Quantile Regression in Survival Analysis
Evaluating Survival Percentiles

Andrea Bellavia

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

VI BIAS Annual Congress
Genova, 30 October 2014
Quantiles and percentiles: why?

- In this presentation we focus on continuous outcomes (time)
- Common summary measures are 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
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.
1. Comparing distributions

- We now want to evaluate gender differences in alcohol consumption

<table>
<thead>
<tr>
<th>Average alcohol consumption among men and women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men: 14.3 grams/day</td>
</tr>
<tr>
<td>Women: 6.7 grams/day</td>
</tr>
</tbody>
</table>

- 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.
2. Focusing on specific percentiles

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

- Mean = 25.3 Kg/m²; 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/m²) and obese (>30 Kg/m²) 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/m² in the example), or the 90th (30 Kg/m²), according to a set of covariates.
Quantile regression: summary

- A percentile approach allows focusing on specific percentiles of interest
- The entire shape of the distribution is taken into account
- Quantile regression is the common way to model percentiles
- Modelling the median is more intuitive and dates back to Boscovich, \(\sim 1700\) (one century before linear regression). Mathematical and computational complexities slowed down its development
Survival Analysis

- In survival analysis the continuous outcome Y is a time variable time T
- T can be defined in different ways (e.g. follow-up time, age)
- The main differences between T and a common variable Y are censoring and skewness
Survival Percentiles

- The percentiles of a time variable $T$ are usually 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 survival curve is a summary of survival percentiles

- Because of censoring we do not observe all percentiles
The 25th survival percentile and the 50th survival percentile (median survival) are shown in the figure.
Group comparison: survival curves

- The survival curves show differences in all survival percentiles
Why to focus on survival percentiles?

- We aim to summarise the association between an exposure and survival. In survival analysis this is usually done by estimating Hazard Ratios (HR) of the event between exposed and unexposed.
- This is appealing, because we only need one number.
- However, as with any other continuous variable (see slide 6), we risk to lose a lot of information.
- In survival analysis, different features make a percentile approach useful.
1. The percentile is intuitive

- It combines information on time and risk
- For example, if the 10th percentile of survival is 10 years, that means that the first 10% of the study population died within 10 years
- Interpretation is simpler than other commonly reported measures
- Median survival (50th percentile) = 11.4 years
- Mortality rate = 12336 cases/909684 person-years = 0.014
2. The percentile facilitates individual interpretation

“Your hazard ratio is 0.72 assuming proportional hazards.”
2. The percentile facilitates individual interpretation

“90% of patients like you have at least 10 more years to live.”
3. Measures of association can be expressed in terms of absolute time differences (PD=percentile differences)

- HR=1.48 - Men died at higher rate than women
- 50th PD=2.2 - Median survival was 2 years longer for women
3.1 Example

- Cancer with high fraction of mortality (e.g. pancreatic cancer)
- Survival after pancreatic cancer diagnosis
3.1 Is physical activity a risk factor?

<table>
<thead>
<tr>
<th>Level of Physical activity</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>1</td>
<td>Ref.</td>
</tr>
<tr>
<td>Walk 1-30 min/day</td>
<td>0.95</td>
<td>(0.92, 0.97)</td>
</tr>
<tr>
<td>Walk 31-60 min/day</td>
<td>0.87</td>
<td>(0.82, 0.93)</td>
</tr>
<tr>
<td>Walk &gt;60 min/day</td>
<td>0.86</td>
<td>(0.82, 0.91)</td>
</tr>
</tbody>
</table>

- Mortality risk after diagnosis is decreasing by increased levels of physical activity (HRs are adjusted for potential confounders)
3.1 Adjusted median survival differences

<table>
<thead>
<tr>
<th>Level of Physical activity</th>
<th>50th PD (days)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive</td>
<td>0</td>
<td>Ref.</td>
</tr>
<tr>
<td>Walk 1-30 min/day</td>
<td>6</td>
<td>(3, 10)</td>
</tr>
<tr>
<td>Walk 31-60 min/day</td>
<td>18</td>
<td>(9, 27)</td>
</tr>
<tr>
<td>Walk &gt;60 min/day</td>
<td>22</td>
<td>(14, 30)</td>
</tr>
</tbody>
</table>

