Notes on Martingalization

[These notes were written for my 1999 NSF proposal and describe some proposed work on asymptotic inference on the quantile regression process.]

Koenker and Machado [6] describes some early steps in a long march toward a complete theory of inference for the quantile regression process. The initial steps are important because a.) they have clarified the crucial role of the Bessel process and its natural role extending classical Kolmogorov-Smirnov results to situations in dimension greater than one, and b.) they yield a theory for the Wald and rankscore tests that successfully accommodate the linear location-scale model under the null.

What is to be done? The theory available in Koenker and Machado enables us to test a wide variety of simple null hypotheses of the form $H_0 = \{\beta(\tau) = 0 : \tau \in \mathcal{T}\}$ where $\mathcal{T}$ denotes closed interval contained in (0, 1). But as in the closely related literature on goodness of fit, we are often interested in composite hypotheses that involve further estimation of parameters. For example, in the iid error linear model we have $\beta(\tau) = \beta + F_u^{-1}(\tau)e_1$, so all the slope coordinates are constant, independent of τ. This is obviously an important hypothesis, but it falls outside the domain of [6] since it involves the fixed slope nuisance parameters.

The classical situation, described by Durbin [4], involves the Kolmogorov-Smirnov test. Suppose we wish to test that the random sample $\{y_1, \ldots, y_n\}$ on Y comes from the df F_0. From an asymptotic standpoint it proves convenient to consider the process

$$v_n(\tau) = \sqrt{n}(F_0(F_n^{-1}(\tau)) - \tau), \quad \tau \in [0, 1],$$

where $F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$. For notational convenience we will denote $G_n(\tau) = F_0(F_n^{-1}(\tau))$ in what follows.

The process $v_n(\tau)$ converges weakly to a Brownian bridge process, and the Kolmogorov-Smirnov statistics $\xi_n = \sup|v_n(t)|$ and $\xi_n^{\pm} = \sup v_n(t)^{\pm}$ can be used to conduct inference. If, however, we wish to test that F_0 is an element of some parametric family $\mathcal{F}_\theta$, with unknown parameter θ the situation is somewhat more complicated. We can, of course, estimate θ, and set $G(\tau, \hat{\theta}_n) = F(y, \hat{\theta}_n)$ so $G(\tau, \theta_0) = \tau$, and consider the process

$$u_n(\tau) = \sqrt{n}(G_n(\tau) - G(\tau, \hat{\theta}_n)).$$

But when we do this we find, under classical maximum likelihood conditions on the family $\mathcal{F}_\theta$, that $v_n(\tau)$ converges weakly to a Gaussian process with mean zero and covariance function,

$$Ev_n(\tau_1)v_n(\tau_2) = \tau_1 \wedge \tau_2 - \tau_1\tau_2 - g_0(\tau_1)'\mathcal{J}^{-1}g_0(\tau_2)$$

where

$$g_0(\tau) \equiv g(\tau, \theta_0) = \left.\frac{\partial F(y, \theta)}{\partial \theta}\right|_{\substack{y = F^{-1}(\tau,\theta) \\ \theta = \theta_0,}}$$

and $\mathcal{J} = E\nabla_\theta \log f(y, \theta_0)\nabla_\theta \log f(y, \theta_0)'$, the Fisher information for θ. The last term of the covariance function reflects the contribution of $\hat{\theta}_n$. Had we been able to use θ_0 instead of $\hat{\theta}_n$, we would have had only the leading Brownian bridge terms.

The presence of this final term considerably complicates the development of a valid test procedure. Durbin discusses a sample splitting strategy also suggested by Rao, but criticizes it for its sensitivity to the convention employed for the splitting. Similar criticism could, of course, be leveled at various bootstrapping schemes that might be used to construct critical values. Special cases, notably for $\mathcal{F}_0$ Gaussian, have been considered by various authors, but a general approach eluded my attention until I received Bai [1], which seemed to contain a key result required to move beyond the theory developed in Koenker and Machado into the realm of composite nulls.

To my chagrin this device had appeared considerably earlier in Khmaladze [5], but seems to have been largely ignored in the statistics literature until recently, when it has been prominently featured in several papers [10, 7, 8] by Stute and his colleagues. The idea is quite simple and attractive. I will first try to briefly explain it in the context of Durbin's problem alluded to above, since this is the immediate context treated by Khmaladze. I will then attempt to explain why the idea is so appealing in the context of quantile regression inference. I should stress at this point that the applications of Bai to problems of parametric inference in time-series models are also extremely promising, but are focused on quite different objectives than those described below.

