Quantile Regression in Survival Analysis

Andrea Bellavia

Unit of Biostatistics, Institute of Environmental Medicine
Karolinska Institutet, Stockholm
http://www.imm.ki.se/biostatistics
andrea.bellavia@ki.se

March 18th, 2015

Outline

1. Quantile regression
2. Survival analysis
3. Quantile regression in survival analysis
4. Recent developments
5. Advantages of evaluating survival percentiles in medical research (this afternoon)

1. Quantile Regression

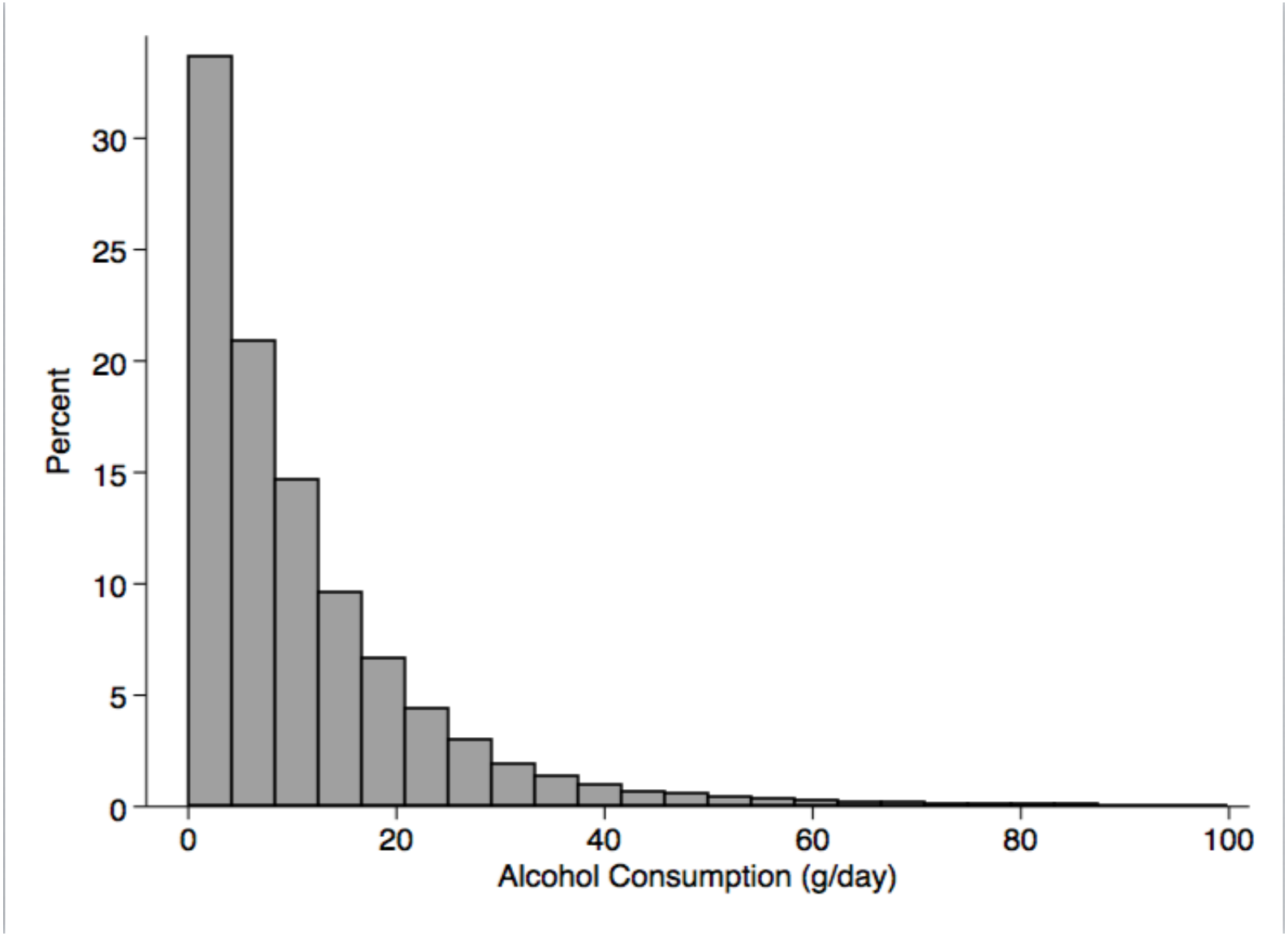
Quantile regression - why?

- To summarize a continuous variable we commonly use Mean and Standard Deviation (SD)
- Example: alcohol consumption, in grams/day, in a study population of ∼70.000 participants

Mean and SD of daily alcohol consumption

Mean: 11 grams/day
Standard Deviation: 12 grams/day

- The histogram depicts the entire distribution


Percent
30
25
20
15
10
5
0
0
20
40
60
80
100
Alcohol Consumption (g/day)

- The distribution of Alcohol consumption is skewed
- In this situation Mean and SD don't provide a complete summary
- Percentiles can complement the information on the entire distribution
- 25% consume less than 3 g/day
- 50% consume less than 7 g/day
- 75% consume less than 15 g/day

a) Comparing distributions

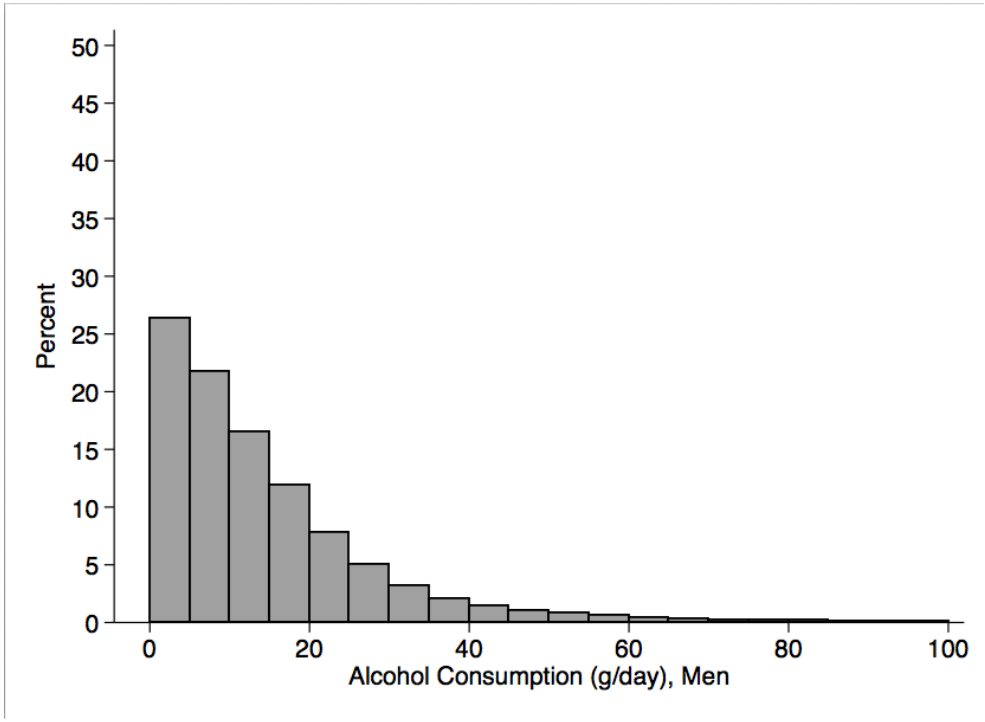
- We now want to evaluate gender differences in alcohol consumption

