IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 215–223
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL721

Is ignorance bliss: Fixed vs. random censoring

Stephen Portnoy[*]

Department of Statistics, University of Illinois

Abstract:
While censored data is sufficiently common to have generated an enormous field of applied statistical research, the basic model for such data is also sufficiently non-standard to provide ample surprises to the statistical theorist, especially one who is too quick to assume regularity conditions. Here we show that estimators of the survival quantile function based on assuming additional information about the censoring distribution behave more poorly than estimators (like the inverse of Kaplan–Meier) that discard this information. This phenomenon will be explored with special emphasis on the Powell estimator, which assumes that all censoring times are observed.

1. Introduction

In many situations where censored observations appear, it is not unreasonable unreasonable to assume that the censoring values are known for all observations (even the uncensored ones). For example, one of the earliest approaches to censored regression quantiles was introduced by work of Powell in the mid 1980's. Powell [10] assumed that the censoring values were constant, thus positing observations of the form $Y^* = \min(Y, c)$ (where Y^* is observed and Y is the possibly unobserved survival time that is assumed to obey some linear model). More generally, we may be willing to assume that we observe a sample of censoring times $\{c_i\}$ and a sample of censored responses $Y_i^* = \min(Y_i, c_i)$, a model that could apply to a single sample. In this case, one could use the empirical distributions of the $\{Y_i^*\}$ and $\{c_i\}$ and take the ratio of empirical survival functions to estimate the survival function of Y. This is asymptotically equivalent to applying the Powell method on a single sample.

Despite some optimality claims in Newey and Powell [5], it turns out that the Kaplan–Meier estimate is better (asymptotically, and by simulations in finite samples) even though it does not use the full sample of $\{c_i\}$ values. More generally, even in multiple regression settings, the censored regression quantile estimators (Portnoy [7]) are better in simulations than Powell's estimator, even for the constant censoring situation for which Powell's estimator was developed. Remarkably, in the one sample case, replacing the empirical function of $\{c_i\}$ by the true survival function (assuming it is known) yields an even less efficient estimator. Thus, it appears that discarding what appears to be pertinent information improves the estimators. Here we will try to quantify and explain this conundrum.

Department of Statistics, University of Illinois, 725 S. Wright St., Champaign IL 61801, U. S. A. e-mail: sportnoy@illinois.edu

[*]Research partially supported by NSF Award DMS-0604229

AMS 2000 subject classifications: Primary 62N02, 62J05; secondary 62B10

Keywords and phrases: regression quantiles, conditional quantile, Powell estimator.

The basic phenomenon was first brought to my attention by Roger Koenker at coffee some time ago. A specific example concerned asymptotics for estimators of the survival quantile function in a single censored sample. Though the basic asymptotic results all appeared in the standard literature, the computations were combined as problems on a take-home exam for advanced econometric students to emphasize the rather surprising result that the more information an estimator incorporates, the poorer the asymptotic behavior. In fact, a recent treatment closely related to the one sample case of Section 2 below appears in Wang and Li [11].

Specifically, consider the one-sample model: Y_i are i.i.d. with cdf $F(x)$, c_i i.i.d. with cdf $G(x)$, and we observe $Y_i^* = \min\{Y_i, c_i\}$. The problem is to estimate $Q(\tau) \equiv F^{-1}(\tau)$ (nonparametrically). There are a number of estimators that converge in distribution at rate $n^{-1/2}$ to an asymptotic normal distribution with mean $Q(\tau)$ and asymptotic variance:

$$(1) \qquad a\,Var \equiv \frac{F(\xi_\tau)(1-F(\xi_\tau))}{f^2(\xi_\tau)} v(\xi_\tau) \qquad \xi_\tau = Q(\tau) = F^{-1}(\tau),$$

where $f(x)$ is the density for $F(x)$, and where $v(x)$ depends on the estimator.

