Censored Quantile Regression and Survival Models

Roger Koenker

University of Illinois, Urbana-Champaign

Aarhus: 22 June 2010

Quantile Regression for Duration (Survival) Models

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

$$h(T_i) = x_i^\top \beta + u_i$$

where h is a monotone transformation, and

- T_i is an observed survival time,
- x_i is a vector of covariates,
- β is an unknown parameter vector
- $\{u_i\}$ are iid with df F.

The Cox Model

For the proportional hazard model with

$$\log \lambda(t|x) = \log \lambda_0(t) - x^\top \beta$$

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

$$\log(-\log(S(t|x))) = \log \Lambda_0(t) - x^\top \beta$$

so, evaluating at $t = T_i$, we have the model,

$$\log \Lambda_0(T) = x^\top \beta + u$$

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

The Bennett (Proportional-Odds) 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,

$$\log \Gamma(t|x) = \log \Gamma_0(t) - x^\top \beta,$$

we have, similarly,

$$\log \Gamma_0(T) = x^\top \beta + u$$

for u iid logistic with $F_0(u) = (1 + e^{-u})^{-1}$.

Accelerated Failure Time Model

In the accelerated failure time model we have

$$\log(T_i) = x_i^\top \beta + u_i$$

so

$$\begin{aligned} P(T > t) &= P(e^u > te^{-x\beta}) \\ &= 1 - F_0(te^{-x\beta}) \end{aligned}$$

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.

Beyond the Transformation Model

The common feature of all these models is that after transformation of the observed survival times we have:

- a pure location-shift, iid-error regression model
- covariate effects shift the center of the distribution of $h(T)$, but
- covariates cannot affect scale, or shape of this distribution

An Application: Longevity of Mediterrean Fruit Flies

In the early 1990's there were a series of experiments designed to study the survival distribution of lower animals. One of the most influential of these was:

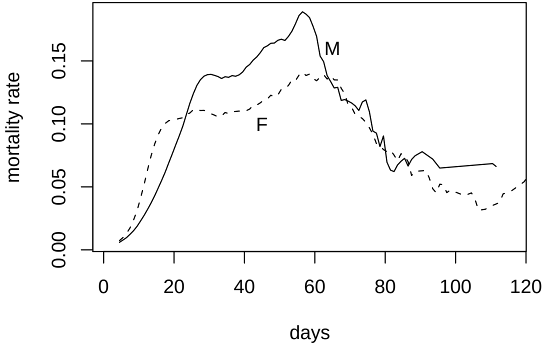
> Carey, J.R., Liedo, P., Orozco, D. and Vaupel, J.W. *(1992) Slowing of mortality rates at older ages in large Medfly cohorts, Science,* **258**, *457-61.*

- 1,203,646 medflies survival times recorded in days
- Sex was recorded on day of death
- Pupae were initially sorted into one of five size classes
- 167 aluminum mesh cages containing roughly 7200 flies
- Adults were given a diet of sugar and water *ad libitum*

Major Conclusions of the Medfly Experiment

- Mortality rates declined at the oldest observed ages. contradicting the traditional 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.
- There was strong evidence for a crossover in gender specific mortality rates.

Lifetable Hazard Estimates by Gender



Smoothed mortality rates for males and females.

Medfly Survival Prospects

Lifespan (in days)	Percentage Surviving	Number Surviving
40	5	60,000
50	1	12,000
86	.01	120
146	.001	12

Initial Population of 1,203,646

Medfly Survival Prospects

Lifespan (in days)	Percentage Surviving	Number Surviving
40	5	60,000
50	1	12,000
86	.01	120
146	.001	12

Initial Population of 1,203,646

Human Survival Prospects*

Lifespan (in years)	Percentage Surviving	Number Surviving
50	98	591,000
75	69	413,000
85	33	200,000
95	5	30,000
105	.08	526
115	.0001	1

