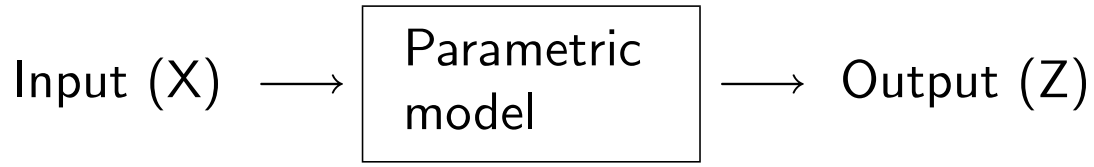
Robust Designs for Nonlinear Quantile Regression

Linglong Kong and Doug Wiens

U. Alberta, Edmonton

- Model Robustness



  - Choose design points x_i at which to observe Z; aim for accuracy (small biases) and efficiency (small variance).

  - The design is tailored to a particular fitted model, e.g. $Z = \theta' \boldsymbol{x} +$ random error.

  - The 'best' design for a slightly wrong model can be much more than slightly sub-optimal. (Box and Draper 1959 etc.)

  - Although we will fit the assumed model, we should design for protection against biases arising from any of a range of nearby models. ('All models are wrong ... ', G. Box.)

- Quantile Regression
  - The solution $q(\boldsymbol{x}) = \theta'\boldsymbol{x}$ to $P\left(Z_{|\boldsymbol{x}} \leq q(\boldsymbol{x})\right) = \tau$ is the τ-regression quantile:

$$q(\boldsymbol{x}) = G^{-1}_{Z_{|\boldsymbol{x}}}(\tau).$$

  - Kong and Wiens ('KW', JASA 2015) obtained designs for 'quantile regression' in 'approximate linear models'; with some modifications our results apply to nonlinear problems.
  - Quantile regression is resistant to y-outliers, not to x-outliers (not a problem with designed studies).
  - More efficient than LSE under non-normal distributions; no moment assumptions made (e.g. Cauchy errors are possible).
  - Provides a satisfying picture of the manner in which the response is affected by the covariates.

- Assume that the τ-quantile of the output Z at input $\boldsymbol{x}$ is a possibly non-linear function $F(\boldsymbol{x};\boldsymbol{\beta})$:

$$\tau = P_{Z_{|\boldsymbol{x}}}\left(Z \leq F(\boldsymbol{x};\boldsymbol{\beta})\right).$$

- The asymptotic properties of the estimates in this nonlinear framework are the same as for the linear model obtained by taking a first order expansion around a local parameter (initial or prior estimate - Phase 2 drug trials?) $\boldsymbol{\beta}_0$:

$$Y = Z - F_0(\boldsymbol{x}) \approx \boldsymbol{f}'_0(\boldsymbol{x})\,\boldsymbol{\theta} + \text{ error},$$

where $F_0(\boldsymbol{x}) = F(\boldsymbol{x};\boldsymbol{\beta}_0)$, $\boldsymbol{f}_0(\boldsymbol{x}) = \frac{\partial}{\partial\boldsymbol{\beta}}F(\boldsymbol{x};\boldsymbol{\beta})_{|\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ and $\boldsymbol{\theta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$.

  - Now the mathematics is done in terms of $(Y, \boldsymbol{x})$ and $\boldsymbol{\theta}$; for this the results in KW apply exactly.

- **Example** Dette and Trampisch ((‘DT’) JASA 2012), report an experiment carried out by Cressie and Keightley (‘CK’) 1979):
  Response Z = amount of estrogen bound to a receptor, x = amount of hormone; Michaelis-Menten response:

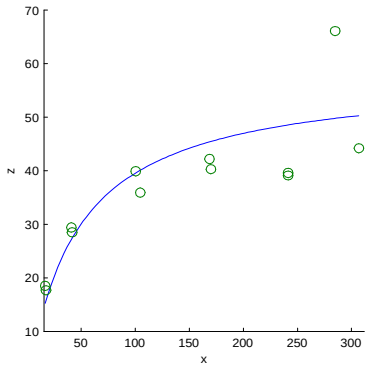
$$z = F(x; \boldsymbol{\beta}) = \frac{\beta_1 x}{\beta_2 + x}. \tag{1}$$

  CK employ a linearizing transformation introduced by Scratchard (1948):

$$\frac{z}{x} = \frac{\beta_1}{\beta_2} - \frac{z}{\beta_2}$$

  and estimate the slope and intercept by least squares, thus obtaining $\boldsymbol{\beta}_0 = (57.98, 46.43)'$. For estimation of the conditional median ($\tau = .5$) in the presence of a symmetric error distribution, $\boldsymbol{\beta}_0$ should provide a reasonable initial estimate.

