A Bayesian hierarchical approach to deal with exposure measurement error in the analysis of the cancer risk associated with exposure to ionizing radiation - An application to the French cohort of uranium miners

Sabine Hoffmann
supervised by Sophie Ancelet (IRSN) and
Chantal Guihenneuc (Université Paris Descartes)

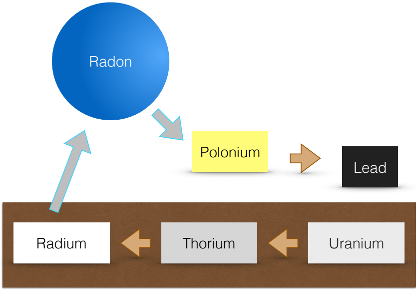
Institut de Radioprotection et de Sûreté Nucléaire

November 27, 2014

Context

- Radon is a noble and radioactive gas which is ubiquitous in soils and rocks
- It is known to be the second leading cause of lung cancer (Samet 2000) responsible for about 2% of cancer deaths in Europe (Darby 2005)

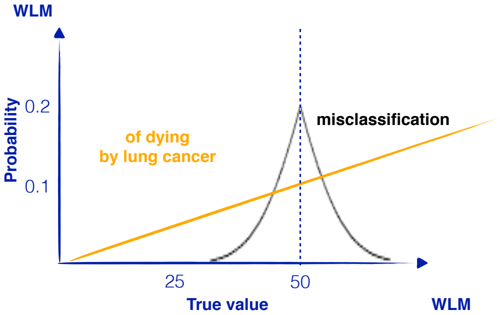
Cohorts of uranium miners present an important source of information on this association



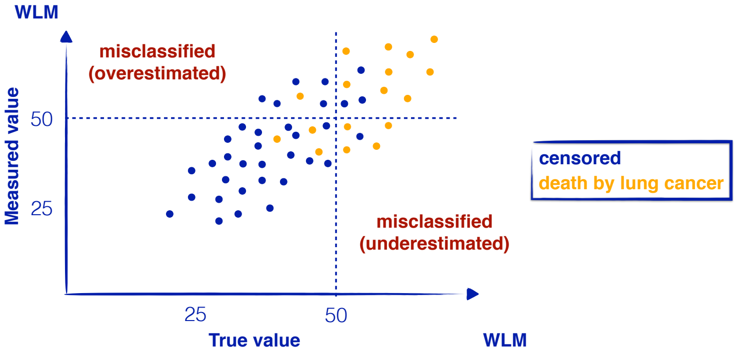
Context

- Studies analysing the lung cancer risk associated with exposure to radon in uranium miners rarely account for exposure measurement error
- Non differential exposure measurement error may cause bias in risk estimates, a distortion in the exposure-risk relationship and a loss in power
- Results obtained on the exposure risk relationship taking into account measurement error are mainly based on simulation studies using Poisson regression

Categorisation in Poisson regression



Categorisation in Poisson regression



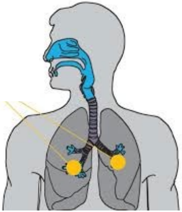
Context

- Regression calibration uses two disjoint steps to estimate true exposure and risk parameters
- The Bayesian hierarchical approach provides a natural way of combining exposure and parameter uncertainty in a coherent framework

Aims

$$\overbrace{\text{Radon}}^{\text{Measurement error}}$$

Obtain a measurement corrected estimate of
the lung cancer risk associated
with cumulative exposure to Radon



The French Uranium Miners' Cohort

1946

5086 uranium miners employed for at least one year at CEA-COGEMA

mean follow-up: 35 years (0.1-60 years)

mean cumulated exposure: 36.61 WLM (0.003 - 960.11)