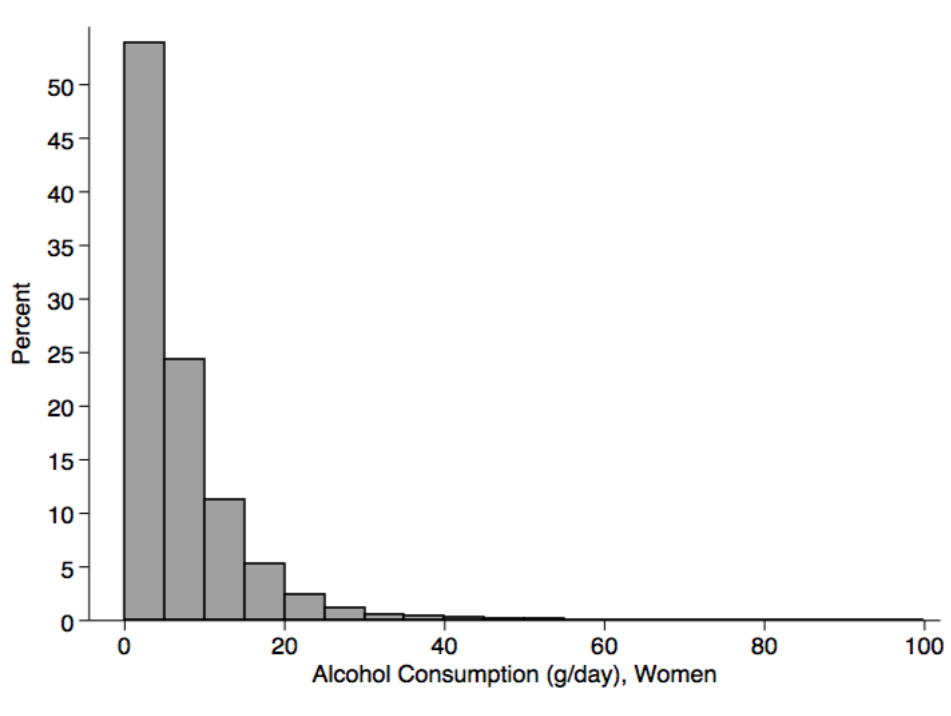
Average alcohol consumption among men and women

Men: 14.3 grams/day
Women: 6.7 grams/day

- On average, men drink double than women

What do we miss?



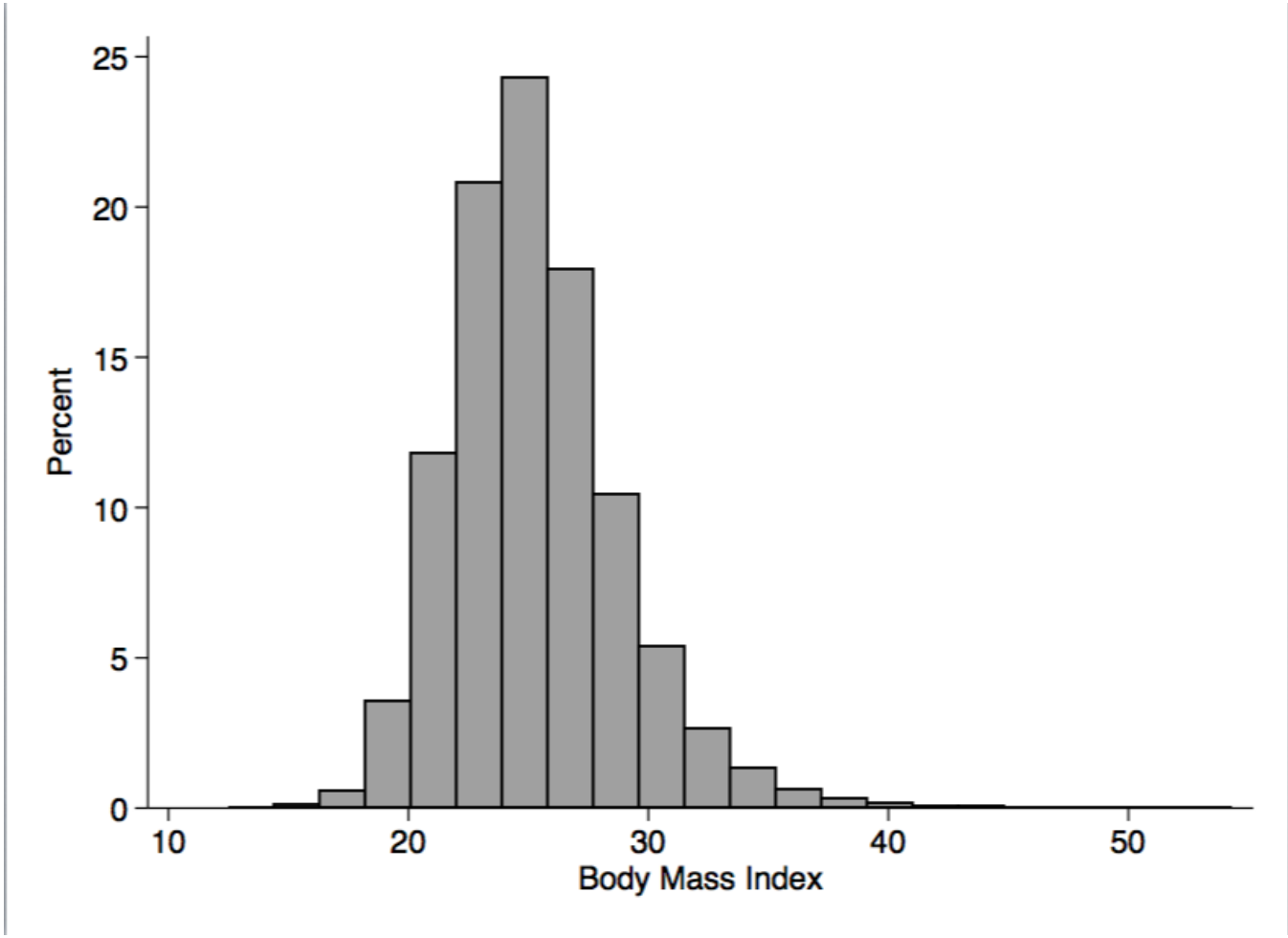


- We miss changes in the shape of the distribution
- If the mean consumption among women is 7 point lower doesn't necessarily mean that the entire distribution is shifted by 7 points on the left

- All common statistical methods for the comparison of two groups are based on a mean comparison (t-test, ANOVA, linear regression)
- Percentile-based approaches allow comparing the entire distribution of alcohol consumption between men and women
- Quantile regression (Koenker, 1978) is the most common approach

