Reappraising Medfly Longevity:
A Quantile Regression Survival Analysis

Roger Koenker
and
Olga Geling

University of Illinois, Urbana-Champaign

http://www.econ.uiuc.edu
Carey Medfly Experiment

- 1,203,646 medflies survival times recorded in days
- Sex determined at time of death
- Pupae were sorted into one of five size classes
- 167 aluminum mesh cages of roughly 7200 flies each
- Adults were given a diet of sugar and water *ad libitum*
Life Cycle of Medfly
Medfly Findings

- Mortality rates actually *declined* at the oldest observed ages. contradicting the view that aging is an inevitable, monotone process of senescence.

- The right tail of the survival distribution was, at least by human standards, remarkably long.

- The experiment provided strong evidence for a crossover in gender specific mortality rates.
Figure 1: Raw mortality rate data for the full sample with a smoothed (7-day geometric moving average) estimate of the hazard function.
**Medfly Survival Prospects**

<table>
<thead>
<tr>
<th>Lifespan (in days)</th>
<th>Percentage Surviving</th>
<th>Number Surviving</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>5</td>
<td>60,000</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>12,000</td>
</tr>
<tr>
<td>86</td>
<td>.01</td>
<td>120</td>
</tr>
<tr>
<td>146</td>
<td>.001</td>
<td>12</td>
</tr>
</tbody>
</table>

**Human Survival Prospects* **

<table>
<thead>
<tr>
<th>Lifespan (in years)</th>
<th>Percentage Surviving</th>
<th>Number Surviving</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>98</td>
<td>591,000</td>
</tr>
<tr>
<td>75</td>
<td>69</td>
<td>413,000</td>
</tr>
<tr>
<td>85</td>
<td>33</td>
<td>200,000</td>
</tr>
<tr>
<td>95</td>
<td>5</td>
<td>30,000</td>
</tr>
<tr>
<td>105</td>
<td>.08</td>
<td>526</td>
</tr>
<tr>
<td>115</td>
<td>.0001</td>
<td>1</td>
</tr>
</tbody>
</table>

* Estimated Thatcher (1999) Model
From Fly to Man
Figure 2: Smoothed mortality rates for males and females.
Some References


Survival/Transformation Models

A wide variety of survival analysis models, Doksum and Gasko (1990), may be written as,

\[ h(T_i) = x_i' \beta + u_i \]

where \( h \) is a monotone transformation,
\( T_i \) is an observed survival time,
\( x_i \) is a vector of covariates,
\( \beta \) is an unknown parameter vector
\( \{u_i\} \) are iid with df \( F \).
Cox Model

For the proportional hazard model with

$$\log \lambda(t|x) = \log \lambda_0(t) - x'\beta$$

the conditional survival function in terms of the integrated baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ as,

$$\log(-\log(S(t|x))) = \log \Lambda_0(t) - x'\beta$$

so

$$\log \Lambda_0(T) = x'\beta + u$$

for $u_i$ iid with df $F_0(u) = 1 - e^{-e^u}$. 
Bennett Model

For the proportional odds model, where the conditional odds of death
\( \Gamma(t|x) = F(t|x)/(1 - F(t|x)) \) are written as,

\[
\log \Gamma(t|x) = \log \Gamma_0(t) - x'\beta,
\]

we have

\[
\log \Gamma_0(T) = x'\beta + u
\]

for \( u \) iid logistic with \( F_0(u) = (1 + e^{-u})^{-1} \).
In the accelerated failure time model we have

$$\log(T_i) = x_i'\beta + u_i$$

so

$$P(T > t) = P(e^u > te^{-x_\beta}) = 1 - F_0(te^{-x_\beta})$$

where $F_0(\cdot)$ denotes the df of $e^u$, and thus,

$$\lambda(t|x) = \lambda_0(te^{-x_\beta})e^{-x_\beta}$$

where $\lambda_0(\cdot)$ denotes the hazard function corresponding to $F_0$. In effect, the covariates act to rescale time in the baseline hazard.
The simplest version of quantile regression is the two-sample treatment-control model, Lehmann (1974) treatment response model:

“Suppose the treatment adds the amount $\Delta(x)$ when the response of the untreated subject would be $x$. Then the distribution $G$ of the treatment responses is that of the random variable $X + \Delta(X)$ where $X$ is distributed according to $F$.”
Lehmann QTE as a QQ-Plot

Doksum (1974) defines $\Delta(x)$ as the “horizontal distance” between $F$ and $G$ at $x$, i.e.

$$F(x) = G(x + \Delta(x)).$$

Then $\Delta(x)$ is uniquely defined as

$$\Delta(x) = G^{-1}(F(x)) - x.$$

Changing variables so $\tau = F(x)$ we have the quantile treatment effect (QTE):

$$\delta(\tau) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau).$$
The Lehmann QTE is naturally estimable by

\[ \hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau) \]

where \( \hat{G}_n \) and \( \hat{F}_m \) denote the empirical distribution functions of the treatment and control observations. Consider the quantile regression model

\[ Q_{Y_i}(\tau|D_i) = \alpha(\tau) + \delta(\tau)D_i \]

where \( D_i \) denotes the treatment indicator, and \( Y_i = h(T_i) \), e.g. \( Y_i = \log T_i \), which can be estimated by solving,

\[ \min \sum_{i=1}^{n} \rho_{\tau}(y_i - \alpha - \delta D_i) \]
Figure 3: Guinea Pig Survival Curves: Lehmann QTE of injection of a virulent form of tubercle bacilli.
Computational Considerations