31 December 2007

2924 miners alive

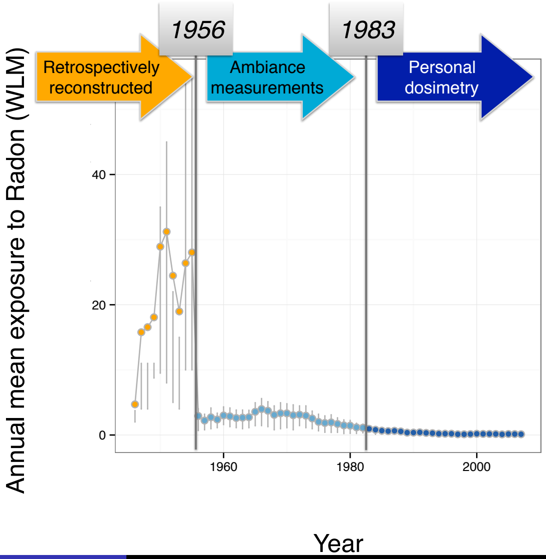
1935 miners dead

211 miners dead by lung cancer (11%)

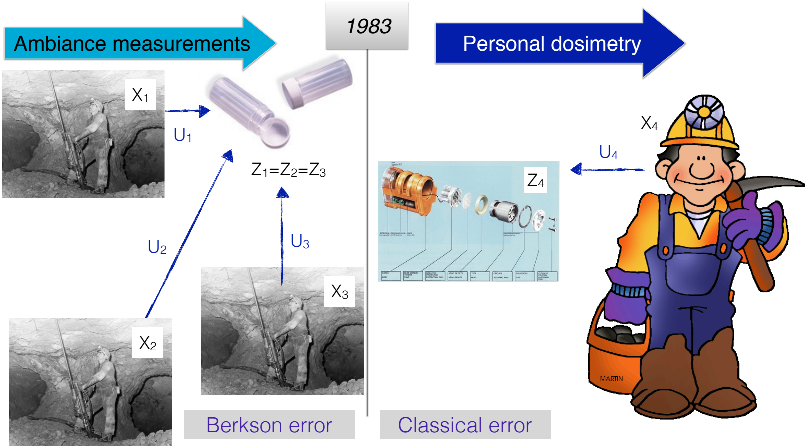
187 miners > 85 years

40 miners lost to follow-up

Radon Exposure in the Cohort



Radon Exposure in the Cohort



Modelling measurement error

Richardson & Gilks (1993) propose to account for measurement error using conditional independence models:

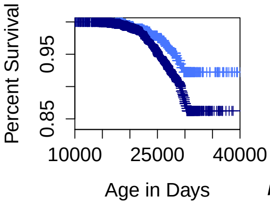
- **disease model:** $[Y_i|\boldsymbol{X_i}, \beta]$
- **measurement model:** $[\boldsymbol{X_i}|\boldsymbol{Z_i}, \sigma]$

where Y_i is the outcome, $\boldsymbol{X_i}$ denotes the vector of true exposure and $\boldsymbol{Z_i}$ the vector of observed exposure for miner i

The disease model

- $Y_i = \min(T_i, C_i)$
- $\delta_i = \mathbb{1}_{[T_i < C_i]}$
- Age at first exposure A_i as potential modifying factor

Two potential hazard structures:

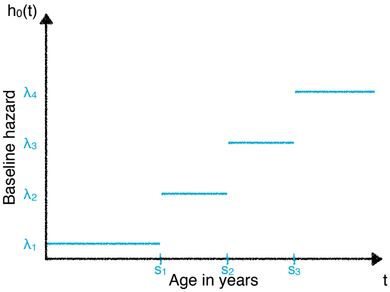


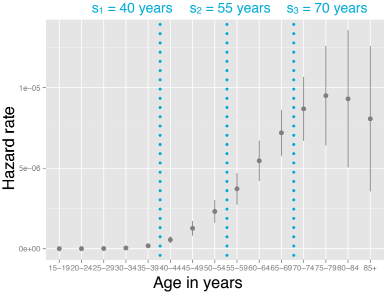
$$h_i(t) = h_0(t)\exp(\beta \sum_{t_j} X_{i,t_j}(t)^{t_j-5} \exp(\gamma A_i(t)))$$

$$h_i(t) = h_0(t)(1 + \beta \sum_{t_j} X_{i,t_j}(t)^{t_j-5} \exp(\gamma A_i(t)))$$

