Modeling Geometric Rates with Quantile Regression

Matteo Bottai, Sc.D.

Unit of Biostatistics
Karolinska Institutet
Stockholm, Sweden

A Motivating Example

A cohort of 100 subjects is followed up for 2 days.

Day	Alive
0	100
1	20
2	16

Geometric rate $= 1 - (16/100)^{1/2} = 0.60$
The probability of dying in a day's time is 0.60.
Suppose 100 subjects die at a constant daily rate of 0.60.
Then $100(1 - 0.60)^2 = 16$ are alive at day 2.

A Motivating Example

A cohort of 100 subjects is followed up for 2 days.

Day	Alive
0	100
1	20
2	16

Geometric rate $= 1 - (16/100)^{1/2} = 0.60$
The probability of dying in a day's time is 0.60.
Suppose 100 subjects die at a constant daily rate of 0.60.
Then $100(1 - 0.60)^2 = 16$ are alive at day 2.

Incidence rate $= 84/(80 \cdot 1 + 20 \cdot 2) = 0.70$ deaths/person-day
At a rate of 0.70, $100(1 - 0.70)^2 = 9 \neq 16$.

The Geometric Rate

Let T be a continuous time variable with support on $\mathcal{R}_+$.

Let $S(t) \equiv P(T > t)$ be the survival function.

The geometric rate over the time interval $(0, t)$ is

$$g(0, t) = 1 - S(t)^{1/t}$$

Geometric and Incidence Rates

The geometric rate is

$$1 - S(t)^{1/t}$$

and the incidence rate is

$$\frac{1 - S(t)}{\int_0^t S(u)du}$$

For example, if $S(t) = \exp(-\lambda t)$,

Geometric rate $= 1 - \exp(-\lambda)$
Incidence rate $= \lambda$

The two rates are constant but different from one another.

A Conjecture

Conjecture. *There does not exist a survival function* $S(t) \equiv P(T > t)$ *such that*

$$1 - S(t)^{1/t} = \frac{1 - S(t)}{\int_0^t S(u)du}$$

for all $t \in (0, \infty)$.

Instantaneous Geometric Rates and Hazards

The geometric rate over shrinking intervals $(t, t+h)$

$$\begin{aligned}\lim_{h\downarrow 0} 1 - \left[\frac{S(t+h)}{S(t)}\right]^{1/h} &= \lim_{h\downarrow 0} 1 - \exp\left[\frac{\log S(t+h) - \log S(t)}{h}\right] \\ &= 1 - \exp[d\log S(t)/dt] \\ &= 1 - \exp[-f(t)/S(t)] \\ &= 1 - \exp[-h(t)]\end{aligned}$$

where $f(t)$ is the PDF and $h(t) \equiv f(t)/S(t)$ the hazard function.

Instantaneous Geometric Rates and Hazards

The geometric rate over shrinking intervals $(t, t+h)$

$$\begin{aligned}\lim_{h\downarrow 0} 1 - \left[\frac{S(t+h)}{S(t)}\right]^{1/h} &= \lim_{h\downarrow 0} 1 - \exp\left[\frac{\log S(t+h) - \log S(t)}{h}\right] \\ &= 1 - \exp[d\log S(t)/dt] \\ &= 1 - \exp[-f(t)/S(t)] \\ &= 1 - \exp[-h(t)]\end{aligned}$$

where $f(t)$ is the PDF and $h(t) \equiv f(t)/S(t)$ the hazard function.

The limit of the incidence rate is

$$\lim_{h\downarrow 0} \frac{S(t) - S(t+h)}{\int_t^{t+h} S(u)du} = \frac{f(t)}{S(t)} = h(t)$$

The instantaneous geometric rate and the hazard are different.

Geometric Rates over Adjacent Time Intervals

The geometric rate is between two time points t_1 and t_2 is

$$g(t_1, t_2) = 1 - [S(t_2)/S(t_1)]^{1/(t_2 - t_1)}$$

The geometric rates can be concatenated as follows

$$g(0, t_2) = 1 - [1 - g(0, t_1)]^{t_1} \cdot [1 - g(t_1, t_2)]^{t_2 - t_1}$$

The value $S(t_2)/S(t_1)$ is the Kaplan-Meier step over (t_1, t_2).
The geometric rate is a weighted average of Kaplan-Meier steps.

The Problem

Poisson regression models the incidence rate

No regression method so far models the geometric rate

Geometric Rate Over Proportions of Events

If $P(T \leq t) = p$,

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

where $Q(p)$ is the quantile function.

