Yu, K., Chen, C.W., Reed, C. & Dunson, D. (2013). Bayesian variable selection in quantile regression. *Stat. Interface*, **6**, 261–274.
Yu, K. & Moyeed, R.A. (2001). Bayesian quantile regression. *Stat. Probabil. Lett.*, **54**, 437–447.

[*Received November 2015, accepted November 2015*]

International Statistical Review (2016), 84, 3, 362–367 doi:10.1111/insr.12180

Discussion: Should a Working Model Actually *Work*?

Xiao-Li Meng

Department of Statistics, Harvard University, Cambridge, MA, USA
E-mail: meng@stat.harvard.edu

1 All Models are Wrong, But Some are Harmful

Defining precisely the term *working model* is impossible. But here is a working definition: a working model is a model we adopt for particular purposes with the knowledge that it may be flawed in some other aspects. All models are working models then, following George Box's mantra that all models are wrong, but some are useful. Unfortunately, the complement of 'useful' in this mantra is not merely 'useless', because all models are wrong, but some are harmful. It is of course impossible either to define precisely what Box meant by 'useful' or what amounts to as 'harmful' without contextual information; one person's poison can easily be another person's medicine. Furthermore, one could–and many do–question what constitutes a *true model*, when the very concept of *reality* has been intensely debated by physicists, let alone philosophers; see Peat (2002) for a fascinating account of this debate.

But I surmise that most of us would agree that, as a practical guideline, a model is harmful if it *routinely* leads to misleading results for which the model is designed, especially when we have *both* empirical verification and theoretical understanding of the harm. This overdue paper by Yang, Wang and He (YWH) provides exactly that for the asymmetric Laplace (AL) likelihood for Bayesian quantile regression. The paper is overdue because much harm has already been done, judging from the list of literature cited in YWH. I am therefore pleased and puzzled as a discussant: pleased because finally the harm can be stopped (hopefully), but puzzled by the fact that this seriously flawed model has been on the Bayesian track for so long, when there were clear warning signs from the moment it was built (and there were no Volkswagen technicians involved, as far as I can tell), as I'll discuss in Section 2.

Tongue in cheek or not, a working model needs to have a reasonable chance to *work*, that is, to be a reasonable depiction of the data being analysed. This seemingly tautological requirement actually has some wisdom teeth in it (which can be taken out once the wisdom is retained). It implies that a working model should *not* be determined, *solely or even largely*, by particular analysis interests, mathematical simplicity or beauty, or estimation or computational procedures. Life would be too easy if nature would generate data according to our particular interests or our analysis or computational ability.

But apparently, this minimal requirement has been overlooked, and AL is by no means the only example. As a reference point, the popular exponential random graph models for social networks specify the probability of observing a network X, typically represented by a matrix of 0s and 1s, as

$$P_\theta(X = x) \propto e^{\theta^\top S(x)}, \tag{1}$$

where θ is a vector of parameters, and $S(x)$ is known as 'important features', 'graphical statistics', or 'network attributes'. Adopting an exponential family is natural–it is mathematically convenient, and it can also serve well as an approximation to many non-exponential families by Taylor expanding a log-density with a sufficient number of terms.

What becomes problematic is when the choice of the sufficient statistics $S(x)$ in (1) is made *only* according to an investigator's limited and specific interest, as in a number of applications (citations are omitted to avoid reducing my social network). That is, the 'sufficient statistics' $S(x)$ would be set automatically to be some dyad counts (e.g. the number of edges) or triad counts (e.g. the number of triangles), depending on whether dyadic or triadic relationships are the interest of the study, without leaving the model any room for accommodating lack of fit because (1) is entirely determined by the chosen $S(x)$.

To accurately model a social network is a challenging task because the very nature of a social (or other) network is that everything is connected to one degree or another. Iterative contemplation and strong assumptions are typically needed in order to create meaningful 'internal replications' that are central to statistical inference (see Liu and Meng, 2016, for a discussion). Statistically and scientifically principled modeling strategies have been developed (e.g. Hoff et al., 2002; Hoff, 2005), but it is not unexpected that convenient models such as exponential random graph models gained popularity in a field where it is not easy to invalidate a model empirically because of lack of external replications. But this is not the case for quantile regression, where we do have reasonable replications (e.g. individual subjects), and hence, we can examine empirically the distributional shapes of our data or of various residuals after fitting a (working) model. It is therefore puzzling that models such as AL have gained popularity, when there are obvious signs that they are harmful, at least for Bayesian analysis, which relies critically on the entire likelihood, working or not.