- The lower mortality risk results in a difference of less than one month of survival between physically active and inactive subjects, after taking into account potential confounders.
4. Estimation of survival percentiles minimise data extrapolation

- Because of censoring we do not observe all percentiles
- Some common measures require assumptions on the unobserved percentiles (e.g. life expectancy)
- A percentile approach allows focusing only on observed percentiles
In this situation focusing on median survival would require data extrapolation. We can look up to the 20th percentile.
5. Focus on early percentiles

- Sometimes we are not interested in median survival
- Example - Time from diagnosis to death
- We wish to understand who is more likely to die soon after diagnosis
- Focusing on early percentiles (like the 5th) is a valid option
6. Time-varying effects

- When there is proportionality of the hazards, COX regression is the best choice.
- This situation is similar to the shift of 7 points in the alcohol consumption distribution (slide 6).
- If this assumption does not hold, by summarising the association with one measure (HR) we lose important information.
- There are advanced extensions of Cox regression to handle this situation (such as inclusion of time-varying covariates).
- Evaluating survival percentiles is an intuitive, flexible, and simpler approach.
6.1 Bladder cancer example (Byar, 1980)

- Study on bladder cancer
- Tumors were removed from 86 patients
- Subsequently RCT was started, with patients assigned to either placebo (48) or drug thiopeta (38)
- Follow-up: 60 months
- The outcome of interest was tumor recurrence
6.1 Bladder cancer example

- Summarize the study
- Incidence Rate Ratio (IRR)=0.59, 95% CI:0.31, 1.10
- Tumor recurrence rate in the treatment group was 41% lower than in the placebo group
- Hazard Ratio (HR)=0.70, 95% CI: 0.38, 1.26
- Hazard of tumor recurrence was 30% lower in the treatment group
6.1 Survival Curve by treatment group
6.1 Percentile differences (PDs)

- 5th PD = -0.4 months; 95% CI: -3.4, 2.6
- 10th PD = -0.0 months; 95% CI: -2.9, 2.9
- 25th PD = -1.2 months; 95% CI: -3.2, 5.7
- 50th PD = 10.0 months; 95% CI: -7.9, 27.9
6.2 Extreme situation (simulated data)

- Effect of a (protective) exposure on mortality after diagnosis
- 2000 subjects with diagnosis of cancer A
- 1000 exposed, 1000 unexposed
- 1536 subjects die within 3 years of follow-up (738 exposed and 798 exposed)
- Is mortality lower in the unexposed group? Are the groups dying at the same rate?
- HR=0.98; 95% CI: 0.89-1.09
25th PD = -0.31 years; 95% CI: -0.37, -0.25
The first 25% of exposed subjects died 4 months later than the first 25% of unexposed
Unadjusted Survival Percentiles

- Survival percentiles are depicted in the survival curve
- The most common estimator of the survival curve is the non-parametric Kaplan-Meier method
- KM estimator was used throughout this presentation (slides 14, 15, 16, 19, 22, 23, 25, 28, 33, 36)
- 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
Adjusted survival percentiles

- Common situation in epidemiology
- There are two main approaches:
  1) Estimate a multivariable parametric (AFT, flexible parametric survival), or semi-parametric (COX) model. Back-calculate the survival function. Derive adjusted survival percentiles
- Computational and mathematical complexity, plus relying on the original model assumption, limit this application
Adjusted survival percentiles

- 2) 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

- A Laplace regression model establishes a linear relationship between the $p$th survival percentile and a set of covariates

$$T_i(p) = x_i' \beta(p)$$

- Laplace regression estimates are interpreted as multivariable-adjusted differences in survival percentiles according to level of a given exposure
- We have used Laplace regression in slides 22, 26, 34, and 36
Laplace regression - Features

- Directly models the percentile
- Allows including interaction and covariate transformations
- Fit multiple percentiles in one fit, and test within and between percentiles
- Available only in Stata (so far) [net install laplace, from(http://www.imm.ki.se/biostatistics/stata) replace]
Adjusted survival curves

- By estimating all the observed percentiles we can derive adjusted survival curves
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.
- Focusing on survival percentiles has different advantages.
- The survival curves depict 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.” AJE. 2014.
Acknowledgments

- Nicola Orsini
- Matteo Bottai
- Paolo Frumento
- Andrea Discacciati
- Alicja Wolk