The most classical estimator of the quantile function is the inverse of the Kaplan–Meier estimator. This inversion is trivial since the Kaplan–Meier estimator is monotonic; and its asymptotic variance is well-known to have

$$(2) \qquad v_{KM}(x) = \frac{(1-F(x))}{F(x)} \int_0^x \frac{\mathrm{d}F(w)}{(1-F(w))^2(1-G(w))}.$$

There are (at least) two alternatives that are appropriate if all the c_i-values are observed. The first is of particular interest here and was developed by Powell [10]: define the quantile function estimate, $\hat{Q}(\tau)$ to minimize the following (non-convex) objective function over β:

$$(3) \qquad \sum_{i=1}^{n} \rho_\tau(\min\{Y_i^*, c_i\} - \min\{\beta, c_i\}).$$

This was originally introduced for fixed (constant) censoring in linear models for the conditional median, but it was quickly recognized that the definition worked whenever all c_i-values were known. The asymptotic variance of $\hat{Q}(\tau)$ is given by (1) with

$$(4) \qquad v_{POW}(x) = (1-G(x))^{-1}.$$

An alternative with exactly the same asymptotic variance is the "synthetic" estimator (for example, see Leurgans [3]). Note that the c.d.f. for the observed value $\min\{Y_i, c_i\}$ is $H(x) = 1-(1-F(x))(1-G(x))$. Thus, if all $\{c_i\}$ are known, we can use the empiric c.d.f.'s ($\hat{H}$ for the observations, and $\hat{G}$ for the censoring times) to estimate F:

$$(5) \qquad \hat{F}(x) = 1-(1-\hat{H}(x))/(1-\hat{G}(x)).$$

This function can be inverted (with perhaps some difficulty because of possible non-monotonicity) to provide an estimate of the quantile function, whose asymptotic variance can be shown directly to coincide with that of the Powell estimator. To provide complete notation, define $v_{\hat{G}}(x) \equiv v_{POW}(x)$.

Finally, suppose we actually know G; that is, we have additional information. Then we can replace $\hat{G}$ by G in (5) and invert. Here

$$v_G(x) = \frac{1-(1-F(x))(1-G(x))}{F(x)(1-G(x))}.$$

For estimation of the median, $\xi = Q(1/2) = F^{-1}(1/2)$,

$$\begin{aligned} v_{KM} &= \int_0^{\xi} \frac{\mathrm{d}F(w)}{(1-F(w))^2(1-G(w))} \\ v_{POW} = v_{\hat{G}} &= (1-G(\xi))^{-1} \\ v_G &= (1+G(\xi))/(1-G(\xi)) \end{aligned}$$

For $\tau = 1/2$, it is immediate that:

$$v_{KM}(\xi_\tau) \le v_{Pow}(\xi_\tau) = v_{\hat{G}}(\xi_\tau) \le v_G(\xi_\tau).$$

These inequalities hold for all τ: $v_{\hat{G}}(\xi_\tau) \le v_G(\xi_\tau)$ since $(1-G(x)) \le 1$ in the numerator of $v_G(x)$. To show $v_{KM}(\xi_\tau) \le v_{\hat{G}}(\xi_\tau)$, note that $(1-G(x)) \ge (1-G(w))$ in the denominator of the integral in (2), and the integral can be computed directly to provide a cancellation of the initial factors.

To provide some specific calculations where the integral in v_{KM} can be computed, let $1-F(x) = e^{-x}$ and let G have density $g(x) = c\,e^{-c(x-a)}$ for $x \ge a$. Figure 1 shows efficiencies for median estimators with respect to the asymptotic variance of the Kaplan–Meier estimator for $a = 1.8$ as a function of c. The unobservable estimate, $med\{Y\}$, is also plotted for comparison. Note that it is only slightly more efficient than Kaplan–Meier.