The geometric rate over the time interval $(0, t)$

$$g(0, t) = 1 - S(t)^{1/t}$$

is equal to the geometric rate over the proportion interval $(0, p)$

$$g(0, p) = 1 - (1 - p)^{1/Q(p)}$$

A Proposition

Proposition. *The geometric rate*

$$g(0, p) = 1 - (1 - p)^{1/Q(p)}$$

is the $(1 - p)$*-quantile of the transformed time variable*

$$T^* = 1 - (1 - p)^{1/T}$$

That is $P[T^* \leq g(0, p)] = 1 - p$.

The above proposition follows directly from the fact that for a fixed p the function $1 - (1 - p)^{1/t}$ is monotonically decreasing in t for $t > 0$.

Proof of the Above Proposition

$$
\begin{aligned}
P[T^* \le g(0,p)] &= P[1-(1-p)^{1/T} \le 1-(1-p)^{1/Q(p)}] \\
&= P[-(1-p)^{1/T} \le -(1-p)^{1/Q(p)}] \\
&= P[(1-p)^{1/T} \ge (1-p)^{1/Q(p)}] \\
&= P[\log(1-p)/T \ge \log(1-p)/Q(p)] \\
&= P[1/T \le 1/Q(p)] \\
&= P[T \ge Q(p)] \\
&= 1 - P[T < Q(p)] \\
&= 1 - p
\end{aligned}
$$

For discrete time variables, the proportion to be used in quantile regression estimation may be set to $(1-p+\epsilon)$ instead of $(1-p)$, where $\epsilon > 0$ is a sufficiently small positive quantity.

A Regression Method for Geometric Rates

Suppose the conditional geometric rate given covariates is

$$g(0,p|x) = x'\beta_p$$

for a set of covariates $x \in \mathcal{R}^k$ and a parameter $\beta_p \in \mathcal{R}^k$.
The value $x'\beta$ is the conditional $(1-p)$-quantile of T^*

$$Q_{T^*}(1-p|x) = x'\beta$$

Given a sample t_i and x_i, $i = 1, \dots, n$,

$$\hat{\beta} = \arg\min (t_i^* - x_i'\beta)[1-p-I(t_i^* \le x_i'\beta)]$$

where $t_i^* = 1-(1-p)^{1/t_i}$.

The estimate can be obtained with quantile regression.
$\hat{\beta}$ shares the properties of quantile regression estimators.

Interpretation of the Regression Coefficient

Consider the regression model

$$g(0,p|x) = x'\beta_p$$

The coefficient β_p is similar to that of any regression method.
It represents the change in the rate for one-unit increase in x.

The rate difference between two covariate patterns x_0 and x_1 is

$$g(0,p|x_1) - g(0,p|x_0) = (x_1 - x_0)'\beta_p$$

Modeling Geometric Rate Ratios

Often rate ratios are preferred to rate differences.
Suppose

$$\log g(0,p|x) = x'\gamma_p$$

for a vector $\gamma_p \in \mathcal{R}^k$.

The rate ratio between two covariate patterns x_0 and x_1 is

$$g(0,p|x_1)/g(0,p|x_0) = \exp[(x_1 - x_0)'\gamma_p]$$

The transform $\log[1-(1-p)^{1/t}]$ is monotonic in t for $t > 0$.
The coefficient γ_p can be estimated with quantile regression.
The dependent variable is $t_i^* = \log[1-(1-p)^{1/t_i}]$.

Survival in a Cohort of Men from Sweden

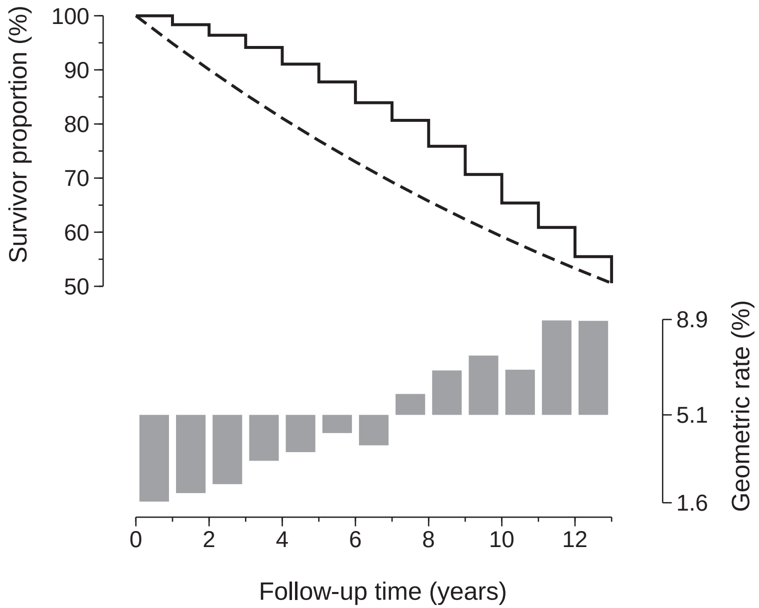
A cohort was recruited in Sweden (Zhen et al, 2014).
Participants were follow-up from Jan 1, 1998, to Jan 1, 2011.