Modelling baseline hazard

- No assumptions about the form of the baseline hazard
- The piecewise constant hazard model is a convenient nonparametric model

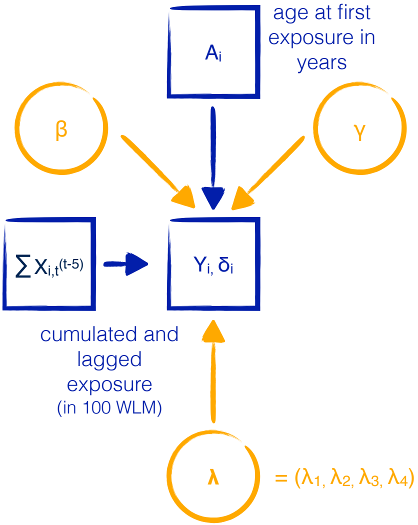




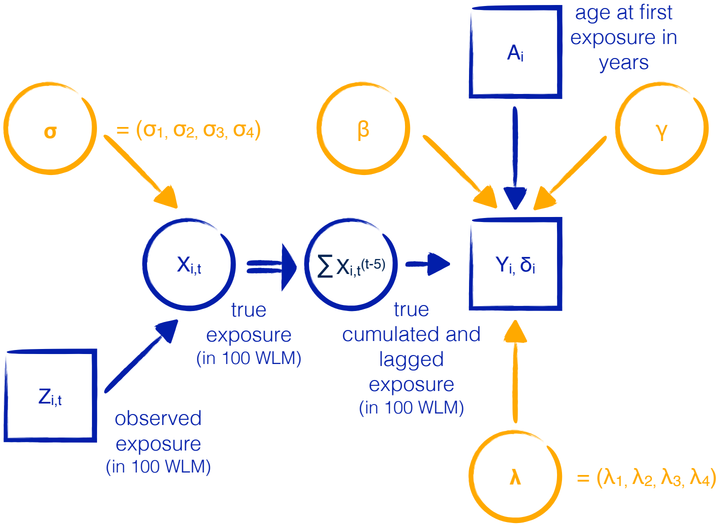
The measurement model

- Neglect classical measurement error 1983 - 2007
- Berkson model during the period 1946 -1982:
  $X = Z \cdot U$ where $\mathbb{E}(U|Z) = 1$ and U log-normal distributed (Heid 2002)
- Categorisation of measurement error in the cohort by Allodji et al. (2012):
  Define σ_1^2, σ_2^2, σ_3^2, σ_4^2 as error variances for the periods 1946-1955, 1956-1974, 1975-1977 and 1978-1982
- $U \sim \mathcal{LN}\left(-\frac{\sigma^2_{p_{iq}}}{2}, \sigma^2_{p_{iq}}\right)$

Integrating measurement error



Integrating measurement error



Bayesian inference

A Bayesian approach to model measurement error in a cohort of uranium miners

Sabine Hoffmann

Introduction

Material and Methods

Results

Discussion and conclusion

$$\pi(\boldsymbol{\theta}|y, Z, A) \propto \pi(\beta)\pi(\gamma)\pi(\boldsymbol{\sigma})\pi(\boldsymbol{\lambda})\prod_{i=1}^{n} L\left(y_i | \sum_{q=1}^{Q_i} X_{iq}, A_i, \beta, \gamma, \lambda\right)$$

$$\cdot \prod_{i=1}^{n}\prod_{q=1}^{Q_i} f\left(X_{iq}|Z_{iq}, \boldsymbol{\sigma}\right)$$

