Universiteit Utrecht

SMOKING IS CONTAGIOUS!

A simple diffusion model of smoking behavior of 12 year olds

Benjamin Rosche | benjamin.rosche@gmail.com | Utrecht University

The research questions:

Is smoking affected by peer influence?

Are boys and girls equally susceptible to peer influence?

Data

- Netherlands 2003/04; 126 classes in 14 schools; varying educational tracks
- 12 year olds; first grade of secondary school

Methods

- *Discrete time survival analysis:*
  $\rightarrow \lambda(t) = \lambda_0(t) * \exp(\alpha_0 + \alpha_1 * X_1 + \beta_2 * X_2(t) + \cdots)$
  $\rightarrow$ baseline hazard rate $\lambda_0(t)$
  $\rightarrow$ time-constant covariate $\alpha_1 * X_1$, e.g. 'female'
  $\rightarrow$ time-dependent covariate $\beta_2 * X_2(t)$, e.g. 'N friends smoking t-1'
- *Dependent variable:* hazard rate / conditional probability of starting to smoke at time point t, given no smoking at time point t-1
  $\rightarrow \lambda(t) = \Pr(t-1 \leq T \leq t \mid T \geq t-1) \,/\, \Delta t$
- *Predictors:* Popularity (indegree), gender-mix of friend network, female, N friends smoking t-1, t-1.

Parameter estimation in a Bayesian framework

- In *"classical"* statistics – using the maximum likelihood approach to derive parameter estimates – we construct a likelihood function of the data and derive parameters that make the occurrence of the data most likely. By contrast, in the *Bayesian framework*, we do not derive one single estimate but a probability distribution for each parameter. The reason for this is that we want to represent prior uncertainty about model parameters with a probability distribution, and update this prior uncertainty with current data to produce a posterior probability distribution that contains less uncertainty.
- As in the Bayesian framework we estimate a probability distribution for each parameter (i.e. posterior distribution) instead of obtaining a single point estimate, we have to summarize our knowledge of the parameters. To obtain summarizing quantities (e.g. mean, variance of a parameter), we need to derive the integral of the respective probability distribution. For many distributions, especially multivariate distributions, integrals may not be easy to compute. However, we can sample from the posterior distribution using modern sampling methods (e.g. MCMC) and thereby obtain the summarizing quantities. One of the most powerful MCMC methods is Metropolis-Hastings (MH) sampling.
- The MH algorithm is an algorithm that generates samples from a probability distribution using the full joint density function. Hence, in contrast to Gibbs sampling, we do not need to know the (full) conditional density. The MH algorithm generates a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired posterior distribution. These sample values are produced iteratively, with the distribution of the next sample being dependent only on the current sample value (which makes it a MCMC method). Specifically, at each iteration, the algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and current value is reused in the next iteration). The probability of acceptance is determined by comparing the function value of the current and candidate sample value, respectively, with respect to the desired distribution.

Posterior predictive distributions

- A flexible approach of examining model fit is the use of posterior predictive distributions (PPD). The PPD for a model is the distribution of future observations that could arise from the model under consideration. If the model has adequately captured the data-generation process, future data simulated from the model should look much like the actually observed (current) data. Deviations imply model misfit. In order to determine whether the simulated and the observed data are similar, formal tests using Bayesian p-values can be conducted.
- To give one example, I examined the extent to which the distribution of replicated values of y match the distribution of the original y. More specifically, I calculated the proportion of correctly predicted pupils starting to smoke (comparing the original data-set to each of the replicated data-sets). On average only 7 percent of the pupils starting to smoke were predicted correctly. This came as no surprise as the model clearly misses important covariates not related to social influence. For instance, socio-economic status of the family can be expected to play an important role with respect to starting to smoke.

