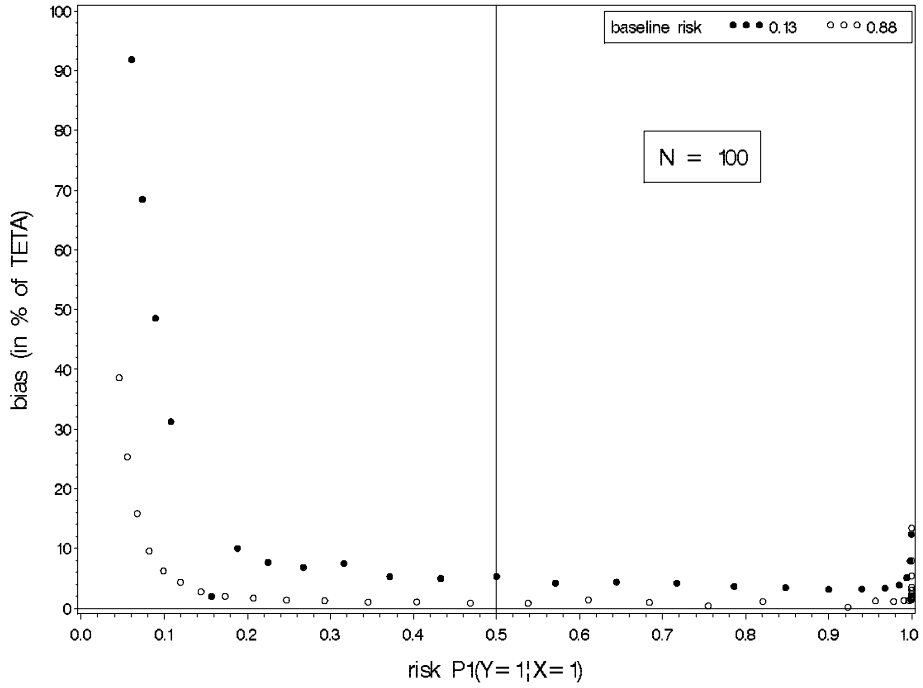


Figure 3: Bias estimates of Cox regression parameter estimates for different true values of the parameter of a binary explanatory variable *x* and sample size *n*=100 plotted versus true risk for *x*=1; graphs shown for two different baseline risks a) as absolute values and b) in percent of the true parameter.