Expanding $G(\tau, \hat{\theta})$ around $\theta = \theta_0$ we have,

$$G(\tau, \hat{\theta}) = \tau + (\hat{\theta} - \theta_0)' g(\tau, \theta^*),$$

for some $\theta^* = \lambda\theta_0 + (1-\lambda)\hat{\theta}$ and $\lambda \in (0,1)$. Assuming that we have a Bahadur representation of the form,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \int_0^1 h(s, \theta_0) dv_n(s) + o_p(1),$$

we may reexpress $u_n(\tau)$ as,

$$u_n(\tau) = v_n(\tau) - g(\tau, \theta_0)' \int_0^1 h(s, \theta_0) dv_n(s) + r_n(\tau),$$

and show that it converges weakly to the Gaussian process

$$u(\tau) = v(\tau) - g(\tau, \theta_0)' \int_0^1 h(s, \theta_0) dv(s)$$

where v as above denotes a Brownian bridge. The appearance of the last term clearly implies that the limiting behavior of $u_n(\tau)$ depends upon the family $\mathcal{F}_\theta$, and perhaps even on the particular value θ_0, thus rendering tests based on functionals of $u_n(\tau)$ evidently distribution unfree. This is the Durbin problem. Under conventional regularity conditions the representation is not invertible, indicating that in general one cannot transform $u(t)$ into $v(t)$.

Khmaladze proposes an alternative representation for $u(\tau)$ that does permit transformation of the parametric empirical process, $u_n(\tau)$, under the null hypothesis, into standard Brownian motion. The underlying idea is closely tied to the classical Doob-Meyer decomposition of $G_n(\tau)$. It is easy to see that $G_n(\tau)$ is Markov: $n\Delta G_n(\tau) = n[\Delta G_n(\tau + \Delta\tau) - G_n(\tau)]$ is binomial with sample size $n(1 - G_n(\tau))$, $p = \Delta\tau/(1-\tau)$, and initial condition $G_n(0) = 0$. Thus, the conditional expectation of $\Delta G_n(\tau)$ with respect to the natural filtration $\{\mathcal{F}_\tau^{G_n} : 0 \le \tau \le 1\}$ is,

$$E[\Delta G_n(\tau) | \mathcal{F}_\tau^{G_n}] = \frac{1 - G_n(\tau)}{1 - \tau} \Delta\tau.$$

This suggests the representation,

$$G_n(t) = \int_0^t \frac{1 - G_n(s)}{1 - s} ds + m_n(t)$$

where the process $\{m_n(t), \mathcal{F}_t^{G_n}\}$ is a martingale. It follows that the empirical process $v_n(t) = \sqrt{n}(G_n(t) - t)$ may be represented as,

$$w_n(t) = v_n(t) + \int_0^t \frac{v_n(s)}{1-s} ds.$$

This is the classical Doob-Meyer representation of the empirical process, and there is a corresponding representation of the limiting Brownian bridge process in terms of standard Brownian motion. Khmaladze essentially extends this idea to a much broader class of transformations applicable to the parametric empirical process $u_n(t)$.

Let $g(t) = (t, g_1(t), \ldots, g_l(t))'$ be a vector of real valued functions on $[0,1]$, and suppose that the vector of derivative functions $\dot{g}(t)$ are linearly independent, so that the matrix

$$C(t) = \int_t^1 \dot{g}(s)\dot{g}(s)' ds$$

is non-singular for each $t \in (0,1)$. The Doob-Meyer representation may be regarded as a special case of the following much more general device,

$$w(t) = v(t) - \int_0^t \dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(r) dv(r) ds.$$

To specialize to the Doob-Meyer case, choose, $g(s) = s$, so $C(s) = 1 - s$, and note that,

$$\int_s^1 \dot{g}(r) dv(r) = v(1) - v(s) = -v(s).$$

The power of the general version of this device is apparent if we return to the parametric empirical process representation,

$$u_n(t) = v_n(t) - g(t, \theta_0)' \int_0^1 h(s, \theta_0) dv_n(s) + r_n(t).$$

Applying the martingale transformation to $v_n(t)$ yields

$$w_n(t) = v_n(t) - \int_0^t \dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(r) dv_n(r) ds$$

with $w_n(t) \Rightarrow w(t)$. Now consider applying the same same transformation to $u_n(t)$, with $g(t) = (t, g_0(t)) \equiv (t, g(t, \theta_0))$, so we obtain,

$$\tilde{u}_n(t) = u_n(t) - \int_0^t \dot{g}(s)' C^{-1}(s) \int_s^1 g(r) du_n(r) ds.$$

Khmaladze proves that $\tilde{u}_n(t) \Rightarrow w(t)$, so in effect the martingale transformation has purged the parametric empirical process $u_n(t)$ of the contribution of $\hat{\theta}_n$ in its first order asymptotic representation, and tests based on $\tilde{u}_n(t)$ can be constructed that are now asymptotically distribution free.