  – None of the designs compared here depends (in an essential manner) on τ.



Data gathered by Cressie and Keightley (1979) with least squares response curve $F(x;\boldsymbol{\beta}_0)$ obtained via the Scratchard linearization.

- The poor fit suggests a need for robustness of some form.

- Linear approximation is $(Z - F_0(\boldsymbol{x}) =)\ Y = \boldsymbol{f}_0'(\boldsymbol{x})\,\boldsymbol{\theta} +$ random error for

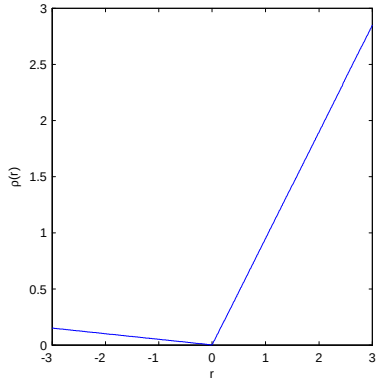
$$\boldsymbol{f}_0'(x) = \left(\frac{x}{\beta_2 + x}, -\frac{\beta_1 x}{(\beta_2 + x)^2}\right)_{|\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

  As 'design space' we take a grid χ of $N = 100$ equally spaced points spanning $[1, 400]$; we will choose $n = 20$, not necessarily distinct, points $x \in \chi$, at which to observe Y.

- The experimenter, acting as though the model is correct and the errors are homoscedastic, computes the quantile regression estimate

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{t}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{f}_0'\left(\boldsymbol{x}_i\right) \boldsymbol{t} \right),$$

where $\rho_\tau(\cdot)$ is the 'check' function $\rho_\tau(r) = r(\tau - I(r < 0))$.



Check function $\rho_\tau(r) = r(\tau - I(r < 0))$; $\tau = .95$.

- The $(Y, \boldsymbol{x}, \boldsymbol{\theta})$ formulation is only an approximation, partly because of the linearizing, and also possibly because the original $(Z, \boldsymbol{x}, \boldsymbol{\beta})$ model (1) may itself have been misspecified, either with respect to the local parameter, or the functional form of the assumed Michaelis-Menten response $F(\boldsymbol{x}; \boldsymbol{\beta})$. We suppose that in fact the model is

$$Y = \boldsymbol{f}_0'(\boldsymbol{x})\,\boldsymbol{\theta} + \delta_n(\boldsymbol{x}) + \sigma(\boldsymbol{x})\,\varepsilon, \tag{2}$$

for some 'small' model error δ_n. We define the 'true' parameter by

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{t}} \frac{1}{N} \sum_{i=1}^{N} E_{Y|\boldsymbol{x}} \left[\rho_\tau \left(Y - \boldsymbol{f}_0'(\boldsymbol{x}_i)\,\boldsymbol{t} \right) \right]; \tag{3}$$

carrying out this minimization and taking a first order approximation results in the orthogonality of the 'model residuals' $\delta_n(\boldsymbol{x}_i)$ and the regressors:

$$\left(n^{-1/2} g_\varepsilon(0) + O(1) \right) \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{f}_0(\boldsymbol{x}_i) \sqrt{n} \delta_n(\boldsymbol{x}_i) = \boldsymbol{0}. \tag{4}$$

- We seek designs for (2) which are robust against increased mean squared errors of the predicted conditional quantiles $\hat{Y}_\tau = \boldsymbol{f}_0'(\boldsymbol{x})\,\hat{\boldsymbol{\theta}}_\tau$:

$$\begin{aligned} MSE &= E\left[\{\text{predicted value} - \text{true value}\}^2\right] \\ &= E\left[\left\{\hat{Y}_\tau(\boldsymbol{x}_i) - Y_\tau(\boldsymbol{x}_i)\right\}^2\right] \end{aligned}$$

  engendered by δ_n or by nonconstant $\sigma(\cdot)$.

- For the asymptotics, the effect of δ_n must drop at the same rate as standard error (*Reason*: mse = s.e.^{2}+ bias^2), and so we assume the existence of a bounded limit:

$$\delta_0(\boldsymbol{x}) = \lim_{n\to\infty} \sqrt{n}\delta_n(\boldsymbol{x}), \text{ with } N^{-1}\sum_{i=1}^{N} \delta_0^2(\boldsymbol{x}_i) \le \eta^2, \qquad (5)$$

  for given η^2. We also impose a bound $N^{-1}\sum_{i=1}^{N}\sigma^2(\boldsymbol{x}_i) \le \sigma_0^2$ for a given σ_0^2 ($= 1$ w.l.o.g.).