* Estimated Thatcher (1999) Model

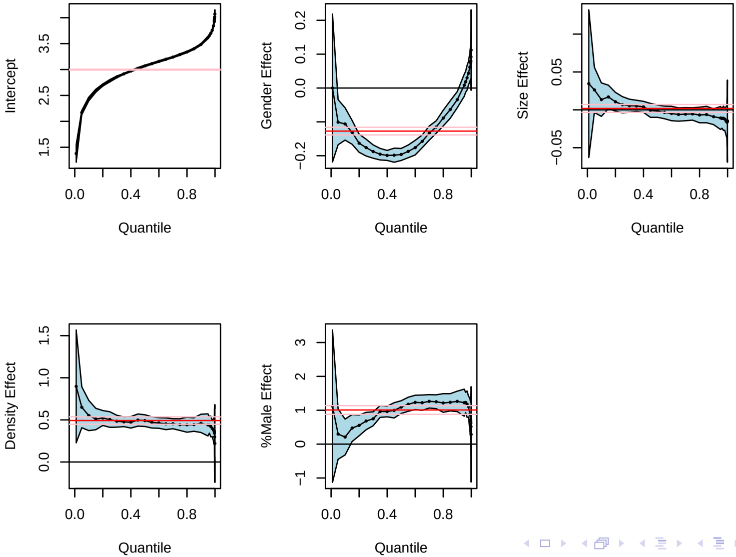
Quantile Regression Model (Geling and K (JASA,2001))

Criticism of the Carey et al paper revolved around whether declining hazard rates were a result of confounding factors of cage density and initial pupal size. Our basic QR model included the following covariates:

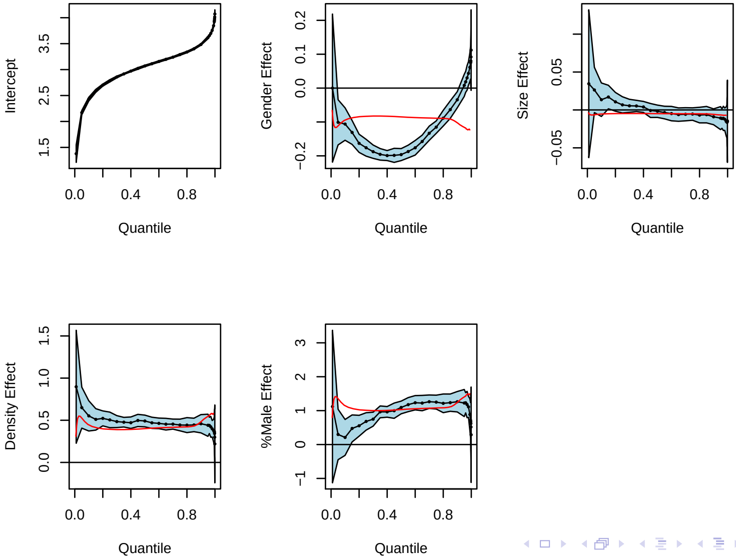
$$\begin{aligned} Q_{\log(T_i)}(\tau|x_i) &= \beta_0(\tau) + \beta_1(\tau)\text{SEX} + \beta_2(\tau)\text{SIZE} \\ &+ \beta_3(\tau)\text{DENSITY} + \beta_4(\tau)\%\text{MALE} \end{aligned}$$

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

Base Model Results with AFT Fit



Base Model Results with Cox PH Fit



What About Censoring?

There are currently 3 approaches to handling censored survival data within the quantile regression framework:

- Powell (1986) Fixed Censoring
- Portnoy (2003) Random Censoring, Kaplan-Meier Analogue
- Peng/Huang (2008) Random Censoring, Nelson-Aalen Analogue

Available for R in the package `quantreg`.