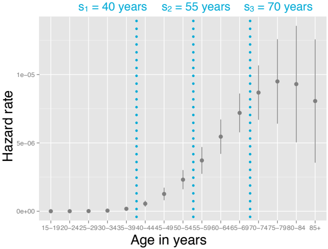
Bayesian inference

$$\pi(\boldsymbol{\theta}|y, Z, A) \propto \pi(\beta)\pi(\gamma)\pi(\boldsymbol{\sigma})\pi(\boldsymbol{\lambda})\prod_{i=1}^{n} L\left(y_i|\sum_{q=1}^{Q_i} X_{iq}, A_i, \beta, \gamma, \lambda\right)$$
$$\cdot \prod_{i=1}^{n}\prod_{q=1}^{Q_i} f\left(X_{iq}|Z_{iq}, \boldsymbol{\sigma}\right)$$

- Markov Chain Monte Carlo method to sample from the joint posterior distribution
- Metropolis-Hastings algorithm implemented in Python
- Model comparison based on Deviance Information Criterion (DIC)

Prior distributions

- $\pi(\beta)$: $\beta \sim \mathcal{N}(0, 10^4)$
  truncated at zero (to guarantee $h_i > 0$)
- $\pi(\gamma)$: $\gamma \sim \mathcal{N}(0, 10^4)$
- $\pi(\boldsymbol{\lambda})$: $\lambda_j \sim \mathcal{G}(\alpha_{0j}, \lambda_{0j})$ for each component j, $j = 1, \dots, 4$



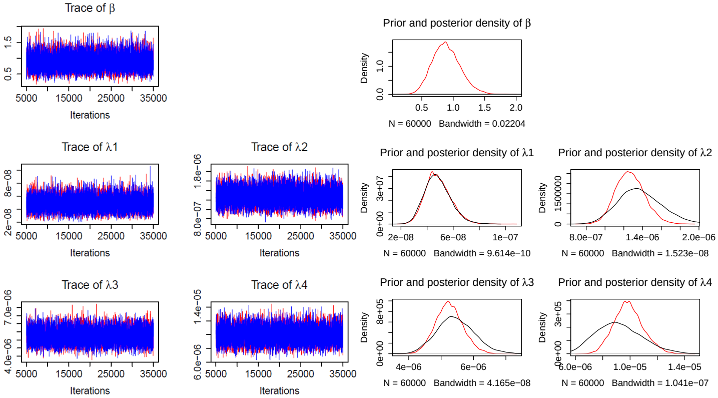
- $\pi(\boldsymbol{\sigma})$: No validation sample available, use point estimates based on Allodji et al. (2012)
  $\Rightarrow \sigma_1^2 = 0.94$, $\sigma_2^2 = 0.47$, $\sigma_3^2 = 0.42$, $\sigma_4^2 = 0.33$

Comparing the different models

Model	DIC
Intercept only	5458.41
Excess relative risk model $X_i^{cum}(t)$	5433.39
Excess relative risk model $X_i^{cum}(t) \cdot \exp(A_i(t))$	5436.02
Cox-like model $X_i^{cum}(t)$	5443.64
Cox-like model $X_i^{cum}(t) \cdot \exp(A_i(t))$	5445.03

Model	β	γ	λ_1 in 10^{-8}	λ_2 in 10^{-6}	λ_3 in 10^{-6}	λ_4 in 10^{-6}
ERR $X_i^{cum}(t)$	0.89 (0.23)	-	4.85 (0.97)	1.27 (0.15)	5.25 (0.42)	9.90 (1.07)
	[0.49;1.37]	-	[3.15 ;6.94]	[0.99;1.58]	[4.46; 6.12]	[7.92;12.14]
ERR $X_i^{cum}(t)\cdot$	0.85 (0.23)	0.14 (0.32)	4.87 (0.98)	1.28 (0.15)	5.28 (0.43)	9.94 (1.07)
$\exp(A_i(t))$	[0.45; 1.34]	[-0.53; 0.71]	[3.13 ;6.91]	[0.99 ;1.60]	[4.49;6.15]	[7.92;12.17]