I analyzed 3,280 healthy 70-79 years-old men.
At the end of the follow-up time, 1,659 men were still alive.

The average annual geometric mortality rate was

$$1 - (1659/3280)^{1/13} = 0.051$$

Survival in a Cohort of Men from Sweden



Mortality Rates and Physical Activity

I estimated mortality rates across levels of physical activity.
The geometric-rate regression model was the following

$$\log g(0, 0.25|x) = \gamma_0 + \gamma_1 \text{PA2} + \gamma_2 \text{PA3} + \gamma_3 \text{PA4}$$

The PA's were indicators of physical activity level.
The lowest level, PA1, was the referent group.

The coefficient γ was estimated with quantile regression.
I estimated the 0.75-quantile of $T^* = \log[1 - (1 - 0.25)^{1/T}]$.

Adjusted Mortality Rates

A second geometric-rate regression model was also estimated

$$\log g(0, 0.25|x) = \gamma_0 + \gamma_1 \text{PA2} + \gamma_2 \text{PA3} + \gamma_3 \text{PA4} + \text{covariates}$$

I included the following additional categorical covariates:

Age (70-71, 72-73, ..., 78-79 years)
Alcohol consumption (four groups)
Smoking habits (four groups)
Waist circumference (four groups)
Fruit and vegetables consumption (four groups)

Mortality Rate Ratios

The table shows the estimated rates for the first 25% of deaths.

Physical Activity	Quantile (years)	Annual Rate (%)	Rate Ratio	
			Crude	Adjusted
Very low	6.3 (5.5 7.0)	4.5 (4.1 4.9)	1.00 (referent)	1.00 (referent)
Low	8.7 (7.9 9.4)	3.3 (3.0 3.6)	0.73 (0.64,0.82)	0.74 (0.65,0.84)
High	9.6 (8.9 10.4)	2.9 (2.7 3.2)	0.66 (0.58,0.75)	0.74 (0.65,0.83)
Very high	9.8 (9.1 10.6)	2.9 (2.6 3.2)	0.64 (0.57,0.73)	0.73 (0.64,0.82)

The mortality rate decreased over levels of physical activity.
In the most active it was 36% smaller than in the least active.

Final Remarks

Geometric rates are not used in biomedical sciences

They are applied in demography and bank accounts.

Incidence rates are different from geometric rates.

Geometric rates are quantiles of a transformed time variable.

A quantile of any variable is a geometric rate of a transform.

With censoring, use Kaplan-Meier and censored Q-regression.

Assumptions can be made to improve efficiency:

$$S(t) = \exp[-(\lambda t)^{\theta}] \Leftrightarrow g(0,p) = 1 - (1-p)^{\lambda[-\log(1-p)]^{-1/\theta}}$$

References

- Bottai M, Cai B and McKeown RE. Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, 2010, 29: 309–17.
- Bottai M and Zheng J. Laplace regression with censored data. *Biometrical Journal*, 2010, 52: 487–503.
- Bottai M and Orsini N. A command for Laplace regression. *Stata Journal*, 2013, 13: 302–314.
- Koenker R, *Quantile Regression*. Cambridge University Press, 2005.
- Zheng Selin J, Orsini N, Ejdervik Lindblad B, Wolk A. Long-Term Physical Activity and Risk of Age-Related Cataract: A Population-Based Prospective Study of Male and Female Cohorts. *Ophthalmology*, 2014 (in press, doi: 10.1016/j.ophtha.2014.08.023).
- Portnoy S. Censored regression quantiles. *Journal of the American Statistical Association*, 2003, 98: 1001–1012.
- Koenker R. Censored quantile regression redux. *Journal of Statistical Software*, 2008, 27: 1–24.