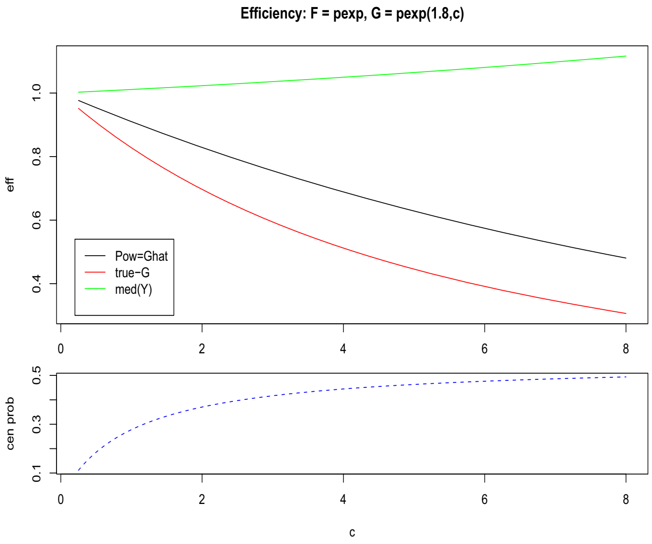

FIG 1. *One Sample Efficiencies for exponential distributions.*

Several remarks can be offered.

- Newey and Powell [5] establish asymptotic optimality of the Powell estimator among all "regular" estimators. Unfortunately, their regularity conditions preclude estimators like the Kaplan–Meier estimator that does not admit an asymptotic expansion whose second-order term is independent of the first-order term. Thus, the fact that the Powell estimator performs more poorly than the Kaplan–Meier estimator does not contradict their result.
- By convex optimization (specifically, the Generalized Method of Moment Spaces – see Collins and Portnoy [1]), it is possible to find the range of values for the efficiency of the Powell estimator for any given amount of censoring. Specifically, if p is the probability of censoring, then the efficiency of the Powell estimator is greater than $(1-2p)$, and this efficiency can be attained. This bound is plotted in the lower panel of Figure 1.
- When a is nearly log(2) (the median for the negative exponential distribution), the probability of censoring is nearly .5, and serious computational difficulties can occur in samples of size 50 or bigger for Powell and the "synthetic" estimators. For the Powell estimator, the problem seems to be the multimodality of the objective function, an issue that will be discussed for regression quantiles later. For synthetic estimators, the ratio of survival functions is *not* monotonic, which may lead to computational problems in the inversion.

2. Regression comparisons

Since the Powell estimator was intended for the case of linear regression quantiles, the results of the previous section may not seem unduly surprising. Nonetheless, we show here that the message is remarkably similar in the regression case. Unfortunately, current asymptotic theory for quantile regression estimator that require only conditional independence of the duration and censoring variables do not admit tractable formulas for asymptotic variances (see Portnoy and Lin [?], and Peng–Huang [6]). Thus we will restrict to simulation comparisons. Since the methods of Portnoy [7] and Peng–Huang [6] appear to be quite similar, we will also focus on comparisons between the Powell method and the CRQ ("censored regression quantiles") method of Portnoy [7].

It is important to note that CRQ requires all estimable regression quantiles to be linear (in the parameters). The Powell estimator does not impose this requirement, positing a linear model *only* for the quantile of interest. Thus will we consider cases where the conditional quantiles are not linear, expecting that the Powell estimator should do better in such cases. Thus, it is surprising that even for moderately large samples ($n = 400$), the CRQ method still outperforms the Powell estimator, often quite substantially. This appears to be due to computational difficulties associated with fact that the Powell method is fitting a nonlinear response function ($\max\{x_i'\beta, c_i\}$), and so its objective function turns out to be multimodal.

Because of the computational difficulties involved in minimizing the Powell objective function, we will restrict to cases with a single explanatory variable, for which the Powell estimator can be computed by exhaustive search. Specifically, the Powell estimator can be taken to be an "elemental" estimator; that is, one interpolating exactly p observations (at least when observations are in general position). This holds for the same reason as for ordinary regression quantiles: if an optimal solution is not elemental, the linear parameters can be changed without increasing the objective function until p observations are interpolated. Thus, for simple linear regression, we will employ an algorithm that exhaustively examines all "n-choose-2" elemental solutions and finds the one minimizing the Powell objective function

$$\sum_{i=1}^{n} \rho_\tau(\min\{Y_i^*, c_i\} - \min\{x_i'\beta, c_i\}). \tag{6}$$

While there are approximate algorithms that are much faster (especially in larger problems), these methods depend strongly on a "starting value", and will have rather different distributional properties (depending on the starting value). The Powell estimator will be compared with results from the R-function `crq` using the default "grid" algorithm of Portnoy [7] as implemented in the `quantreg` R-package (Koenker [2]).