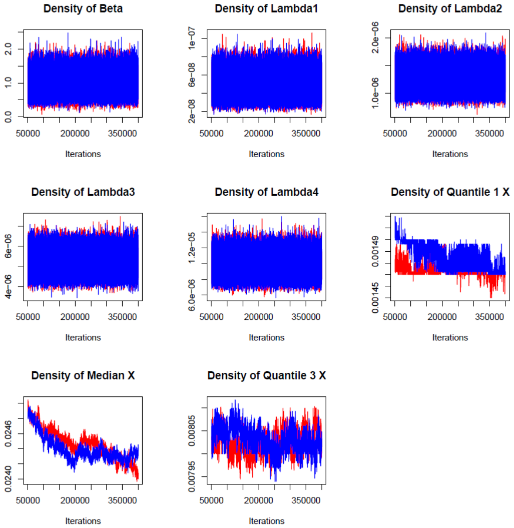
Prior and posterior distributions



Integrating measurement error

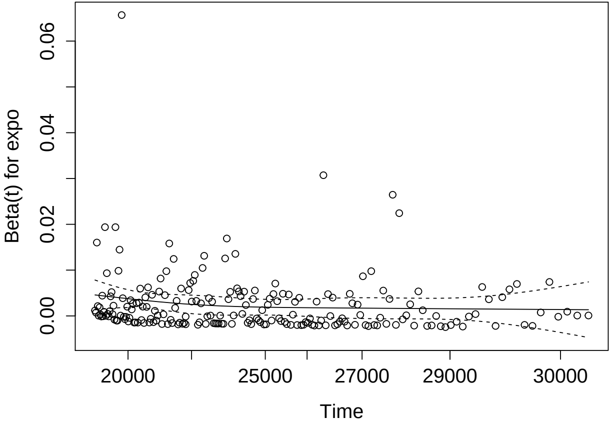
Model	β	λ_1 in 10^{-8}	λ_2 in 10^{-6}	λ_3 in 10^{-6}	λ_4 in 10^{-6}
without measurement error	0.89 (0.23)	4.85 (0.97)	1.27 (0.15)	5.25 (0.42)	9.90 (1.07)
	[0.49;1.37]	[3.15 ;6.94]	[0.99;1.58]	[4.46; 6.12]	[7.92;12.14]
with Berkson measurement error	0.98 (0.25)	4.82 (0.93)	1.26 (0.15)	5.23 (0.41)	9.86 (1.09)
	[0.54; 1.50]	[3.13 ;6.90]	[0.98 ;1.58]	[4.44;6.09]	[7.88;12.08]

Checking the convergence of the full model



Checking the proportional hazards assumption

Harrell test: $p = 0.16 > 0.05$



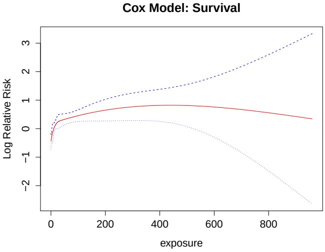
Checking the (log-) linearity of the model

Piecewise linear model for β:

Model	DIC
Intercept only	5458.41
ERR $X_i^{\text{cum}}(t)$	5433.39
ERR piecewise linear	5426.55

$\hat{\beta}_1$ (<100 WLM):
1.49 [0.84;2.23]

$\hat{\beta}_2$ ($\geq$100 WLM):
0.29 [0.01;0.87]



Discussion

- Bayesian approach flexible, possible to account for exterior information
- Taking Berkson type measurement error into account does not seem to substantially change the risk associated with cumulative exposure to radon but the convergence for the full model does not appear to be reached for the latent variables
- Assumption of linearity is violated: model misspecification?
- It was not possible to adjust for tobacco consumption, diesel exhaust, arsenic, asbestos and silica quartz
- Mixture of Berkson and classical error before 1983?