- **Optimality and variational mathematics**: In KW we establish asymptotic normality of the estimate $\hat{\boldsymbol{\theta}}_n$, from which we obtain the MSE matrix of $\hat{\boldsymbol{\theta}}_n$. Our loss function is to be asymptotic, average MSE when the conditional quantile $Y_\tau(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\theta}_\tau + \delta_n(\boldsymbol{x})$, for $\boldsymbol{x} \in \chi$, is incorrectly estimated by $\hat{Y}_n(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n$, i.e.

$$\text{amse} = \lim_n \frac{1}{N} \sum_{i=1}^{N} E\left[\left\{\sqrt{n}\left(\hat{Y}_n(\boldsymbol{x}_i) - Y_\tau(\boldsymbol{x}_i)\right)\right\}^2\right].$$

The amse is evaluated, and then maximized over δ_0 using variational methods. The maximum amse depends on various matrices through the design measure

$$\xi_i = \text{fraction of observations made at } \boldsymbol{x}_i.$$

We find that $\max_{\delta_0}$amse is $\frac{\tau(1-\tau)}{g_{\bar{\varepsilon}}^2(0)} + \eta^2$ times

$$\mathcal{L}_\nu(\xi|\sigma) = (1-\nu)\, tr(\boldsymbol{AS}) + \nu ch_{\max}(\boldsymbol{AT}),$$

where $\nu = \eta^2 \Big/ \left\{\frac{\tau(1-\tau)}{g_{\bar{\varepsilon}}^2(0)} + \eta^2\right\} \in [0,1]$.

$$\mathcal{L}_\nu(\xi|\sigma) = (1-\nu)\, tr(\boldsymbol{AS}) + \nu ch_{\max}(\boldsymbol{AT}) \tag{6}$$

Here

$$\begin{aligned}
\boldsymbol{A} &= N^{-1} \sum_{i=1}^{N} \boldsymbol{f}_0(\boldsymbol{x}_i)\, \boldsymbol{f}_0'(\boldsymbol{x}_i), \\
\boldsymbol{B} &= \sum_{\xi_i > 0} \boldsymbol{f}_0(\boldsymbol{x}_i)\, \boldsymbol{f}_0'(\boldsymbol{x}_i) \left(\frac{\xi_i}{\sigma(\boldsymbol{x}_i)}\right), \\
\boldsymbol{S} &= \boldsymbol{B}^{-1} \left[\sum_{\xi_i > 0} \boldsymbol{f}_0(\boldsymbol{x}_i)\, \boldsymbol{f}_0'(\boldsymbol{x}_i)\, \xi_i\right] \boldsymbol{B}^{-1}, \\
\boldsymbol{T} &= \boldsymbol{B}^{-1} \left[\sum_{\xi_i > 0} \boldsymbol{f}_0(\boldsymbol{x}_i)\, \boldsymbol{f}_0'(\boldsymbol{x}_i) \left(\frac{\xi_i}{\sigma(\boldsymbol{x}_i)}\right)^2\right] \boldsymbol{B}^{-1}.
\end{aligned}$$

- The first component ($tr(\boldsymbol{AS})$) of $\mathcal{L}_\nu(\xi|\sigma)$ arises solely from variation – $\boldsymbol{f}_0'(\boldsymbol{x})\, \boldsymbol{S} \boldsymbol{f}_0(\boldsymbol{x})$ is the asymptotic variance of $\sqrt{n} \boldsymbol{f}_0'(\boldsymbol{x})\, \hat{\boldsymbol{\theta}}_n = \sqrt{n} \hat{Y}_n(\boldsymbol{x})$. A 'classical' (non-robust) design aims to minimize $\mathcal{L}_0$; this is appropriate if one has absolute faith in one's model.