To be specific, we consider the following design for a simulation experiment with three models (two of which are heteroscedastic and nonlinear), two error distributions, and three choices for sample size (n = 50, 100, 200). In each case, we take 1000 replications with the pairs (x_i, Y_i) i.i.d., and take constant censoring with $c = 10$. For all cases, we resample $x_i \sim Unif(0, 4)$ in each replication. The Models are:

Linear: $Y = 5 + 2x + \varepsilon$
Nonlinear: $Y = 2.5x + \max(x, 2)\,\varepsilon$
Heavy Nonlin: $Y = x + 4\max(x, 2)\,\varepsilon$

The error distributions are either $\varepsilon \sim N(0, 1)$ or ε has a location shift of a negative exponential distribution with density $f(x) = \exp\{-(x + a)\}$ where $a \approx -.69$ is chosen to provide $\text{med}(\varepsilon) = 0$.

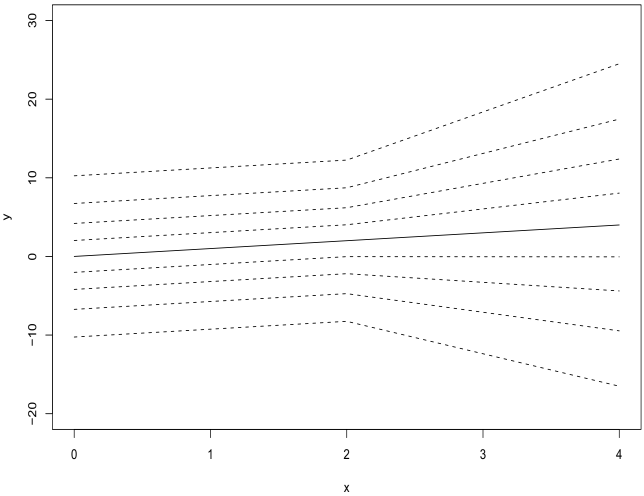

Fig 2. *Deciles for heavy nonlinear model (Normal errors).*

Note that only the first model has all linear regression quantiles; and thus CRQ would be consistent only for this model. The conditional median is linear in all three models; and so the Powell estimator should be consistent in all cases. A plot of the conditional quantiles for the case of "heavy" non-linearity is given in Figure 2. Figures 3 and 4 provide the results of the simulations expressed as the ratios of the median absolute errors for the CRQ estimator over the Powell estimator. Ratios of mean squared errors showed much less efficiency for the Powell estimator.

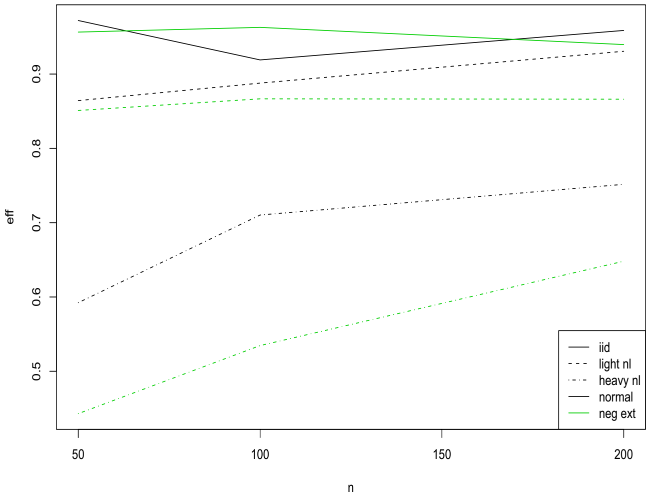

Fig 3. *Efficiency for intercept: MAE(CRQ)/MAE(Powell).*

The following conclusions seem quite clear from the plots:

- The one-sample story appears to hold for regression.
- Heavy nonlinearity hurts Powell (computational problems) more than CRQ (bias) for $n \leq 200$. For larger n, the bias may become more serious. Even if we believe only the median is linear, CRQ seems to be better for moderate nonlinearity and sample size.