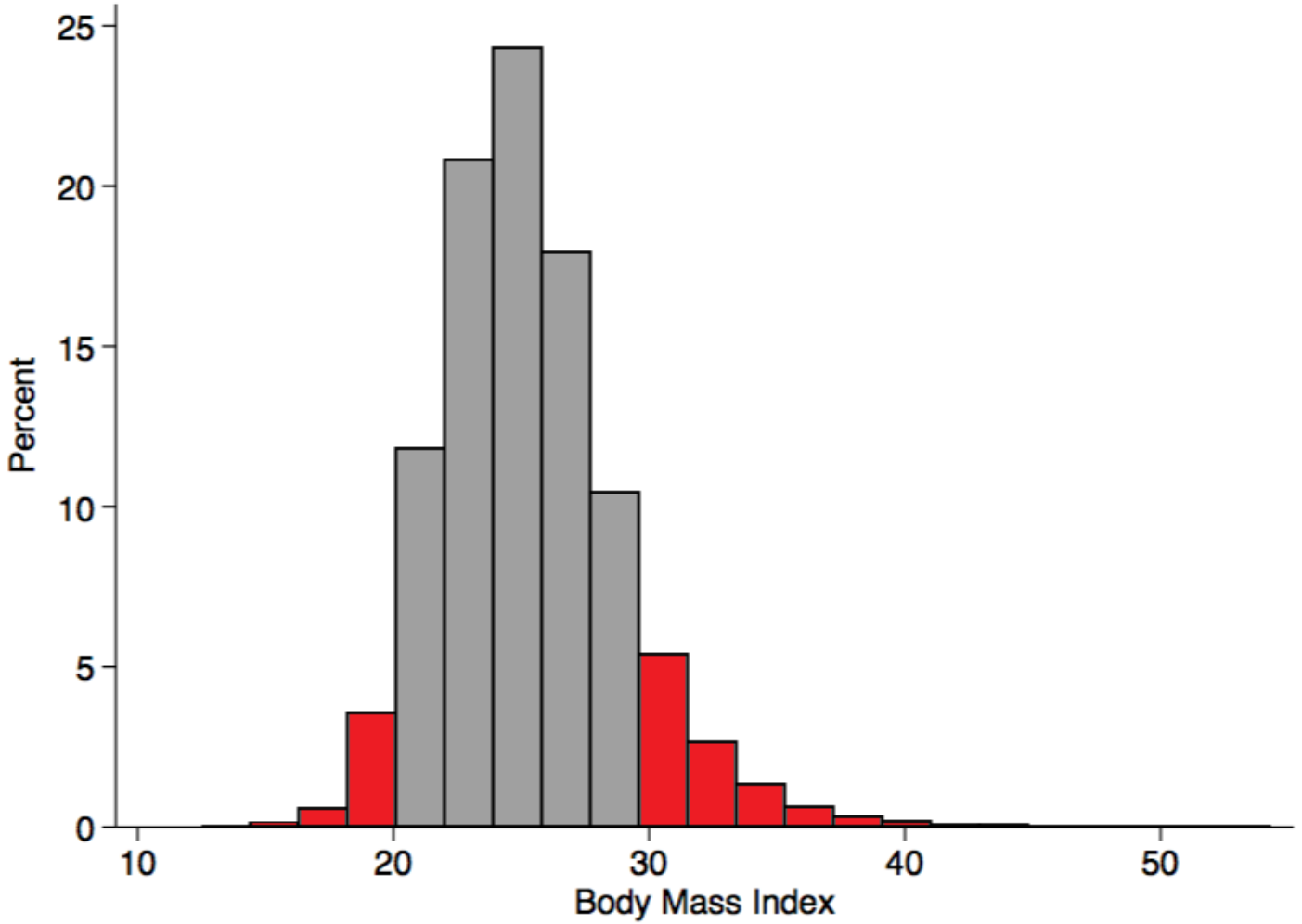
b) Focusing on specific percentiles

- The histogram shows the distribution of body mass index (BMI) in the same population



- Mean = 25.3 Kg/m^2; SD=3.1

- From a public health perspective we are not really interested in evaluating the mean BMI (they are usually healthy). We are more interested in underweight ($<$19.5 Kg/m^2) and obese ($>$30 Kg/m^2) participants



- Linear regression would model changes in the mean BMI according to a set of covariates
- Quantile regression allows evaluating changes in specific percentiles of BMI, such as the 10th (20 Kg/m^2 in the example), or the 90th (30 Kg/m^2), according to a set of covariates

Quantile regression

- Allows focusing on specific percentiles of interest
- The entire shape of the distribution is taken into account
- Quantile regression dates back to Boscovich, 1757
- Mathematical and computational difficulties have slowed its development
- Estimation of conditional quantiles of a distribution was developed by Wagner, 1959. A detailed presentation of the topic is in the book by Koenker, 2005

Some maths

- Given a quantile τ, a response variable Y, and a set of covariates x, a linear model for the conditional τth quantile is:

Linear Quantile Regression

$$Q_{y_i}(\tau|x_i) = x_i^T \beta(\tau)$$

- The p^{th} quantile regression model establishes a linear relationship between x and the p^{th} quantile of Y
- Different applications have been developed, especially in economics

Some maths

- Estimation is conducted by minimizing the Eucledian distance $||y - \hat{y}||$ over all $\hat{y}$
- We can write the quantile-regression distance function as

$$d_\tau(y, \hat{y}) = \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}_i)$$

- This function is differentiable except at the points in which residulas are equal to 0

Properties

- Equivariance to monotonic transformation: let h be a non-decreasing function, then for any Y

$$Q_{h(Y)}(\tau) = h(Q_Y(\tau))$$

- Robustness: quantile regression estimates are not sensitive to outliers
- Standard errors and confidence intervals: the bootstrap procedure is usually preferred to asymptotic procedures

Softwares

- Stata: `qreg`

```
qreg y x z, quantile(.25 .75) reps(100)
```

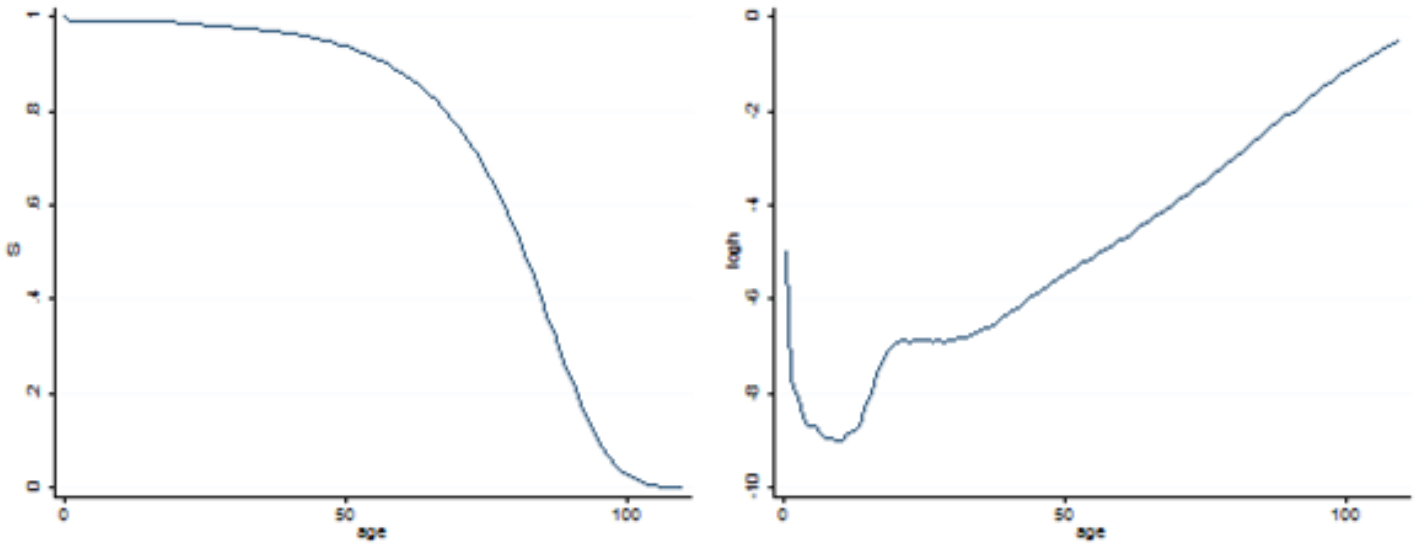
- SAS: Proc QUANTREG
- R: package 'quantreg'

2. Survival Analysis

Review

- In survival analysis we have two quantitites of interest: the event D (usually 0/1), and the time to the event T (a continuous variable)
- T can be defined in different ways (e.g. follow-up time, age)
- The main differences between T and a common continuous variable Y are censoring and skewness