- The second ($ch_{\max}(\boldsymbol{AT})$) arises from bias – the asymptotic bias of $\sqrt{n}\boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\theta}}_n$ is

$$\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{B}^{-1}\left[\sum_{\xi_i>0}\boldsymbol{f}_0(\boldsymbol{x}_i)\delta_0(\boldsymbol{x}_i)\xi_i\right]\left(=\boldsymbol{c}'(\boldsymbol{x})\boldsymbol{d},\ \text{say}\right);$$

this is squared, averaged over χ and maximized over $\boldsymbol{d}=(\delta_0(\boldsymbol{x}_1),...,\delta_0(\boldsymbol{x}_N))'$. This amounts to maximizing a quadratic form $\boldsymbol{d}'\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{c}(\boldsymbol{x}_i)\boldsymbol{c}'(\boldsymbol{x}_i)\right]\boldsymbol{d}$ subject to a bound (from (5)) $\boldsymbol{d}'\boldsymbol{d}\leq N\eta^2$ and a linear constraint

$$\sum_{i=1}^{N}\boldsymbol{f}_0(\boldsymbol{x}_i)\sqrt{n}\delta_n(\boldsymbol{x}_i)\sim(\boldsymbol{f}_0(\boldsymbol{x}_1),...,\boldsymbol{f}_0(\boldsymbol{x}_N))\boldsymbol{d}=\boldsymbol{0}.$$

Hence ... maximum eigenvalue.

- We parameterize the designs by $\nu=\eta^2\Big/\left\{\frac{\tau(1-\tau)}{g_\varepsilon^2(0)}+\eta^2\right\}\in[0,1]$, which may be chosen by the experimenter, representing his relative concern for errors due to bias rather than to variation. Once ν is chosen, the designs do not depend upon τ.

- **Design construction**. We design for the example above, and compare five designs – ES (n equally spaced points spanning $\chi = [1, 400]$) and:

KW1 These attain minimax amse, i.e. minimize (6), for a particular value of ν; each is assessed for $0 \leq \nu \leq 1$. When $\nu = 0$ the loss is the average variance of the predicted values. The minimization is carried out via a genetic algorithm.

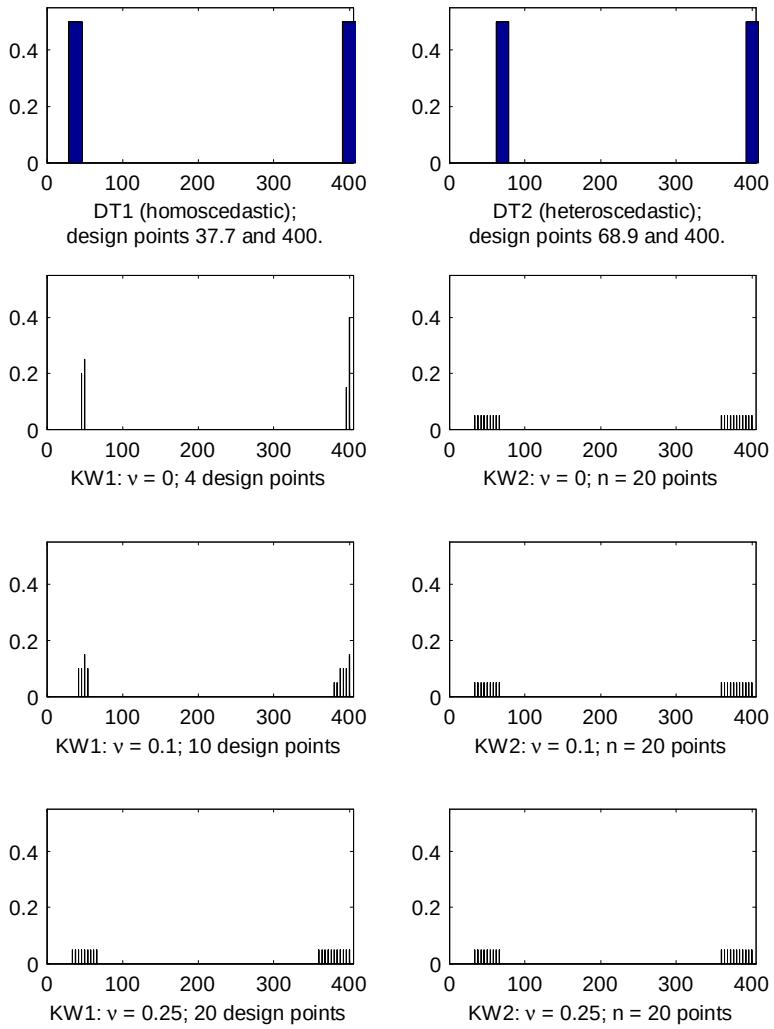
KW2 We have found designs minimizing the maximum amse, with the maximum evaluated not only over δ_0 but also over variance functions $\sigma^2(x_i) \propto \xi_i^r$ for $r \in (-\infty, \infty)$. It turns out that $r = 1$ is least favourable, and that the minimizing design must be supported on n distinct points. These points are found via an exchange algorithm.