Perspectives

- More extensive modelling of the exposure-risk relationship
- Integrate classical error into the measurement model
- Simulation studies to compare the performance of this Bayesian approach to take measurement error into account with classical frequentist methods

Thank you for your attention

Smoking as effect modifying factor

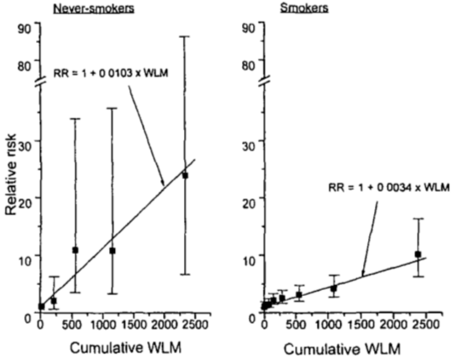


Figure 3. Relative risk (RR) of lung cancer with cumulative radon exposure among lifelong non-smokers and others in the six cohort studies of underground miners for which smoking information was available (based on [18]). Although the increase in relative risk per unit exposure is higher for never smokers than for smokers, the increase in absolute risk will be higher for smokers, as they have much higher rates of lung cancer.

Survival for piecewise constant baseline hazard

$$S_i(y_i; X_i^{\text{cum}}) = \exp\left(-\varphi(X_i^{\text{cum}}, A_i, \beta, \gamma) \sum_{j=1}^{J} \delta_{ij} \left(\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1})\right)\right)$$

The likelihood of the disease model

$$\prod_{i=1}^{n} L\left(y_i \Big| \sum_{q=1}^{Q_i} X_{iq}, A_i, \beta, \gamma, \lambda\right) = \prod_{i=1}^{n} [h_0(y_i)\varphi(X_i^{\text{cum}}(y_i), A_i, \beta, \gamma)]^{\nu_i}$$

$$\cdot \prod_{m=1}^{M_i} \frac{S_m(r_i^m; \omega_{i,m})}{S_m(r_i^{m-1}; \omega_{i,m})}$$

$$= \prod_{i=1}^{n} \prod_{m=1}^{M_i} [h_0(r_i^m)\varphi(X_i^{\text{cum}}(r_i^m), A_i, \beta$$

$$\cdot \frac{S_m(r_i^m; \omega_{i,m})}{S_m(r_i^{m-1}; \omega_{i,m})}$$

Effect of calendar period on baseline hazard

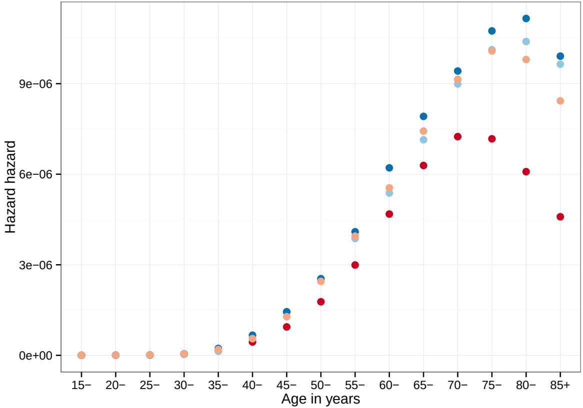
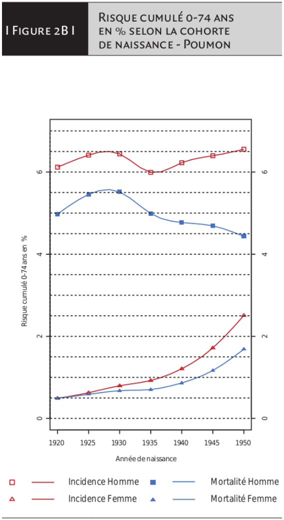