Powell's Approach for Fixed Censoring

Rationale *Quantiles are equivariant to monotone transformation:*

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

Model $Y_i = T_i \wedge C_i \equiv \min\{T_i, C_i\}$

$$Q_{Y_i|x_i}(\tau|x_i) = x_i^\top \beta(\tau) \wedge C_i$$

Data *Censoring times are known for all observations*

$$\{Y_i, C_i, x_i : i = 1, \cdots, n\}$$

Estimator *Conditional quantile functions are nonlinear in parameters:*

$$\hat{\beta}(\tau) = \textit{argmin} \sum \rho_\tau(Y_i - x_i^\top \beta \wedge C_i)$$

Portnoy's Approach for Random Censoring I

Rationale *Efron's (1967) interpretation of Kaplan-Meier as shifting mass of censored observations to the right:*

Algorithm *Until we "encounter" a censored observation KM quantiles can be computed by solving, starting at* $\tau = 0$,

$$\hat{\xi}(\tau) = \textit{argmin}_{\xi} \sum_{i=1}^{n} \rho_{\tau}(Y_i - \xi)$$

Once we "encounter" a censored observation, i.e. when $\hat{\xi}(\tau_i) = y_i$ *for some* y_i *with* $\delta_i = 0$, *we split* y_i *into two parts:*

- $y_i^{(1)} = y_i$ *with weight* $w_i = (\tau - \tau_i)/(1 - \tau_i)$
- $y_i^{(2)} = y_\infty = \infty$ *with weight* $1 - w_i$.

Then denoting the index set of censored observations "encountered" up to τ *by* $K(\tau)$ *we can solve*

$$\min \sum_{i \notin K(\tau)} \rho_{\tau}(Y_i - \xi) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_{\tau}(Y_i - \xi) + (1 - w_i(\tau))\rho_{\tau}(y_\infty - \xi)].$$

Portnoy's Approach for Random Censoring II

When we have covariates we can replace ξ by the inner product $x_i^\top \beta$ and solve:

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - x_i^\top \beta) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - x_i^\top \beta) + (1 - w_i(\tau))\rho_\tau(y_\infty - x_i^\top \beta)].$$

At each τ this is a simple, weighted linear quantile regression problem.

Portnoy's Approach for Random Censoring II

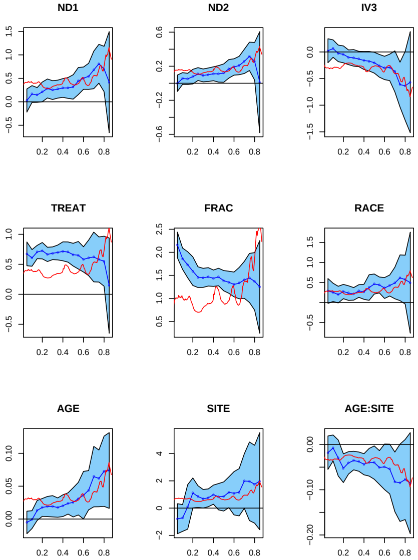
When we have covariates we can replace ξ by the inner product $x_i^\top \beta$ and solve:

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - x_i^\top \beta) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - x_i^\top \beta) + (1 - w_i(\tau))\rho_\tau(y_\infty - x_i^\top \beta)].$$

At each τ this is a simple, weighted linear quantile regression problem. The following R code fragment replicates an analysis in Portnoy (2003):

```
require(quantreg)
data(uis)
fit <- crq(Surv(log(TIME), CENSOR) ~  ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis, method = "Por")
Sfit <- summary(fit,1:19/20)
PHit <- coxph(Surv(TIME, CENSOR) ~  ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis)
plot(Sfit, CoxPHit = PHit)
```

Reanalysis of the Hosmer-Lemeshow Drug Relapse Data



Peng and Huang's Approach for Random Censoring I

Rationale *Extend the martingale representation of the Nelson-Aalen estimator of the cumulative hazard function to produce an "estimating equation" for conditional quantiles.*