One possible reason that CRQ seems so good concerns the fact that the CRQ estimator weights each censored observation depending on the quantile crossing the observation. Since each weight applies only to a small number of observations, the accuracy in estimating the weights may not be very crucial. Also, most censoring occurs near the median, where nonlinearity is smaller.

Some further complementary simulation experiments were run. One used the approximate algorithm for the Powell estimator given in the "quantreg" R-Package (see Koenker [2]). This algorithm is based on work of Fitzenberger and attempts to find a local minimum of the Powell objective function (with the starting value defaulting to the naive regression quantile estimator that ignores censoring). This does correct the worst problems with the Powell estimator in the case of heavy censoring; and in fact this version of the Powell estimator slightly outperforms CRQ for estimating the slope parameter when n = 50. In all other cases, even this version is less efficient than CRQ with efficiencies varying from .6 to .95 over the range of cases in the simulation experiment above. Since this algorithm can

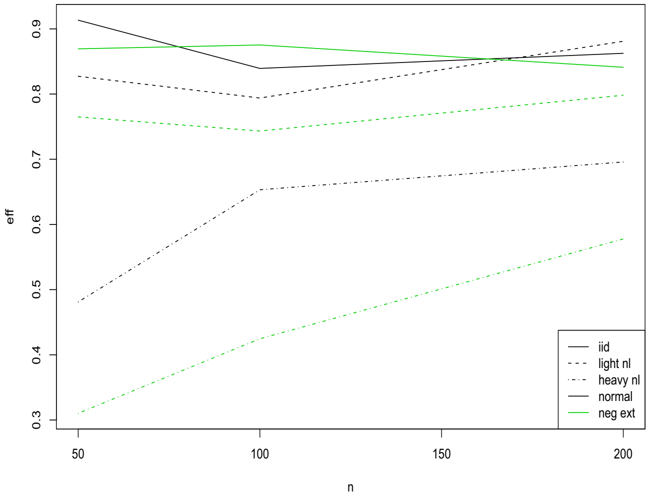

Fig 4. *Efficiency for slope: MAE(CRQ)/MAE(Powell).*

differ from the formal Powell estimator, it is not clear what asymptotic properties it has. Nonetheless, the simulations suggest that even this version is not preferable to CRQ.

Finally, a simulation experiment was run with an alternative estimator suggested by work in Lindgren [4]. This method is based on binning the data (by x-values) into M bins, applying Kaplan–Meier to the data in each bin, and fitting the resulting quantiles by linear least squares. Here we choose $M = 8$ bins equally spaced for $x \in (0, 4)$. Such proposals appear regularly in the literature, but binning difficulties (the curse of dimensionality) seriously degrade such methods beyond the case of simple linear regression. In any event, this estimator performed only slightly better than the Powell estimator, and clearly suffered in comparison with CRQ.

3. Inconsistency of the Powell estimator

As noted above, if only the quantile of interest is linear, the Powell estimator can remain consistent while CRQ is inconsistent. However, the conditions for consistency for these estimators differ in nontrivial ways. The author has obtained several examples where the Powell estimator is inconsistent while CRQ remains consistent (Portnoy [8]). The basic idea is that the use of a nonlinear fit in the Powell estimator permits breakdown in cases where standard regression quantile methods (RQ and CRQ) maintain breakdown robustness. In fact, CRQ can be consistent even though some lower conditional quantiles are nonlinear: specifically, when the lower quantiles are below all censored observations. The examples do appear to violate conditions for known consistency results, and so do not suggest any error in the proof of consistency for the Powell estimator. They do emphasize that the nonlinear nature of the Powell objective function does impose additional regularity conditions.