2 Yellow and Red Lights for the Asymmetric Laplace

It is not unreasonable to approximate the τth quantile of the conditional distribution $p(Y|\mathbf{X})$ by a linear term, $\mathbf{x}^\top\beta(\tau)$, for any τ; again, one can think of this as a first-order Taylor expansion of the quantile in $\mathbf{x}$. It is then also harmless to write our working model as

$$y_i = \mathbf{x}_i^\top\beta(\tau) + \sigma(\tau)\epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

as long as we do *not* impose $\mathrm{E}(\epsilon) = 0$ or $\mathrm{V}(\epsilon) = 1$. To construct a working likelihood, we need to pose some distributional assumptions for ϵ. This is where we can introduce much harm if we pose assumptions without understanding their implications, without making them sufficiently flexible and/or without performing a minimal empirical check.

The AL model apparently was introduced (and followed) without such due diligence. It specifies ϵ as a weighted mixture of two standard exponential variables on $R^- = (-\infty, 0]$ and $R^+ = (0, \infty)$, and hence, they will be denoted by Z^- and Z^+, with expectations -1 and 1, respectively. That is,

$$\epsilon = \xi\frac{Z^-}{1-\tau} + (1-\xi)\frac{Z^+}{\tau}, \quad \text{with } \xi \sim Ber(\tau) \perp \left\{Z^+, Z^-\right\}. \tag{3}$$

- *Yellow light.* The mixture nature should remind us that the resulting working likelihood does not form an exponential family, and there is no sufficient statistics of dimensions lower than the data themselves. İn turn, this should remind us immediately that the corresponding Bayesian analysis will depend on the data in a much more nuanced way than merely via the classical Koenker–Bassett estimator, despite the fact that the AL likelihood was motivated primarily by the desire to leave this estimator untouched. I label this as a yellow light because asymptotically, we may still have an approximately exponential-family likelihood, almost always normal, and this approximation could work well even when the sample size is not that large, as indicated in YWH. Hence, it may not necessarily be a stopping light, but it should be a clear sign for caution, because both the beauty and burden of Bayesian analysis is that it operates with distributions, not point estimators.
- *Red light.* Because a Bayesian analysis takes into account the entire likelihood, not just its mode, if there is a sign that the likelihood is seriously misspecified, then it is a red light. Detecting serious likelihood misspecification is not a difficult task when we have observable replications, as mentioned earlier. Our theoretical or strong prior knowledge can also help us to rule out specifications or at least cast strong doubts about them prior to seeing our data. In particular, the AL has a rather peculiar density shape, with two exponentially decaying but asymmetric tails. Now, how frequently have we observed such shaped histograms in our applied work? Worse, even if my data do have such a distributional shape, why should its decay rates be determined entirely by the particular quantile level τ that I happen choose to study? And what if I want to look at several τ's, such as $\tau = 50\%, 75\%, 90\%$, as in the analysis of the woman's labour force data in YWH? Do I then have to use three mathematically incomparable working models for the same data set? Putting aside the impossibility that all three models are reasonable depictions of our data, shall we minimally examine whether any of them has a chance to fit our data?

3 AL is for Artificial Likelihood, and It Needs a Bartlisation

'Yes, but it does not really matter.' I can imagine such a response from those who care about only the point estimator for the quantile regression. And I understand that AL nicely reproduces the Koenker–Bassett estimator as its maximum likelihood estimator and is easy to handle, analytically and numerically. But these are good reasons only if they are not at the expense of *validity*. As YWH demonstrates, the beauty of the AL likelihood disappears as soon as we go beyond the point estimator. To see a more general picture, let us assume that we have an estimator $\hat{\theta}$, which is the (global) minimiser of a non-negative objective function $R(\theta; \mathcal{D})$, where $\mathcal{D}$ denotes our data. It follows trivially then that $\hat{\theta}$ is the maximum likelihood estimator from the *artificial likelihood* (another AL; so AL is AL!)

$$L_w(\theta|D) \propto e^{-R(\theta;\mathcal{D})}, \tag{4}$$

as long as $\int_{\mathcal{D}} e^{-R(\theta;\mathcal{D})} d\mathcal{D}$ is finite and free of θ. However, if this were the true likelihood for our data, then it is well known that under regularity conditions, including $R(\theta; \mathcal{D})$ being twice differentiable as a function of θ, the inverse of its observed Fisher information, $I_w = R''(\hat{\theta}; \mathcal{D})$, should provide a consistent estimator of the asymptotic sampling variance of $\hat{\theta}$. But it is also well known that under similar regularity conditions, the asymptotic sampling variance of $\hat{\theta}$ is of a 'sandwich' form, which can be consistently estimated by $I_w^{-1} \hat{V} I_w^{-1}$, where $\hat{V}$ is a consistent estimator of $\mathrm{V}(R'(\theta; \mathcal{D}))$, the variance of the 'score function' with respect to the true model of $\mathcal{D}$, not the working model (4).