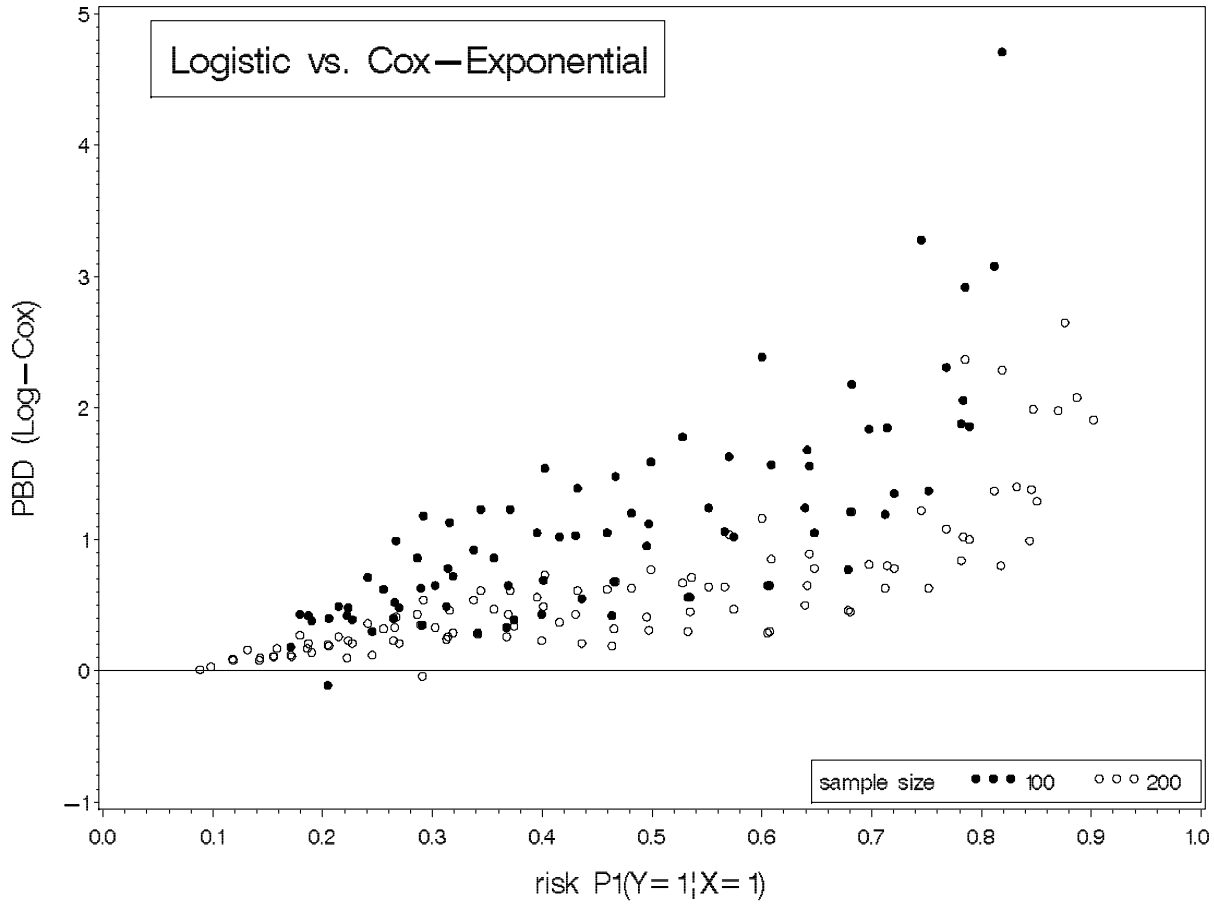


Figure 4: Difference between the percentage biases (see equation 10) in logistic regression and Cox regression of a binary explanatory variable x for identical baseline risks and risks for x=1; estimates are shown for two sample sizes.

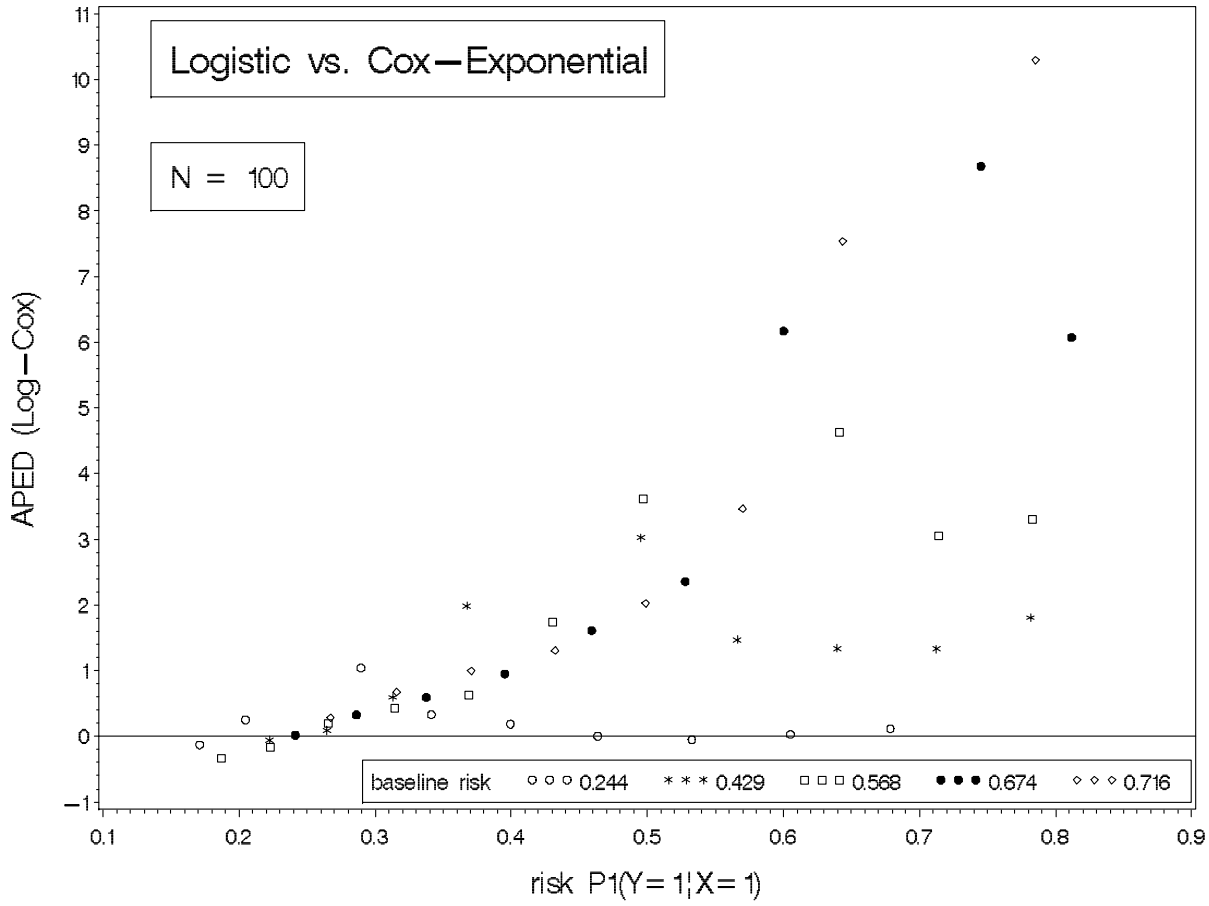


Figure 5: Mean differences of the absolute percentage deviations of true parameters and parameter estimates for single samples (see equation 11) between logistic regression and Cox regression; results are shown for n=100 and different baseline risks and risks for x=1