The problem of estimating $\beta(\tau)$ in our basic model
\[ Q_{h(\tau)}(\tau|x) = x'\beta(\tau) \]
may be formulated as
\[ \hat{\beta}(\tau) = \arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\tau}(h(T_i) - x'_i b) \]
where $\rho_{\tau}(u) = u(\tau - I(u < 0))$, (Koenker and Bassett (1978)). This is a linear program and is a natural extension of the one-sample definition of the $\tau$th sample quantile as any minimizer of the function,
\[ R(a) = \sum_{i=1}^n \rho_{\tau}(y_i - a). \]
**Monotone Equivariance**

In the quantile regression model, for any monotone function, $h(\cdot)$, we have,

$$Q_{h(T)}(\tau|x) = h(Q_T(\tau|x)),$$

which follows immediately from observing that

$$P(T < t|x) = P(h(T) < h(t)|x).$$

This equivariance to monotone transformations of the conditional quantile functions is a crucial feature, allowing us to decouple the potentially conflicting objectives of transformations in conventional conditional mean regression.
The asymptotic behavior of the quantile regression process \( \{ \hat{\beta} : \tau \in (0, 1) \} \) closely parallels the theory of ordinary sample quantiles in the one sample problem. In the classical linear model,

\[
y_i = x_i \beta + u_i
\]

with \( u_i \) iid from \( dfF \), with density \( f(u) > 0 \) on its support \( \{u|0 < F(u) < 1\} \), the joint distribution of \( \sqrt{n}(\hat{\beta}_n(\tau_i) - \beta(\tau_i))_{i=1}^m \) is asymptotically normal with mean 0 and covariance matrix \( \Omega \otimes D^{-1} \).

Here \( \beta(\tau) = \beta + F_u^{-1}(\tau)e_1, e_1 = (1, 0, \ldots, 0)' \), \( x_{1i} \equiv 1, n^{-1} \sum x_ix'_i \to D \), a positive definite matrix, and

\[
\Omega = ((\tau_i \wedge \tau_j - \tau_i \tau_j)/(f(F^{-1}(\tau_i))f(F^{-1}(\tau_j)))_{i,j=1}^m.
\]
When the response is conditionally independent over $i$, but not identically distributed, the asymptotic covariance matrix of
\[ \zeta(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \]
is somewhat more complicated. Let $\xi_i(\tau) = x_i\beta(\tau)$, $f_i(\cdot)$ denote the corresponding conditional density, and define,

\[ J_n(\tau_1, \tau_2) = (\tau_1 \wedge \tau_2 - \tau_1 \tau_2) n^{-1} \sum_{i=1}^{n} x_i x_i', \]
\[ H_n(\tau) = n^{-1} \sum x_i x_i' f_i(\xi_i(\tau)). \]

Under mild regularity conditions on the $\{f_i\}$’s and $\{x_i\}$’s, we have joint asymptotic normality for $(\zeta(\tau_i), \ldots, \zeta(\tau_m))$ with covariance matrix

\[ V_n = (H_n(\tau_i)^{-1} J_n(\tau_i, \tau_j) H_n(\tau_j)^{-1})_{i,j=1}^{m}. \]
Medfly Quantile Regression Models

- Model A:
  - SEX Gender
  - SIZE Pupal Size in mm
  - DENSITY Density of Cage
  - %MALE Initial Proportion of Males

- Model B:
  - BATCH + Pupal Cohort
More References


Figure 4: Quantile regression results for Model A. The lightly shaded region is a 90 percent pointwise confidence band for the corresponding coefficient.
Figure 5: Quantile regression results for Model B.
Figure 6: Quantile regression results for Model B.
Figure 7: Estimated survival and hazard functions for Model A. Survival and hazard functions are illustrated males (solid line) and females (dotted line) with the other covariates evaluated at the experimental sample means. Horizontal line indicates $\hat{S}(t) = .05$. 
Figure 8: Estimated survival and hazard functions for Model B. Survival and hazard functions are illustrated males (solid line) and females (dotted line) with the other covariates evaluated at the experimental sample means. Horizontal line indicates $\hat{S}(t) = .05$. 
A Proportional Hazard Model

To contrast the quantile regression results with the Cox model we estimated the following proportional hazard (PH) model corresponding to model A.

$$\log \lambda(t|x) = \log \lambda_0(t) + 0.2165 \text{SEX} + 0.0124 \text{SIZE}$$

$$- 1.021 \text{ Density} - 2.625 \% \text{Male}.$$ (0.0072) (0.0203)

All four covariate effects constitute a rescaling of the baseline hazard function, $\lambda_0(t)$. 
Figure 9: Estimated Cox quantile treatment effect implicit in the estimated Cox model. Estimated hazard functions for males and females based on the Cox proportional hazard model.
5 Medfly Life Lessons

Males are tough ... but only until 40.
Bigger is better ... but only before 18.
Small is beautiful... after 25.
Crowds are good ... especially of guys.
Life gets safer ... but only after 60.
Quantile Regression Survival Models

- More flexible semiparametric survival models.
- Scalar treatment effects → functions.
- Natural interpretability of quantiles.
- Attractive equivariance and computation.
- Open problems: Inference, censoring, ...