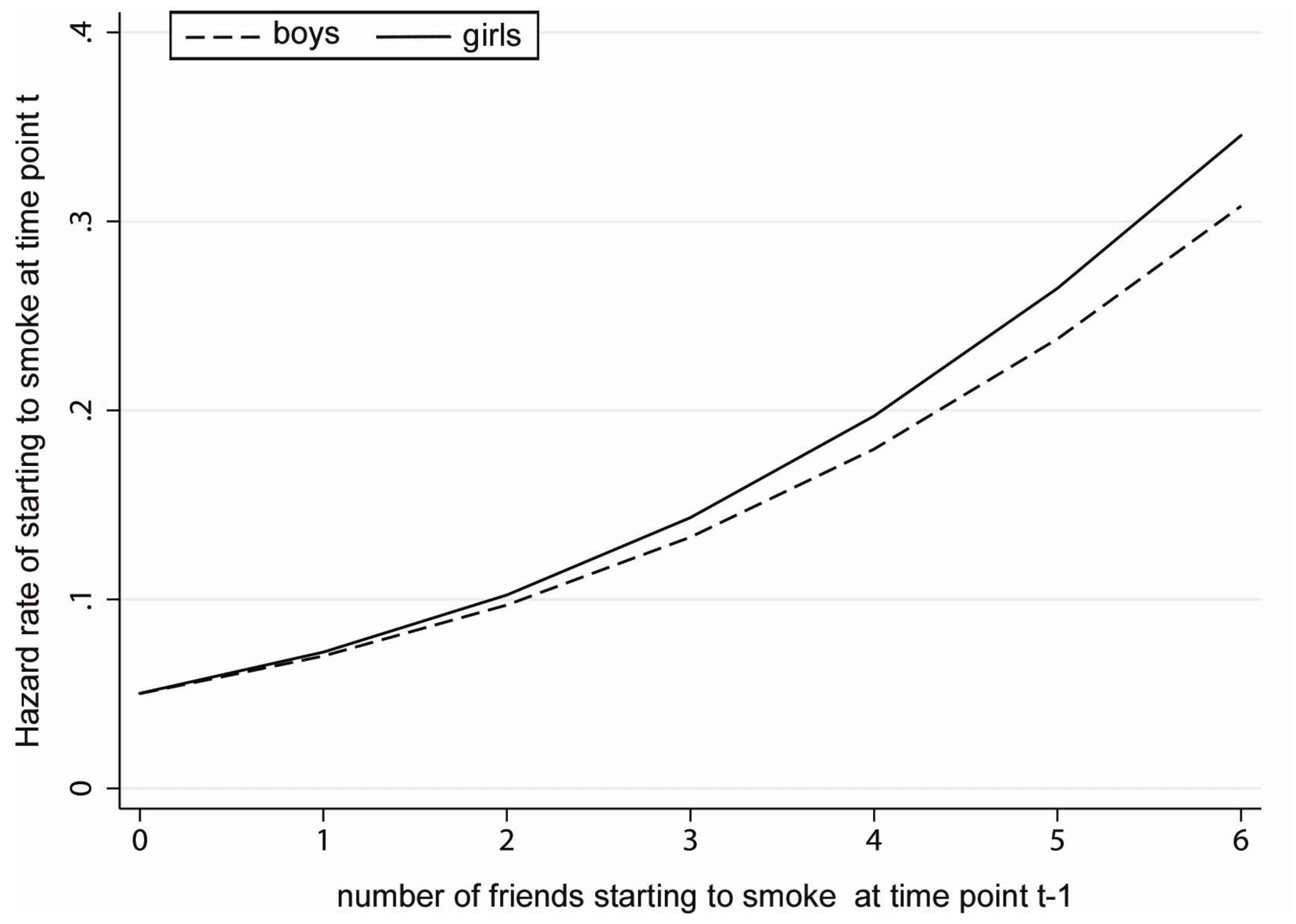
Bayes Factor

- The Bayes factor is another way to investigate model fit. In particular, the Bayes factor can be used to compare two or more (non-nested) models to each other. The Bayes factor compares the posterior probability of each model given the observed data. The posterior odds of one model over the other is $\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} * \frac{p(M_1)}{p(M_2)}$, where the first term on the right hand side of the equation is the ratio of two integrated likelihoods. This is referred to as the Bayes factor. By solving this equation for the Bayes factor, it becomes clear that the Bayes factor is the ratio of the posterior odds to the prior odds.
- To give one example, I calculated the Bayes factor of the hypothesis that friends*gender > 0, against the hypothesis that it equals 0 (H_2). This Bayes factor constitutes a test of whether the gender or the diffusion model best describes the observed data. The prior odds are highly in favor of the gender model as before observing the data far more possible parameter values are in line with it. However, after observing the data, the odds point in the opposite direction: the diffusion model is more than 150 times more likely than the gender model. Put another way, the posterior model probability of the diffusion model is 99.9 percent.