Model *AFT form of the quantile regression model:*

$$Prob(\log T_i \leqslant x_i^\top \beta(\tau)) = \tau$$

Data $\{(Y_i, \delta_i) : i = 1, \cdots, n\}$ $Y_i = T_i \wedge C_i$, $\delta_i = I(T_i < C_i)$

Martingale *We have* $EM_i(t) = 0$ *for* $t \geqslant 0$, *where:*

$$\begin{aligned} M_i(t) &= N_i(t) - \Lambda_i(t \wedge Y_i | x_i) \\ N_i(t) &= I(\{Y_i \leqslant t\}, \{\delta_i = 1\}) \\ \Lambda_i(t) &= -\log(1 - F_i(t|x_i)) \\ F_i(t) &= Prob(T_i \leqslant t | x_i) \end{aligned}$$

Peng and Huang's Approach for Random Censoring II

The estimating equation becomes,

$$En^{-1/2} \sum x_i [N_i(\exp(x_i^\top \beta(\tau))) - \int_0^\tau I(Y_i \geqslant \exp(x_i^\top \beta(u))) dH(u) = 0.$$

where $H(u) = -\log(1-u)$ for $u \in [0, 1)$, after rewriting:

$$\begin{aligned} \Lambda_i(\exp(x_i^\top \beta(\tau)) \wedge Y_i | x_i)) &= H(\tau) \wedge H(F_i(Y_i | x_i)) \\ &= \int_0^\tau I(Y_i \geqslant \exp(x_i^\top \beta(u))) dH(u), \end{aligned}$$

Approximating the integral on a grid, $0 = \tau_0 < \tau_1 < \cdots < \tau_J < 1$ yields a simple linear programming formulation to be solved at the gridpoints.

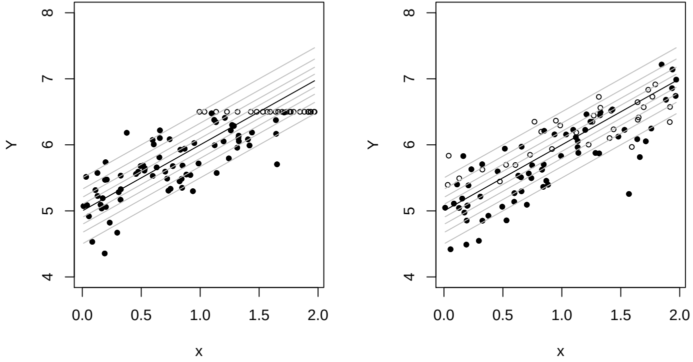
Alice in Asymptopia

It might be thought that the Powell estimator would be more efficient than the Portnoy and Peng-Huang estimators given that it imposes more stringent data requirements. Comparing asymptotic behavior and finite sample performance in the simplest one-sample setting indicates otherwise.

	median	Kaplan-Meier	Nelson-Aalen	Powell	Leurgans $\hat{G}$	Leurgans G
n= 50	1.602	1.972	2.040	2.037	2.234	2.945
n= 200	1.581	1.924	1.930	2.110	2.136	2.507
n= 500	1.666	2.016	2.023	2.187	2.215	2.742
n= 1000	1.556	1.813	1.816	2.001	2.018	2.569
n= ∞	1.571	1.839	1.839	2.017	2.017	2.463

Scaled MSE for Several Estimators of the Median: Mean squared error estimates are scaled by sample size to conform to asymptotic variance computations. Here, T_i is standard lognormal, and C_i is exponential with rate parameter .25, so the proportion of censored observations is roughly 30 percent. 1000 replications.