A heuristic sketch of the argument goes as follows. The linear operator,

$$P_g[h](t) = \int_0^t \dot{g}(s)' C^{-1}(s) \int_s^1 \dot{g}(r) dh(r) ds$$

has the property $P_g[g] = g$, and consequently the transformation $u_n \to u_n - P_g[u_n]$ annihilates the g_0 term in the linear representation of u_n leaving

$$\tilde{u}_n \equiv u_n - P_g[u_n] = v_n - P_g[v_n] + R_n.$$

Since the remainder process R_n converges in $\mathcal{L}_2[0,1]$ to zero, we have that $\tilde{u}_n \Rightarrow w$, i.e., the transformed process $\tilde{u}_n$ converges weakly to standard Brownian motion.

The transformation from u_n to $\tilde{u}_n$ provides an elegant general solution to the Durbin problem at least from the standpoint of conventional first-order asymptotic theory. I am not aware of any published Monte-Carlo work on the original one-sample goodness of fit problem. However, a note by Nikabadze and Stute [8] reports some Monte-Carlo results for a closely related problem in which $G_n(\tau)$ is replaced by the Kaplan-Meier estimator, and these results are quite encouraging.

I will briefly describe one leading example to illustrate how the foregoing theory may be applied in the context of quantile regression inference. Suppose that we would like to test the null hypothesis that we have the linear location-scale shift model

$$y_i = x_i'\theta_1 + (x_i'\theta_2)u_i$$

with $\{u_i\}$ iid from some (perhaps) unknown df F, versus the alternative hypothesis that we have some more general form of linear conditional quantile functions. Traditionally, econometrics, and statistics more generally, has focused almost exclusively on how covariates affect the location and scale of the conditional distribution of $y|x$. Thus, the linear location scale shift model subsumes a broad spectrum of conventional models. Nevertheless, it is, I believe, crucial to explore the possibility that covariates act in more complex ways to alter other features of the conditional distribution, not simply its location ands scale, but its shape as well. This is the objective of the inference strategy I propose to explore.

Under the null we have,

$$Q_{y_i}(\tau|x) = x'\beta(\tau)$$

where $\beta(\tau) = \theta_1 + \theta_2 F^{-1}(\tau)$, so the hypothesis implies that each of the p functions $\beta(\tau) = (\beta_1(\tau), \dots, \beta_p(\tau))'$ can be expressed as an affine function of $F^{-1}(\tau)$. Were we confident about the validity of the null, it is clear that adaptively efficient estimators of (θ_1, θ_2) could be designed. Tests based on the quantile regression process, $\hat{\beta}(\tau)$, offer a systematic way to determine whether such confidence is justified. Such tests do not, however, fall into the class of simple nulls investigated in Koenker and Machado since they involve the unknown nuisance parameters θ_1 and θ_2. It may also be worth remarking that by appropriate interpretation of the vectors $\{x_i\}$ we may extend the present theory into the realm of non-parametric quantile regression using series expansions, B-splines and related methods. Locally polynomial quantile regression using kernel weighting as in Chaudhuri [3] and Welsh [11], could also be explored in this fashion.

Suppose for the moment that $F^{-1}(\tau)$ were known and $x_i \equiv 1, i = 1, \dots, n$. This puts us back into a situation very close to the Durbin problem. We have

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbf{R}} \sum_{i=1}^{n} \rho_\tau(y_i - b)$$

as simply the empirical quantile function based on the sample $\{y_1, \dots, y_n\}$. We know that

$$v_n(\tau) = \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))/s(\tau) \Rightarrow v(\tau)$$

where $s(\tau) = 1/f(F^{-1}(\tau))$ is assumed to be strictly positive, and under the null $\beta(\tau) = \theta_1 + \theta_2 F^{-1}(\tau)$ for real numbers (θ_1, θ_2). As above, $v(\tau)$ denotes the Brownian bridge process.

Let $\tilde{\theta}_n$ denote an estimator of the parameter, $\theta_0 = (\theta_1, \theta_2)'$ satisfying

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = O_p(1).$$

Set $\xi(\tau) = (1, F^{-1}(\tau))$ and write, $\tilde{\beta}(\tau) = \tilde{\theta}'\xi(\tau)$, so

$$\begin{aligned} u_n(\tau) &= \sqrt{n}(\hat{\beta}(\tau) - \tilde{\beta}(\tau))/s(\tau) \\ &= \sqrt{n}(\hat{\beta}(\tau) - \beta(\tau))/s(\tau) - \sqrt{n}(\tilde{\beta}(\tau) - \beta(\tau))/s(\tau) \\ &= v_n(\tau) - \sqrt{n}(\tilde{\theta}_n - \theta_0)'\xi(\tau)/s(\tau). \end{aligned}$$