Results

DV: hazard rate of starting to smoke at time point t	(baseline)	(diffusion)	HR	(gender)
Popularity (indegree)	-0.02 [-0.07; 0.02]	-0.03 [-0.08; 0.02]	0.97	-0.03 [-0.08; 0.02]
Gender-mix of friend network	0.55 [0.26; 0.87]	0.49 [0.18; 0.82]	1.63	0.41 [0.04; 0.72]
Female	0.01 [-0.16; 0.21]	0.03 [-0.16; 0.19]	1.03	0.02 [-0.21; 0.26]
N friends smoking t-1		0.37 [0.27; 0.48]	1.45	0.36 [0.22; 0.54]
Female * N friends smoking t-1				0.02 [-0.22; 0.23]
t-1	-0.04 [-0.16; 0.07]	-0.12 [-0.23; -0.00]	0.89	-0.11 [-0.25; 0.02]
Constant	-2.66 [-2.92; -2.37]	-2.72 [-2.97; -2.47]	0.07	-2.74 [-3.04; -2.44]
N	6697	6697		6697
DIC	3133.1	3091.7		3093.6

Credible intervals in square brackets; HR = hazard ratio



ATTENTION! The hypotheses are:

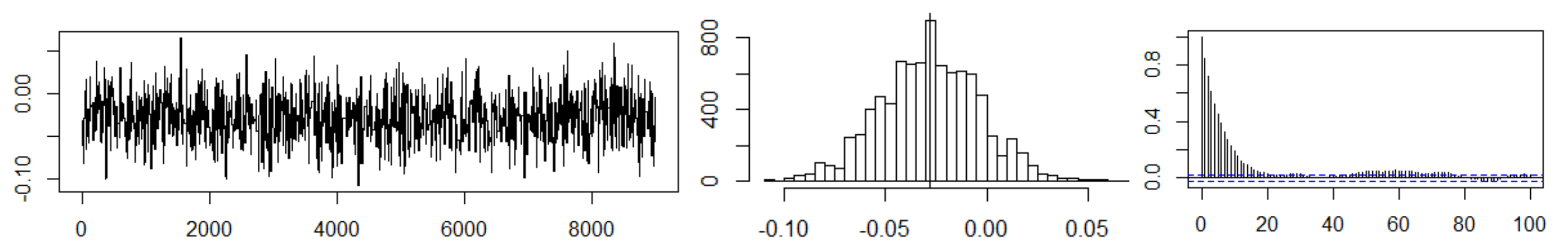
H_1: The more friends start to smoke @ t-1, the more likely a pupil starts as well @ t

H_2: Girls are more susceptible to peer influence in their smoking behavior than boys

- The *diffusion model* shows that starting to smoke is not cool anymore. Popular pupils, who were nominated as a friend by many other pupils, are not more likely to start smoking than unpopular pupils. Furthermore, girls and boys exhibit the same likelihood to start smoking. However, the gender-mix of the friend network matters. Networks with a fifty-fifty balance exhibit 1.63x the likelihood to start smoking than homogeneous groups. More importantly, a peer influence effect is consistent with the data (H_1)! With each additional friend starting to smoke at time point t-1, the likelihood of a pupil starting to smoke increases by a factor of 1.45. The figure above displays this relationship. The probability of an average pupil starting to smoke at time point t given s/he had not started t-1 is 10 percent if two friends started t-1.
- The *gender model* shows that girls are unlikely to be more susceptible to peer influence than boys (H_2). Even though a small positive effect is estimated, the credible interval, the DIC, and the Bayes factor render the simple diffusion model far more credible.

Convergence

Trace plot, histogram, and auto-correlation plot of the parameter 'popularity (indegree)' parameter:



- Although MCMC theory shows that the MH algorithm will, at the limit, converge on the posterior distribution of interest, there is no guarantee that it will converge in any run of finite length. However, the fat caterpillars in the trace plot, the approximately normally distributed parameter distribution in the histogram, and the small autocorrelation of the posterior parameter distribution suggest convergence of the parameter. Similar results are obtained for all other parameters.