Simulation Settings I



Simulations I-A

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
n = 100	-0.0032	0.0638	0.0988	0.0025	0.0702	0.1063
n = 400	-0.0066	0.0406	0.0578	0.0036	0.0391	0.0588
n = 1000	-0.0022	0.0219	0.0321	0.0006	0.0228	0.0344
Peng-Huang						
n = 100	0.0005	0.0631	0.0986	0.0092	0.0727	0.1073
n = 400	-0.0007	0.0393	0.0575	0.0074	0.0389	0.0598
n = 1000	0.0014	0.0215	0.0324	0.0019	0.0226	0.0347
Powell						
n = 100	-0.0014	0.0694	0.1039	0.0068	0.0827	0.1252
n = 400	-0.0066	0.0429	0.0622	0.0098	0.0475	0.0734
n = 1000	-0.0008	0.0224	0.0339	0.0013	0.0264	0.0396
GMLE						
n = 100	0.0013	0.0528	0.0784	-0.0001	0.0517	0.0780
n = 400	-0.0039	0.0307	0.0442	0.0031	0.0264	0.0417
n = 1000	0.0003	0.0172	0.0248	-0.0001	0.0165	0.0242

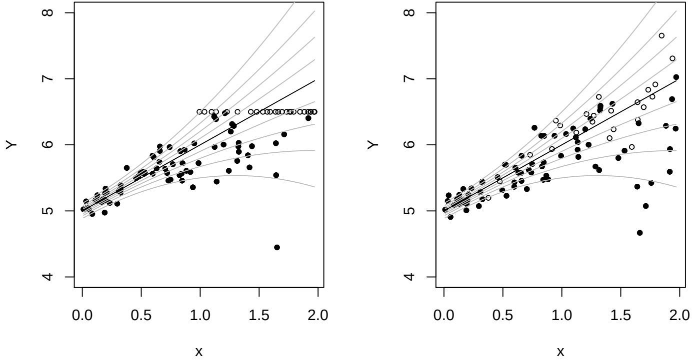
Comparison of Performance for the iid Error, Constant Censoring Configuration

Simulations I-B

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
n = 100	-0.0042	0.0646	0.0942	0.0024	0.0586	0.0874
n = 400	-0.0025	0.0373	0.0542	-0.0009	0.0322	0.0471
n = 1000	-0.0025	0.0208	0.0311	0.0006	0.0191	0.0283
Peng-Huang						
n = 100	0.0026	0.0639	0.0944	0.0045	0.0607	0.0888
n = 400	0.0056	0.0389	0.0547	-0.0002	0.0320	0.0476
n = 1000	0.0019	0.0212	0.0311	0.0009	0.0187	0.0283
Powell						
n = 100	-0.0025	0.0669	0.1017	0.0083	0.0656	0.1012
n = 400	0.0014	0.0398	0.0581	-0.0006	0.0364	0.0531
n = 1000	-0.0013	0.0210	0.0319	0.0016	0.0203	0.0304
GMLE						
n = 100	0.0007	0.0540	0.0781	0.0009	0.0470	0.0721
n = 400	0.0008	0.0285	0.0444	-0.0008	0.0253	0.0383
n = 1000	-0.0004	0.0169	0.0248	0.0002	0.0150	0.0224