Figure: Hazard of lung cancer mortality in French males for the following periods: 1968-1977 (red), 1978-1987 (orange), 1988-1997 (blue), 1998-2005 (lightblue)

Effect of calendar period on lung cancer



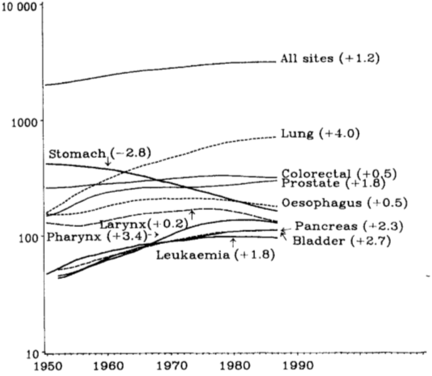


Fig 2—Trends in cancer mortality by site, French males, 1950–87. Age-standardised rates per million population, logarithmic scale. Average annual variation (%) by site shown in parentheses.

Characterisation of the error periods in the French cohort of uranium miners

Sources	Periods			
	1956–74	1975–77	1978–82	1983–99
Natural variations of air-borne radon gas concentration	30.0	21.2	21.2	0.0
Precision of the measurement device	20.0	20.0	20.0	10.0
Approximation of equilibrium factor	29.4	29.4	11.8	0.0
Operator in charge of air samples	2.0	2.0	2.0	0.0
Estimation of working time	4.0	4.0	8.0	0.0
Record-keeping and data transcription	1.5	1.5	1.5	1.0
Combined relative standard uncertainty[a]	46.8	41.7	32.6	10.1

[a] Estimated using the root sum square (RSS) method.

Radon Exposure and its Assessment in the Cohort

Period	1946-1955	1956-1974	1975-1977	1978-1982	1983-1999
Ventilation	poor	medium	good		good
Exposure Assessment	reconstructed in 1981	ambient measurement by scintillation flasks and and F $\times$ time			personal dosimetric system
Number of measurements	only a few in 1954	$\geq$ one per week (>20.000/year)	several per week (>40.000/year)		continuous: (monthly values)
Mean yearly exposure	21.3	3.0	1.9	1.4	0.4

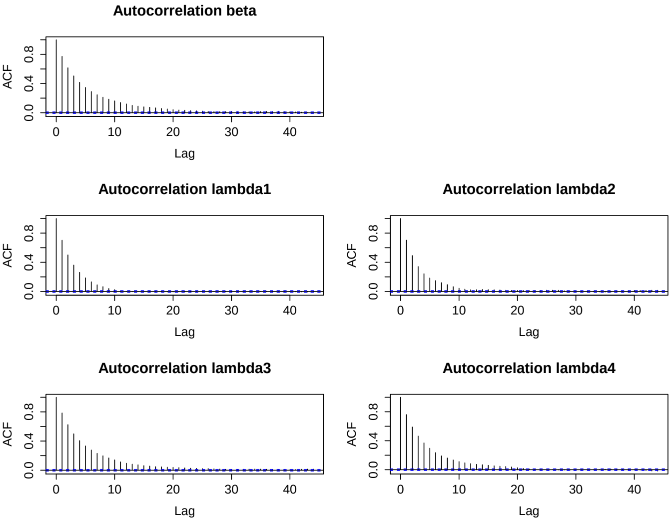
Radon Exposure and its Assessment in the Cohort

Period	1946-1955	1956-1974	1975-1977	1978-1982	1983-1999
Ventilation	poor	medium	good		good
Exposure Assessment	reconstructed in 1981	ambient measurement by scintillation flasks and and F $\times$ time			personal dosimetric system
Number of measurements	only a few in 1954	$\geq$ one per week (>20.000/year)	several per week (>40.000/year)		continuous: (monthly values)
Mean yearly exposure	21.3	3.0	1.9	1.4	0.4