Survival and Hazard curves



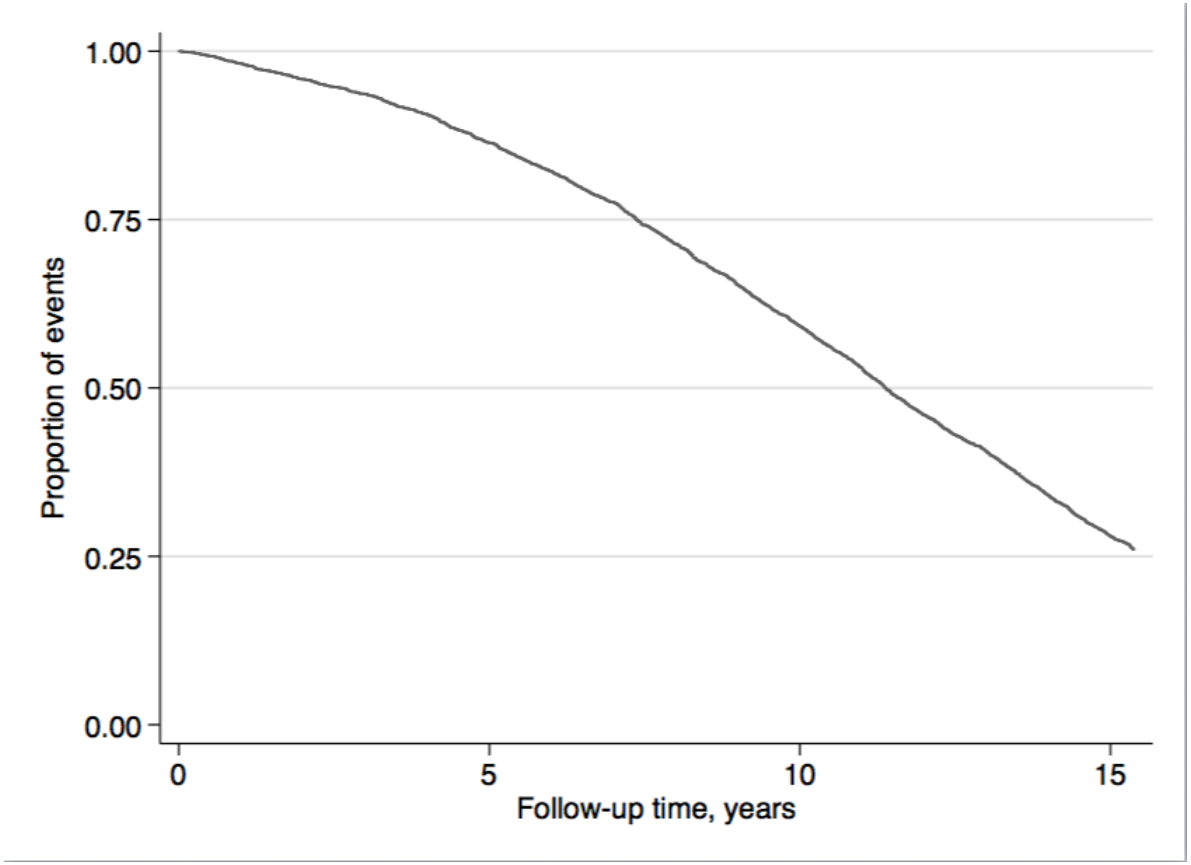
- The hazard can be seen as an instantaneous rate of the event D. Little emphasis is posed on T
- The survival curve combines information on the risk of the event D and the time T
- If we wish to make inference on time (e.g. certain events such as overall mortality) we have to focus on the survival curve

3. Quantile Regression in Survival Analysis

Motivation (1)

- Statistical modeling of the survival curve is not strigthforward
- This is one reason for the extreme popularity of hazard-based methods in survival analysis (e.g. COX)
- Quantile regression can be used in survival analysis to evaluate the (percentiles of the) survival curves

Motivation (2)

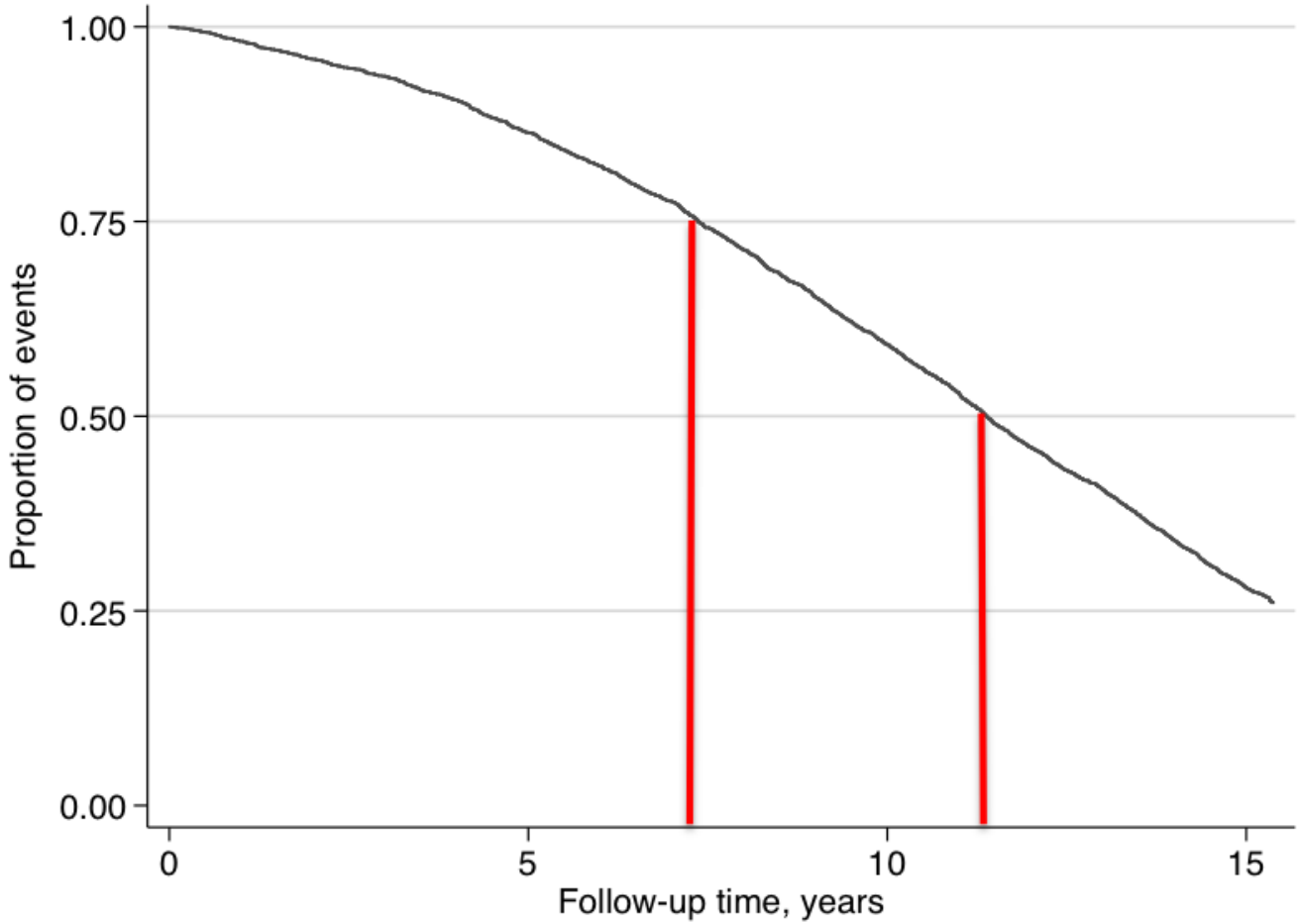


- Because of censoring we often do not observe all the curve
- Mean survival time can not be calculated (that would be the integral under the curve, which is ∞)
- Median survival is a valid alternative

Survival Percentiles

- The percentiles of a time variable T are referred to as survival percentiles
- Example - The minimal value of T is 0, when everyone is alive. The time by which 50% of the participants have died is called 50th survival percentile, or median survival
- In the same way we can define all survival percentiles

Survival Curve



- The 25th survival percentile and the 50th survival percentile (median survival) are shown in the figure

Basic functions

- Let's denote T the time-to-event random variable.