Comparison of Performance for the iid Error, Variable Censoring Configuration

Simulation Settings II



Simulations II-A

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
n = 100	0.0084	0.0316	0.0396	-0.0251	0.0763	0.0964
n = 400	0.0076	0.0194	0.0243	-0.0247	0.0429	0.0533
n = 1000	0.0081	0.0121	0.0149	-0.0241	0.0309	0.0376
Portnoy Q						
n = 100	0.0018	0.0418	0.0527	0.0144	0.1576	0.2093
n = 400	-0.0010	0.0228	0.0290	0.0047	0.0708	0.0909
n = 1000	-0.0006	0.0122	0.0154	-0.0027	0.0463	0.0587
Peng-Huang L						
n = 100	0.0077	0.0313	0.0392	-0.0145	0.0749	0.0949
n = 400	0.0064	0.0193	0.0240	-0.0125	0.0392	0.0493
n = 1000	0.0077	0.0120	0.0147	-0.0181	0.0279	0.0342
Peng-Huang Q						
n = 100	0.0078	0.0425	0.0538	0.0483	0.1707	0.2328
n = 400	0.0035	0.0228	0.0291	0.0302	0.0775	0.1008
n = 1000	0.0015	0.0123	0.0155	0.0101	0.0483	0.0611
Powell						
n = 100	0.0021	0.0304	0.0385	-0.0034	0.0790	0.0993
n = 400	-0.0017	0.0191	0.0239	0.0028	0.0431	0.0544
n = 1000	-0.0001	0.0099	0.0125	0.0003	0.0257	0.0316
GMLE						
n = 100	0.1080	0.1082	0.1201	-0.2040	0.2042	0.2210
n = 400	0.1209	0.1209	0.1241	-0.2134	0.2134	0.2173
n = 1000	0.1118	0.1118	0.1130	-0.2075	0.2075	0.2091

Comparison of Performance for the Constant Censoring, Heteroscedastic Configuration

Simulations II-B

	Intercept			Slope		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
n = 100	0.0024	0.0278	0.0417	-0.0067	0.0690	0.1007
n = 400	0.0019	0.0145	0.0213	-0.0080	0.0333	0.0493
n = 1000	0.0016	0.0097	0.0139	-0.0062	0.0210	0.0312
Portnoy Q						
n = 100	0.0011	0.0352	0.0540	0.0094	0.1121	0.1902
n = 400	0.0002	0.0185	0.0270	-0.0012	0.0510	0.0774
n = 1000	-0.0005	0.0116	0.0169	-0.0011	0.0337	0.0511
Peng-Huang L						
n = 100	0.0018	0.0281	0.0417	0.0041	0.0694	0.1017
n = 400	0.0013	0.0142	0.0212	0.0035	0.0333	0.0490
n = 1000	0.0012	0.0096	0.0139	0.0002	0.0208	0.0310
Peng-Huang Q						
n = 100	0.0044	0.0364	0.0550	0.0322	0.1183	0.2105
n = 400	0.0026	0.0188	0.0275	0.0154	0.0504	0.0813
n = 1000	0.0007	0.0113	0.0169	0.0077	0.0333	0.0520
Powell						
n = 100	-0.0001	0.0288	0.0430	0.0055	0.0733	0.1105
n = 400	0.0000	0.0147	0.0226	0.0001	0.0379	0.0561
n = 1000	-0.0008	0.0095	0.0146	0.0013	0.0237	0.0350
GMLE						
n = 100	0.1078	0.1038	0.1272	-0.1576	0.1582	0.1862
n = 400	0.1123	0.1116	0.1168	-0.1581	0.1578	0.1647
n = 1000	0.1153	0.1138	0.1174	-0.1609	0.1601	0.1639

Comparison of Performance for the Variable Censoring, Heteroscedastic Configuration

Conclusions

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are quite similar.

Conclusions

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are quite similar.
- The martingale representation of the Peng-Huang estimator yields a more complete asymptotic theory than is currently available for the Portnoy estimator.

Conclusions

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are quite similar.
- The martingale representation of the Peng-Huang estimator yields a more complete asymptotic theory than is currently available for the Portnoy estimator.
- The Powell estimator, although conceptually attractive, suffers from some serious computational difficulties, imposes strong data requirements, and has an inherent asymptotic efficiency disadvantage.

Conclusions

- Simulation evidence confirms the asymptotic conclusion that the Portnoy and Peng-Huang estimators are quite similar.
- The martingale representation of the Peng-Huang estimator yields a more complete asymptotic theory than is currently available for the Portnoy estimator.
- The Powell estimator, although conceptually attractive, suffers from some serious computational difficulties, imposes strong data requirements, and has an inherent asymptotic efficiency disadvantage.
- Quantile regression provides a flexible complement to classical survival analysis methods, and is now well equipped to handle censoring.