Though the examples of inconsistency are somewhat pathological, they do suggest cases where fitting a nonlinear response function (*viz.*, the Powell estimator) leads to a (very) incorrect estimate of the true regression line. In fact, the following finite sample simulated example shows that Powell's estimator may be extremely poor even though the data do not appear unreasonable and the CRQ estimates appear quite reasonable.

Specifically, we consider an example where $x \sim$ Unif(0, 4) and $Y_0 \sim 5 + x + 4\max\{x, 2\}\, N(0,1)$. Here censoring is at the constant value, $c = 10$, and so we observe $Y = \min(Y_0, 10)$. The specific data may be generated in R as follows:

```
  # generate powell-crq examples
set.seed(23894291)
for(i in 1:92) {
x <- 4*runif(50)
y0 <- 5 + x + 4*pmax(x,2)*rnorm(50) }
y <- pmin(y0,10)
```

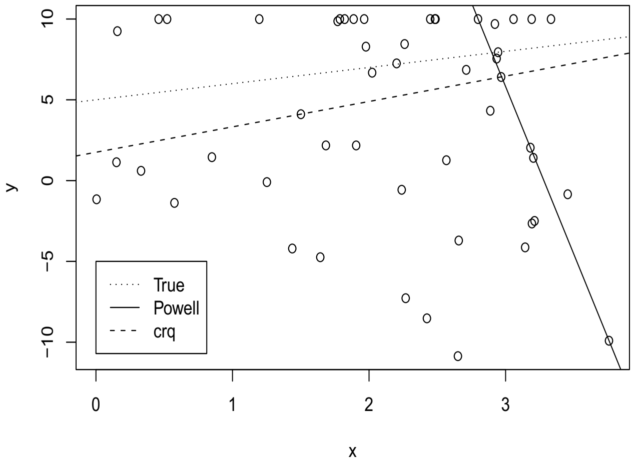

FIG 5. *Example where the Powell estimator is very poor.*

The Powell estimator (with exhaustive search) gives intercept and slope estimates: (68.10, -20.76); while the CRQ method gives: (1.754, 1.571) The data is plotted in Figure 5, and though there is an indication of heteroscedasticity, the data do not appear to contain outliers or other discrepancies that should lead to undue difficulties. Careful examination of the Powell objective function shows several local minima. One local minimum is indeed near the CRQ estimates, but the value at this local minimum is in fact somewhat larger than that at the global minimum, and is only slightly smaller than one local minimum located rather far from either Powell or CRQ.

4. Conclusions

- Beware of stating the results of theorems without stating the conditions. The hypotheses in the Newey-Powell optimality result restrict consideration to a class that ignores the most natural alternatives.

- Is ignorance bliss? No! But seeing superfluous information can encourage one to try to make use of the information in ways that may be detrimental!
- Be careful of using procedures whose computation may be problematic, especially those defined by minimization of a multimodal objective function. Even if such procedures have provable asymptotic properties, finite sample computational difficulties can result in extremely poor performance.

References

[1] Collins, J. and Portnoy, S. (1981). Maximizing the variance of M-estimators using the generalized method of moment spaces. *Ann. Statist.* **9** 567–577.

[2] Koenker, R. (2008). quantreg: Quantile Regression. R package version 4.23. `http://www.r-project.org`.

[3] Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74** 301–309.

[4] Lindgren, A. (1997). Quantile regression with censored data using generalized L_1 minimization. *Comp. Statist. Data Anal.* **23** 509–524.

[5] Newey, W. and Powell, J. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* **6** 295–317.

[6] Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *J. Amer. Statist. Assoc.* **103** 637–649.

[7] Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.* **98** 1001–1012.

[8] Portnoy, S. (2009). Inconsistency of the Powell estimator: examples. Preprint, Department of Statistics, University of Illinois.

[9] Portnoy, S. and Lin, G. (2010) Asymptotics for censored regression quantiles. *J. Nonparametric Statistics* **22** 115–130.

[10] Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32** 143–155.

[11] Wang, J. and Li, Y. (2005). Estimators for the survival function when censoring times are known. *Comm. Statist.: Theory and Methods* **34** 449–459.

[12] Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *J. Amer. Statist. Assoc.* **104** 1117–1128.