Cumulative distribution function

$$F(t) = Pr(T \leq t) = p$$

Survival function

$$S(t) = 1 - F(t) = 1 - p$$

Quantile function

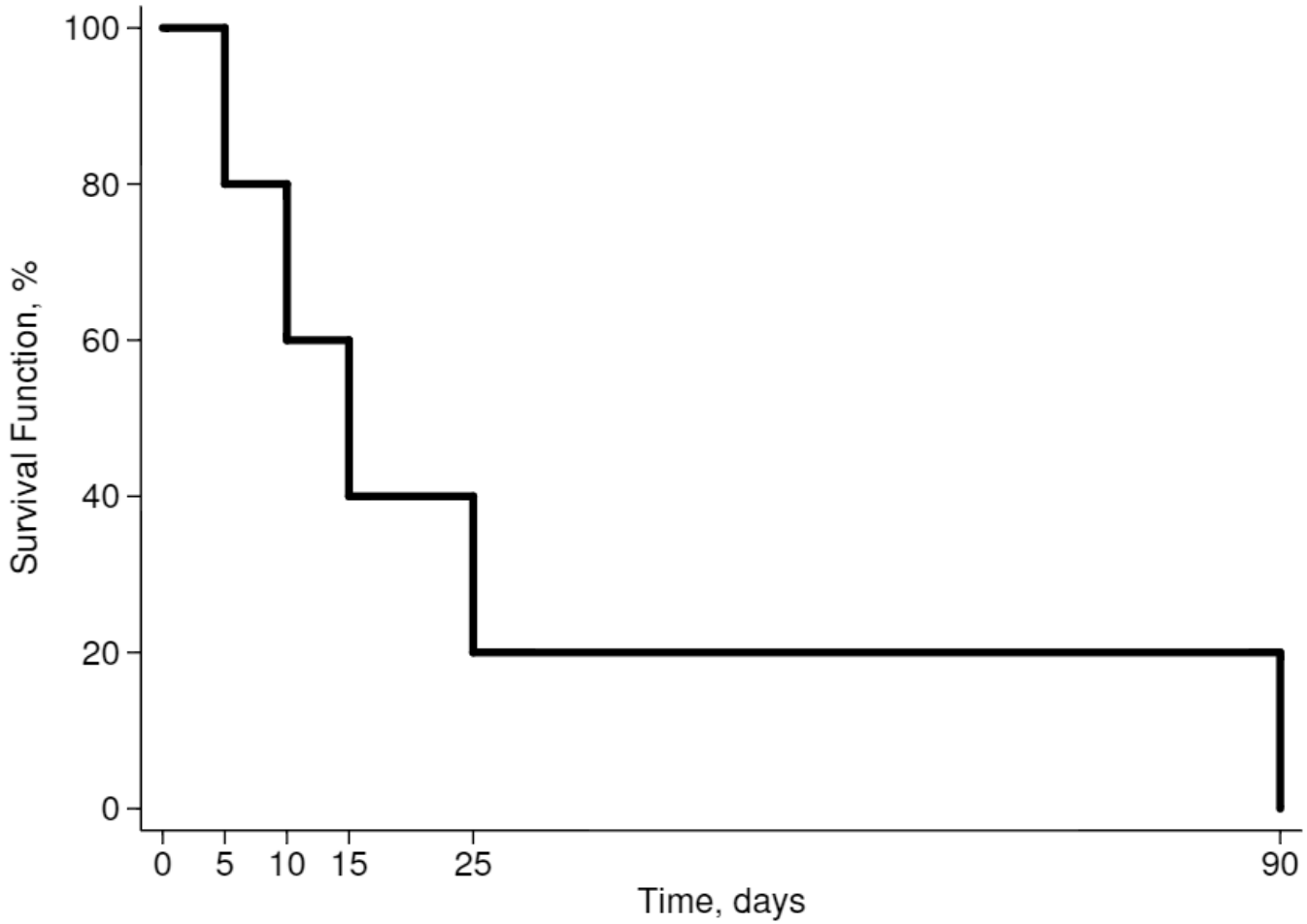
$$Q(p) = F^{-1} = t$$

Relation between Q and CDF

- when T is continuous, $Q(p) = t$ only if $F(t) = p$.
- for a probability p between 0 and 1, the quantile function is the minimum value of time t below which a randomly selected person from the given population will fall $p * 100$ percent of the times.

Hypothetical survival data with no censoring (1)

- A sample of 5 persons experienced the event at $t = 5$, 10, 15, 20, and 90 days.



Hypothetical survival data with no censoring (1)

- A sample of 5 persons experienced the event at $t = 5$, 10, 15, 20, and 90 days.

Cumulative distribution function

$$F(5) = Pr(T \leq 5) = 0.2$$

Survival function

$$S(5) = 1 - 0.2 = 0.8$$

Quantile function

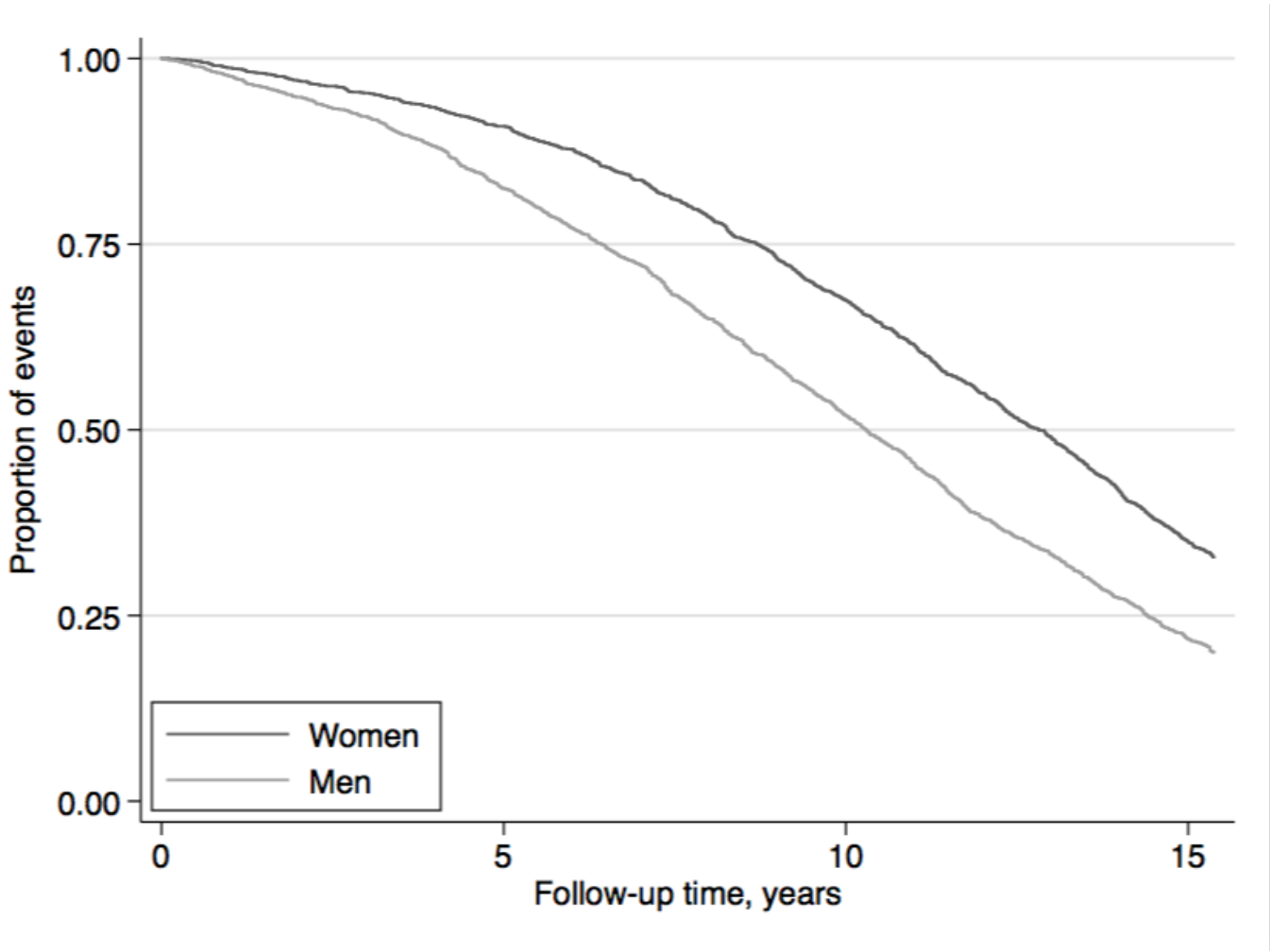
$$Q(0.2) = 5 \text{ days}$$

Interpretation

- 20th-percentile of survival time is 5 days
- 20% (1 out of 5 persons) of the population experienced the event within 5 days
- A person randomly selected from this population has a probability of 20% of experiencing the event within 5 days