DT1 These 'D-optimal' designs minimize the determinant of the asymptotic covariance matrix of the parameter estimates, assuming homoscedastic (i.e., equally varied) errors. They do not depend on the choice of τ and place equal weight on two points, derived explicitly in DT (2012).

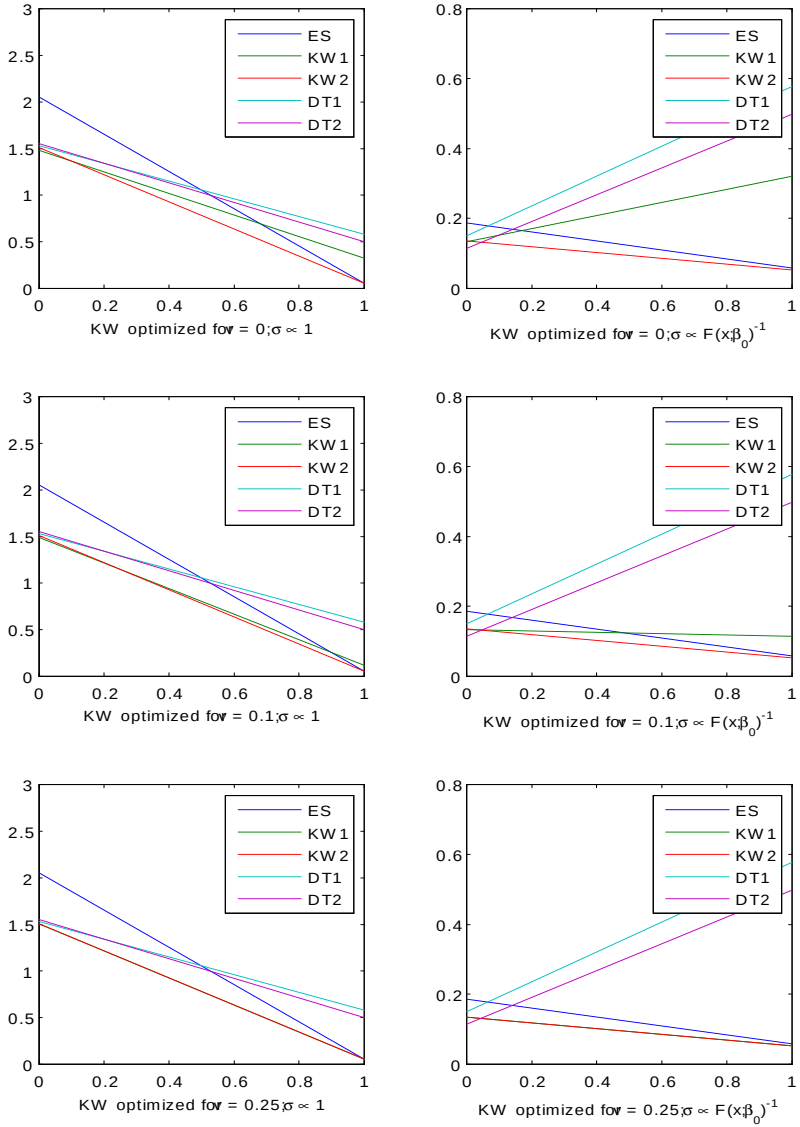
DT2 As for DT1, but derived assuming heteroscedastic errors $\sigma(x) \propto 1/F_0(x)$.

- All designs are assessed in the presence of both homoscedastic errors, and heteroscedasticity of the form above.

- The designs are constructed according to different optimality criteria and goals. Comparisons when $\nu = 0$ reflect efficiency, and there the D-optimal designs should be 'nearly' optimal ... . Comparisons with $\nu > 0$ reflect robustness, for which KW1 and KW2 should be optimal, at least for the value of ν used in their construction.



Designs; KW1&2 constructed for optimality at particular values of ν.



Maximum amse vs. ν.

Conclusions and recommendations

- If the model is in doubt, then substantial reductions in MSE can be attained by employing notions of robustness.

- If the 'classically' optimal design is available, then an easy robustification comes about by spreading its replicates into clusters of design points at distinct but nearby locations.

- If one is willing to do the analytic and numerical work required to obtain optimal robustness, then the n-point ('KW2') designs are much more easily obtained (by the exchange algorithm) than are the minimax amse designs (by the genetic algorithm). In the cases considered here the performance of KW2 was either better than that of KW1, or the difference in performance was negligible.