$\Rightarrow$ Berkson error

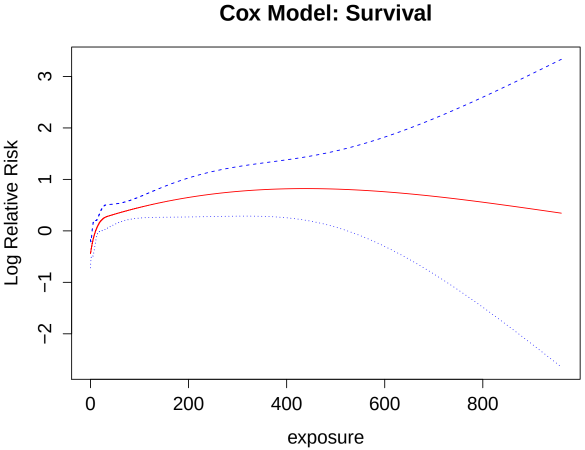
$\Rightarrow$ Classical error

Autocorrelations



Checking log linearity

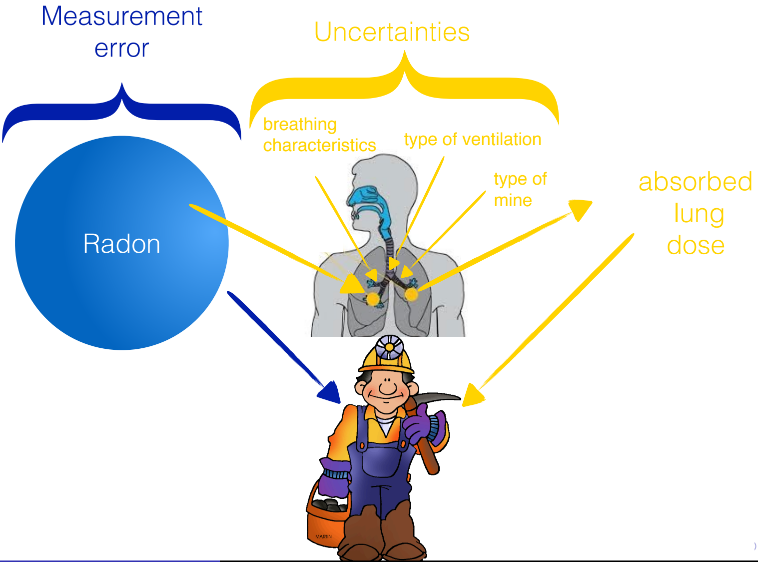
Test linear model against linear-quadratic model:
$p = 0.006 < 0.05$



Principal sources of measurement error in the Cohort

- Natural variations in air-borne radon gas concentration
- Precision of the measurement device
- Approximation of the equilibrium factor
- Human error by the operator in charge of air samples
- Estimation of working time
- Record-keeping and data transcription

Long-term perspectives



The principle of the Metropolis-Hastings algorithm

- Choose an arbitrary starting value $\boldsymbol{\theta}^{(0)}$
- At iteration t and for each component $\theta_l, l = 1, \ldots, p$ of θ
  1. Given $\theta_l^{(t-1)}$ simulate $\theta_l^{\text{cand}} \sim \mathcal{N}(\theta_l^{(t-1)}, \sigma_l^2)$
  2. Compute $\rho(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{\text{cand}}) = \min\left\{\frac{\pi(\boldsymbol{\theta}^{\text{cand}}|y,Z,A)}{\pi(\boldsymbol{\theta}^{(t-1)}|y,Z,A)}, 1\right\}$.
  3. Accept θ_l^{cand} and set $\theta_l^{(t)} = \theta_l^{\text{cand}}$ with probability $\rho(\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\theta}^{\text{cand}})$, otherwise reject θ_l^{cand} and set $\theta_l^{(t)} = \theta_l^{(t-1)}$.

Vector X of true exposures updated by block