Group comparison: survival curves

- The survival curves show differences in all survival percentiles



Estimation of survival percentiles - Univariable

- The most common estimator of the survival curve is the non-parametric Kaplan-Meier method
- SAS: proc lifetest. R: package 'survival'. Stata: `sts graph`, `stqkm`.
- `stqkm` provides differences in survival percentiles with CI. It can be installed by typing:
  `net install stqkm, ///`
  `from(http://www.imm.ki.se/biostatistics/stata) replace`

Estimation of survival percentiles - Multivariable

- Common situation in epidemiological studies, when one needs to adjust for potential confounders
- Methods of quantile regression for censored data
- Recent developments (Powell, Portnoy, Peng-Huang)
- R: package 'quantreg'. SAS: proc quantlife
- Bottai & Zhang introduced Laplace regression in 2010

Laplace regression

- When the time variables T_i may be censored we observe the covariates x_i, $y_i = \min(t_i, c_i)$, and $d_i = I(t_i \leq c_i)$
- The aim is to estimate the τ^{th} conditional quantile of T_i
- A Laplace regression model establishes a linear relationship between a given percentile of T and a set of covariates

$$t_i(\tau) = x_i^{'}\beta(\tau) + \sigma_i(\tau)u_i$$

- u_i follows the Asymmetric Laplace distribution
- Estimation is conducted via maximum-likelihood, and standard errors are preferrabily estimated via bootstrap

Laplace regression - Stata

- The program can be installed from `net install laplace, /// from(http://www.imm.ki.se/biostatistics/stata) replace`

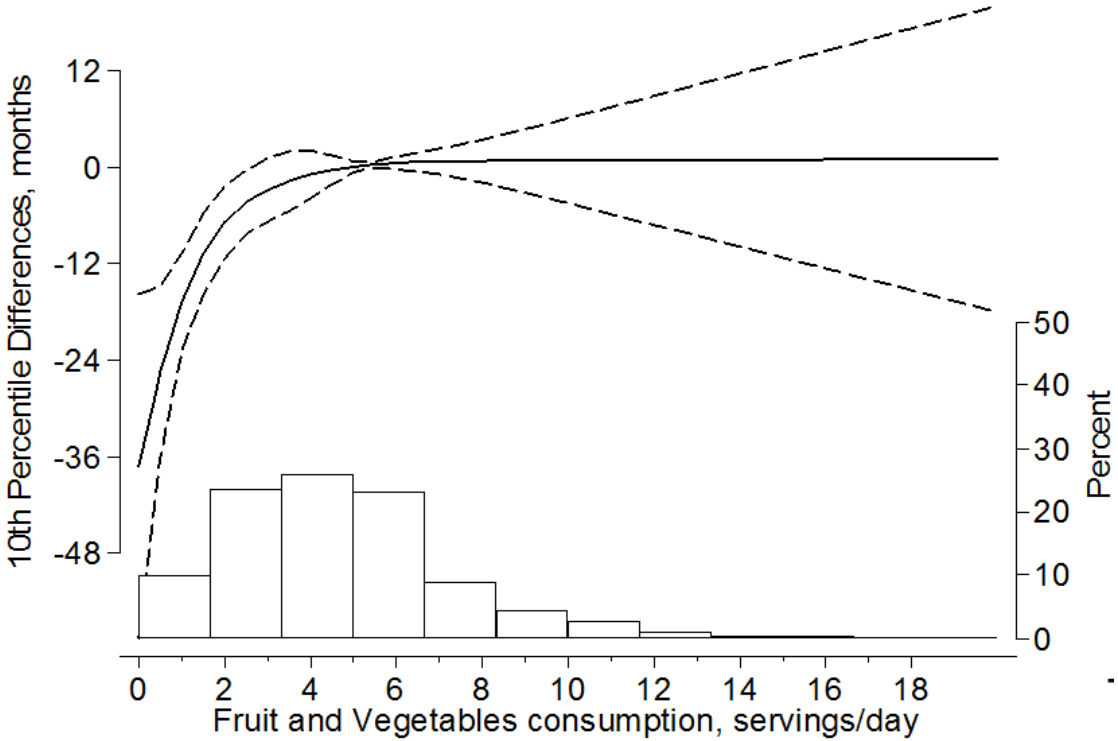
Example

```
sysuse cancer, clear
xi:  laplace studytime i.drug, fail(died)
xi:  laplace studytime i.drug, q(.25 .5 .75) fail(died)
```

Laplace regression - Example (1)

- Study on Fruit and Vegs consumption and survival (AJCN 2013)
- Study population: 71,706 men and women from central Sweden
- Exposure: Fruit and Vegetables consumption, servings/day
- Outcome: Time to death
- Potential confounders: age, gender, smoking, alcohol, physical activity, bmi, energy intake, education
- Follow-up time: 13 years
- Cases: 11,439 deaths (15%)
- Measure of association: differences in the $10th$ survival percentile
- Analysis: FV consumption was flexibly modeled by means of right restricted cubic splines

Laplace regression - Example (2)



- *Lower FV consumption was increasingly associated with shorter survival, up to 3 years for those who never ate FV.*

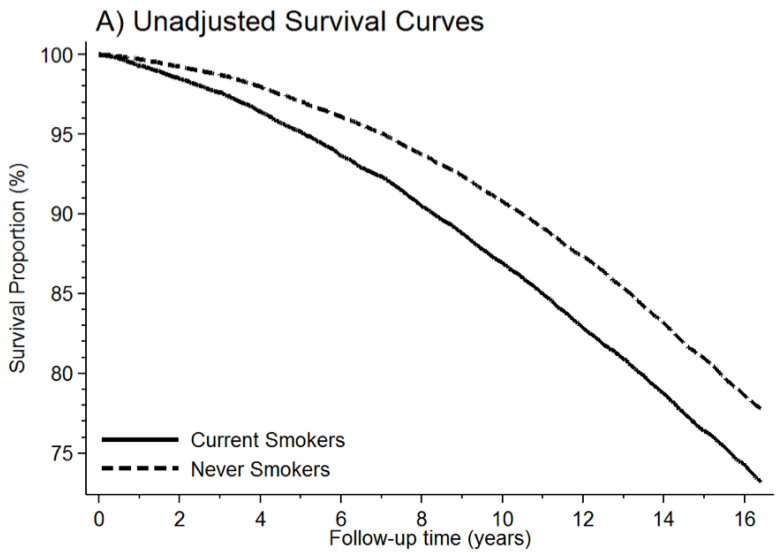
4. Recent developments

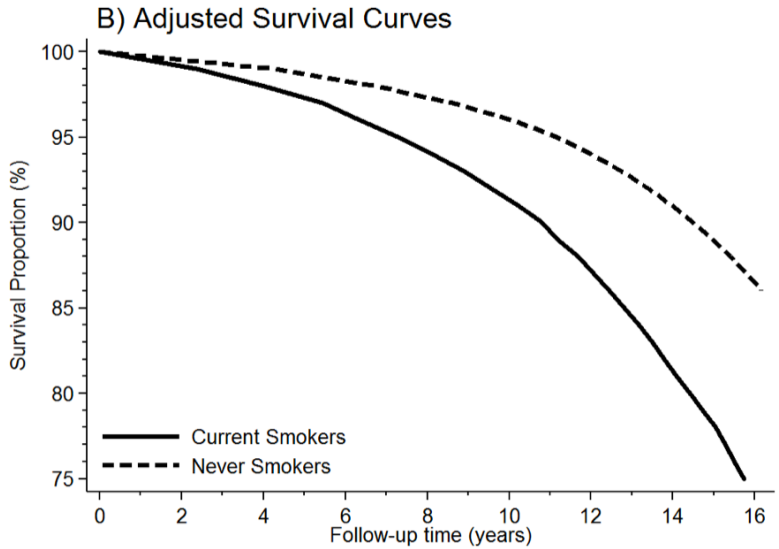
a) Adjusted survival curves

- With Laplace regression we can estimate multivariable-adjusted differences in survival percentiles
- One could focus on all percentiles within the observed range. The Stata command `laplace` allows simultaneously modeling different percentiles
- Coefficients obtained at different survival percentiles can be combined to derive adjusted survival curves
- Published on Epidemiology, 2015