We find ourselves back in precisely the Durbin problem with $g_0(\tau) = \xi(\tau)/s(\tau)$. Applying the Khmaladze martingalization, we obtain

$$\tilde{u}_n = u_n - P_g[u_n]$$

where $g(\tau) = (\tau, \xi(\tau))'/s(\tau))$. In this case,

$$\dot{g}(\tau) = (1, \dot{f}/f, 1 + F^{-1}(\tau) \cdot \dot{f}/f)'$$

where the score function $\dot{f}/f$ is evaluated at $F^{-1}(\tau)$. In the Gaussian case, $F = \Phi$, this yields,

$$\dot{g}(\tau) = (1, -\Phi^{-1}(\tau), 1 - \Phi^{-1}(\tau)^2)'.$$

It should come as no surprise that the result is so strongly reminiscent of the prior discussion of tests based on the empirical distribution function. The weak convergence theory of the empirical distribution function and the corresponding theory for the empirical quantile function are intimately tied together. A convenient reference is is Beirlant and Deheuvels [2] and an extended treatment is available in Shorack and Wellner [9].

The extension to a general p dimensional covariate vector, x_i, is straightforward. We now have a $p \times 2$ matrix of nuisance parameters comprising θ_0, but the $g(\tau)$ function remains the same. In this case we have a p-variate empirical process and the standardization is a little more complicated,

$$u_n(\tau) = \sqrt{n} D_n^{-1/2} H_n(\hat{\beta}(\tau) - \tilde{\beta}(\tau))$$

where $D_n = n^{-1}X'X, H_n = n^{-1}X'\Omega^{-1}X$, $\Omega =$ diag $(x_i'\gamma)$. But the same procedure applied to each coordinate yields

$$\tilde{u}_n = u_n - P_n[u_n] \Rightarrow w_p$$

where w_p denotes a p-variate standard Brownian motion. Obviously, we can transform $\tilde{u}_n$ again to obtain a p-variate Brownian bridge and consider test statistics of the form $\sup ||\tilde{u}(\tau) - \tau\tilde{u}(1)||^2/\tau(1-\tau)$ that have the same squared Bessel process behavior as those investigated previously in Koenker and Machado.

In the more interesting, and realistic, case that $F^{-1}(u)$ is no longer assumed to be known, we seem to have several options. A particularly simple option would be to choose one coordinate from $\hat{\beta}(\tau)$ and let it play the role of $F^{-1}(\tau)$ and proceed as before with a $(p-1)$-variate process. The obvious candidate for the chosen coordinate is the intercept. Alternatively, we may consider various reduced rank schemes which would have the advantage that they would avoid privileging any one coordinate of the vector process $\hat{\beta}(\tau)$.

There is a wide array of other hypotheses on $\hat{\beta}(\tau)$ that fall into the domain of applicability of the martingalization device. And parallel to the theory we have sketched based on the primal quantile regression process, $\hat{\beta}(\tau)$, there is a dual theory based on the regression rankscore process that also merits investigation. There are many issues that have been glossed over in this brief description. There are obviously important remaining problems of nuisance parameter estimation involving $s(\tau)$ and Ω. Although these problems do not play a significant role in the asymptotic theory, they may be of considerable importance for the performance of the tests in moderate sized samples.

References

[1] Bai, J. Testing parametric conditional distributions of dynamic models. preprint, 1998.

[2] Beirlant, J., and Deheuvels, P. On the approximation of P-P and Q-Q plot processes by brownian bridges. *Stat. and Prob. Letters 9* (1990), 241–51.

[3] Chaudhuri, P. Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Mult. Analysis. 39* (1991), 246–269.

[4] Durbin, J. *Distribution theory for tests based on the sample distribution function.* SIAM, 1973.

[5] Khmaladze, E. V. Martingale approach in the theory of goodness-of-fit tests. *Theory of Prob. and its Apps 26* (1981), 240–257.

[6] Koenker, R., and Machado, J. Goodness of fit and related inference processes for quantile regression. *J. of Am. Stat. Assoc.* (1999). forthcoming.

[7] Koul, H., and Stute, W. Nonparametric model checks for time series. *Annals of Stat. 27* (1999), 204–236.

[8] Nikabadze, A., and Stute, W. Model checks under random censorship. *Stat. and Prob. Letters 32* (1997), 249–259.

[9] Shorack, G. R., and Wellner, J. *Empirical Processes with Applications to Statistics.* Wiley: New York, 1986.

[10] Stute, W. Nonparametric model checks for regression. *AnlsStat 25* (1997), 613–641.

[11] Welsh, A. H. Robust estimation of smooth regression and spread functions and their derivatives. *StSinica 6* (1996), 347–366.