Adjusted survival curves (2)

- Adjusted curves calculated by estimating a Laplace regression model for all percentiles from the 1st to the 25th, adjusting for baseline age, body mass index, and sex





Adjusted survival curves (3)

- We have developed a Stata post-estimation command, `laplace_surv` to calculate and draw adjusted or marginal survival curves

Example

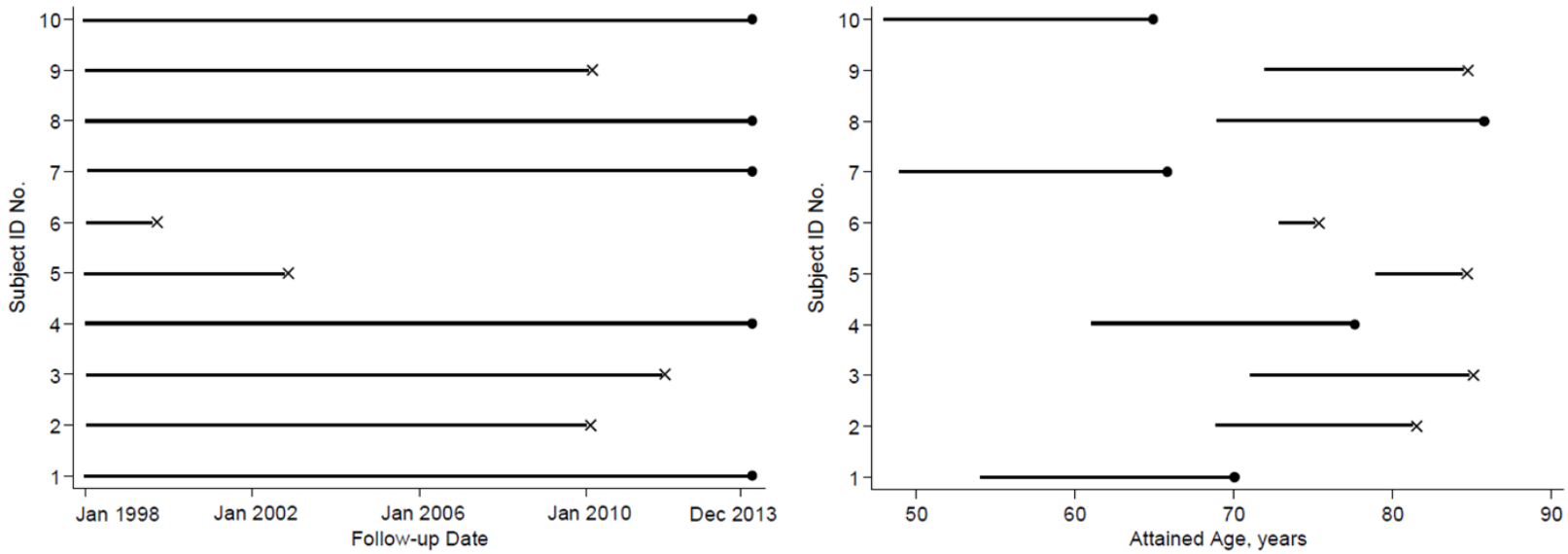
```
laplace time female age, fail(infect) q(1(1)70)
laplace_surv, at1(female=0) at2(female=1) line
```

b) Age as Time-scale

- The time variable can be defined in different ways (e.g. follow-up time, attained age, calendar time . . .)
- All previous slides were focusing on follow-up time, defined as the time from entering the study until event or censoring
- Another common option is to focus on attained age. When data are analyzed with Cox regression this choice is recommended and it is becoming the standard
- Can we model the percentiles of attained age as we model survival percentiles?
- On press in American Journal of Epidemiology, 2015

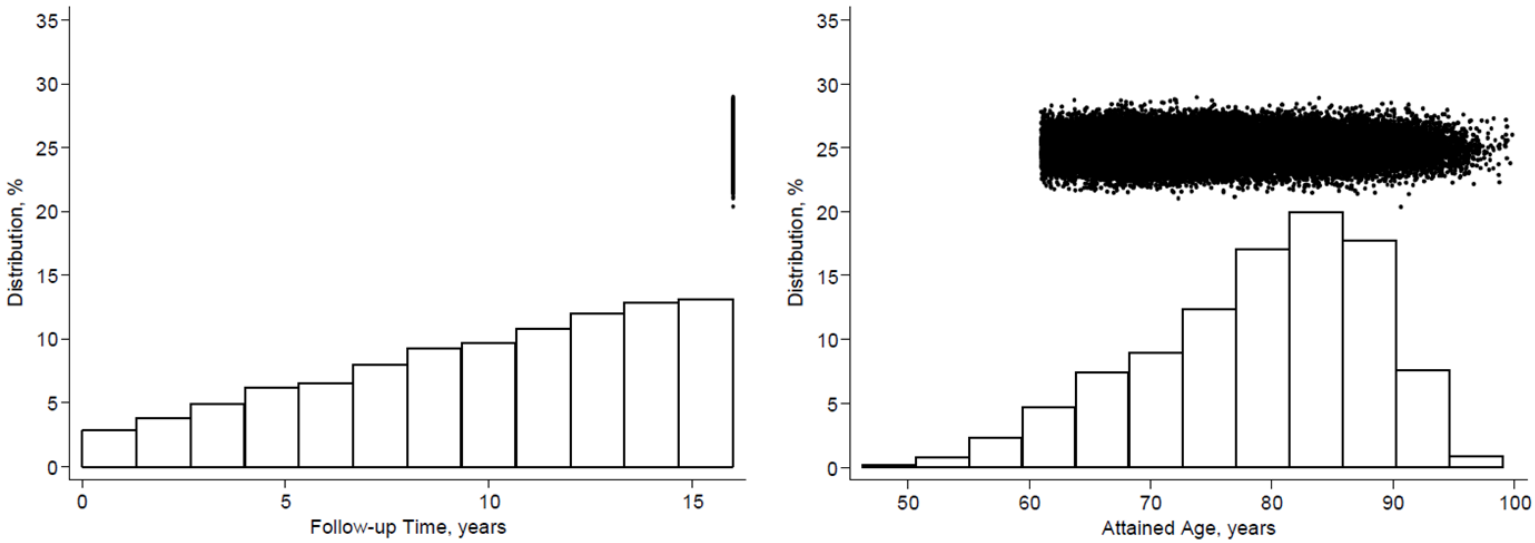
Consequences of changing time-scale

1. We introduce delayed entries, leading to left-truncation



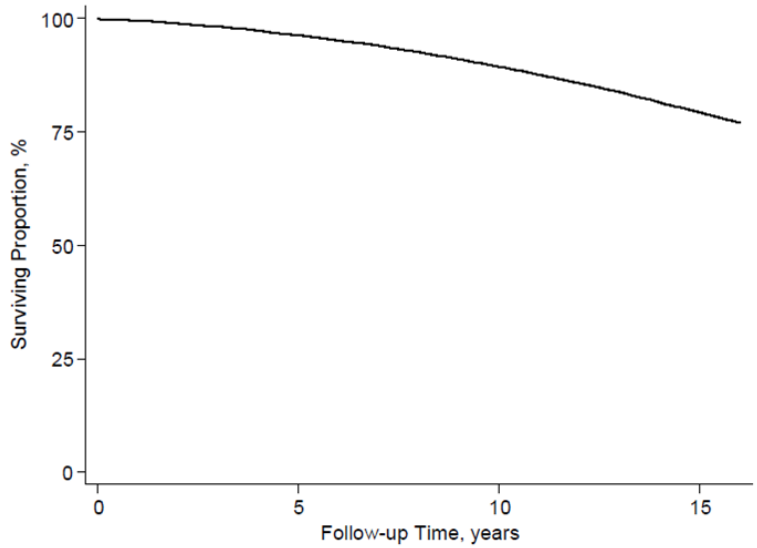
Consequences of changing time-scale

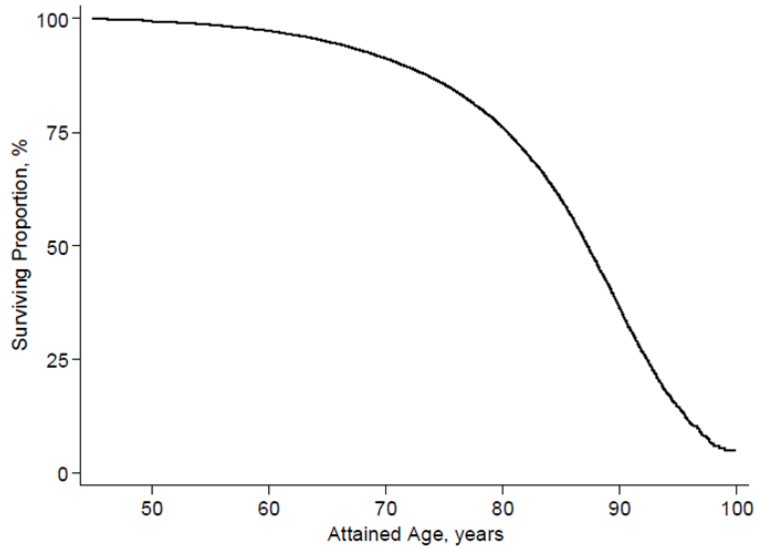
2. Censored observations are spread throughout the time-scale



Consequences of changing time-scale

3. The survival curve can still be calculated but becomes hard to interpret





Laplace regression to model attained age

- We change the time-variable from T_i to A_i, the attained age at event, or censoring, for participants i
- We fit a Laplace regression model on the pth percentile of A_i.
- To get meaningful estimates we can further adjust for a function of age at baseline

$$A_i(p) = \beta_0(p) + \beta_1(p) \cdot f(age\_baseline_i)$$

- This model can be extended to include other covariates and interactions terms

Differences in attained age

- Suppose we are interested in the difference in age at the event between men and women

$$A_i(p) = \beta_0(p) + \beta_1(p) \cdot gender_i + \beta_2(p) \cdot f(age\_baseline_i)$$

- $\beta_1(p)$ represents the difference in the pth percentile of age at event between men and women

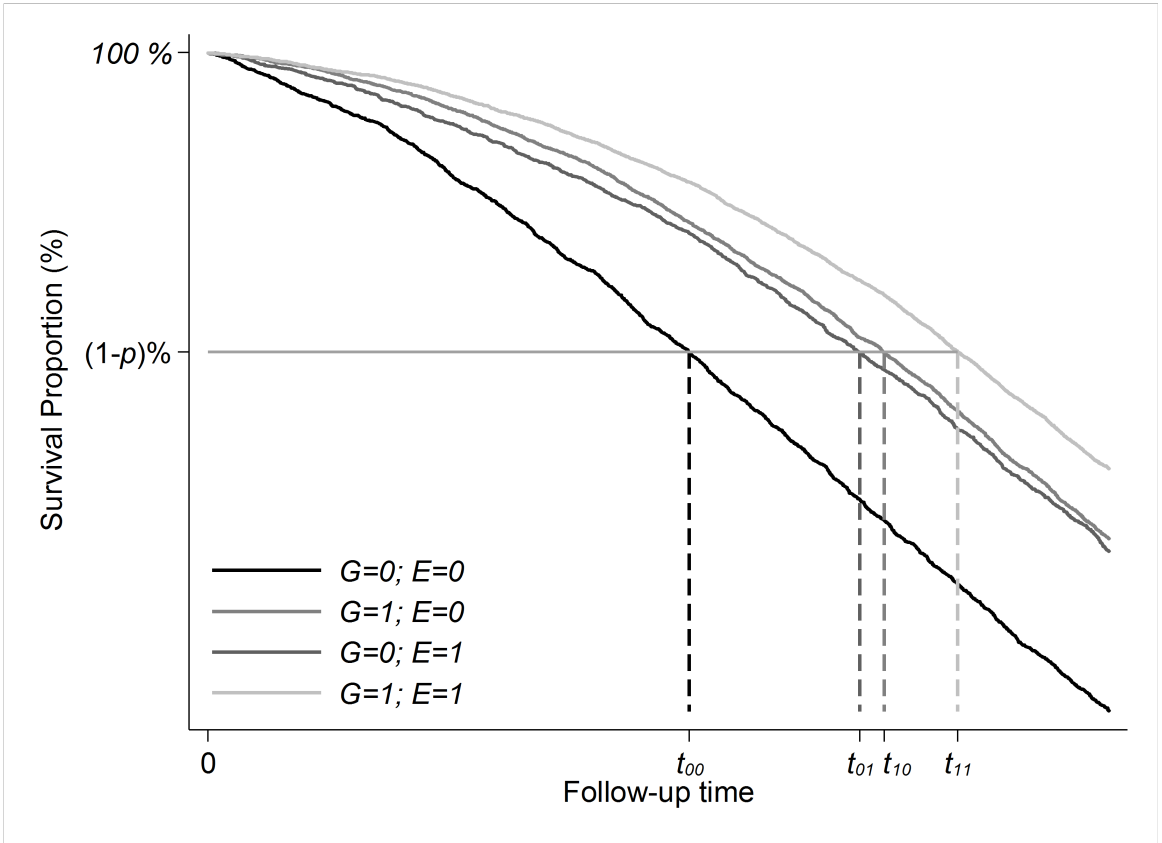
c) Evaluating additive interaction

- Statistical interaction can be evaluated on the additive or the multiplicative scale
- Presentation of both scales is recommended
- In survival analysis, because of the popularity of Cox regression, the multiplicative scale alone is usually presented
- We defined the concept of interaction in the context of survival percentiles and presented how to evaluate additive interaction
- Epidemiology, Under review

Interaction in the context of survival percentiles

- We can define a measure of additive interaction at the pth percentile as:

$$I_p = (t_{11} - t_{00}) - [(t_{10} - t_{00}) + (t_{01} - t_{00})]$$



Evaluating additive interaction with Laplace regression

- Including an interaction term between two exposures G and E will serve as a test for additive interaction

$$T(p|G,E) = \beta_0(p) + \beta_1(p) \cdot G + \beta_2(p) \cdot E + \beta_3(p) \cdot G \cdot E$$

- $\beta_3(p)$ represents the excess in survival due to the presence of both exposures G and E

Summary

- Percentiles provide a complete summary of a continuous outcome
- In survival analysis we focus on survival percentiles, defined as time by which a certain proportion of the participants have experienced the event of interest
- The survival curves depicts all observed survival percentiles and can be estimated with the Kaplan-Meier method
- Adjusted survival percentiles can be estimated with Laplace regression

References

- Beyerlein A. *Quantile Regression - Opportunities and Challenges From a User's Perspective.* American Journal of Epidemiology. 2014.
- Orsini N et al. *Evaluating percentiles of survival.* Epidemiology. 2012.
- Bellavia A et al. *Fruit and vegetable consumption and all-cause mortality: a dose-response analysis.* American Journal of Clinical Nutrition. 2013.
- Bellavia A et al. *Adjusted Survival Curves with Multivariable Laplace Regression.* Epidemiology. 2015.
- Bellavia A et al. *Using Laplace regression to model and predict percentiles of age at death, when age is the primary time-scale.* American Journal of Epidemiology. 2015.
- Bottai M, Zhang J. *Laplace regression with censored data.* Biometrical Journal. 2010.
- Koenker R, Bassett Jr G. *Regression quantiles.* Econometrica: journal of the Econometric Society. 1978.
- Powell JL. *Censored regression quantiles.* Journal of econometrics. 1986.
- Portnoy S. *Censored regression quantiles.* JASA. 2003.
- Peng L, Huang J. *Survival analysis with quantile regression models.* JASA. 2008.