Given the usual asymptotic equivalence between the asymptotic sampling variance and posterior variance, it is then clear that in order for a posterior inference from an artificial likelihood to provide an asymptotically valid inference, minimally we need to require that asymptotically I_w^{-1} and $I_w^{-1}\hat{V}I_w^{-1}$ are the same, or equivalently I_w and $\mathrm{V}(R'(\theta;\mathcal{D}))$ are the same. When an artificial likelihood coincides with the actual likelihood for the data, this requirement is automatically satisfied because of the well-known *second Bartlett identity*; that is, the variance of the score function is the same as the expected Fisher information, where both the variance and expectation calculations are performed under the true model. When this critical identity fails, erroneous confidence intervals or posterior intervals are expected, as YWH's simulation demonstrated.

Of course, because the AL likelihood is not everywhere differentiable, we cannot directly invoke the concept of score function or Fisher information. Nevertheless, the idea of adjusting the asymptotic (posterior) variance–covariance induced by the AL likelihood to the actual 'sandwiched' form (posterior) variance–covariance is the same, and this is exactly what YWH did. This approach of course is not restricted to quantile regression; see for example Müller (2013). It can also be viewed as a form of 'Bartlisation', a process proposed for correcting the *H*-likelihood, another incidence of artificial likelihood (Meng, 2009). That is, when an artificial likelihood fails to admit the crucial second Bartlett identity, we can try to tune or modify it to make the identity hold, either exactly or asymptotically. Perhaps the most well-known and adopted approach is to construct a quasi-likelihood, which enforces the second Bartlett identity via building it into the objective *R* function, albeit this is not always possible (McCullagh, 1983).

However, as shown in Meng (2009), admitting the second Bartlett identity is only a necessary condition for a valid asymptotic inference; the other crucial condition is for the log of the artificial likelihood to be quadratic asymptotically, which is the case for AL, as YWH's (3.1) shows. Fortunately, for many working likelihoods constructed according to some consistent estimators, this condition usually holds. Therefore, the key to ensure such working likelihood functions to be (minimally) useful for Bayesian inference is to carry out a Bartlisation process, even though the resulting adjusted likelihood may still be an artificial one. The promising simulation results reported in YWH demonstrated that this seemingly simplistic Bartlisation adjustment–as it adjusts only for second-order moment–could work well even for modest sample sizes.

4 More Flexible Working Models for Quantile Regression?

Quantile regression is a major branch of the semi-parametric paradigm, and constructing a working likelihood amounts to parameterising and hence restricting a family of semi-parametric models. From a mathematical perspective, the difficulty of constructing a *working* working model analytically then depends on how difficult it is to add on restrictions that are still sufficiently flexible to accommodate many data (distributional) shapes. This in turn depends on whether there is a 'functional independence' between the aspects of a distribution that are restricted by the semi-parametric family and those that are not.

Take moment regression, another major branch of semi-parametric methods, as an example. Suppose we are dealing with continuous variables, and we are modelling only the mean, as in the vast majority of applications. Because the shape of a continuous distribution is functionally independent of its location, we have essentially unlimited flexibility to accommodate our data shape, for example, via using a particular location family. This task becomes increasingly difficult, however, as we add more restrictions on higher-order moments, because we rapidly lose the 'functional independence' between the shape and the moments (e.g. variance–covariance matrices can already affect the distributional shape).

Take hazard regression (i.e. Cox regression), yet another major branch of semi-parametric modeling, as another example. It may seem rather restrictive, as there is a one-to-one correspondence between the hazard function and the cumulative distribution function (CDF). However, the semi-parametric regression that Cox (1972) proposed leaves the entire baseline hazard function unrestricted, which provides great flexibility for further restricting the distributional shape to fit the data. Again, this is possible because of the functional independence between the baseline hazard function and the incremental changes in the hazard functions (as long as the proportional hazard assumption is reasonable).

In comparison with moment regression and hazard regression, quantile regression seems to be most flexible, because it restricts only one value of a CDF, instead of restricting its integral (as for the moment regression) or derivative (as for the hazard regression). Indeed, it is trivial to show that a CDF $F(\epsilon)$ has zero as its τth quantile, as in (3), if and only if it can be written as

$$F(\epsilon) = \begin{cases} \tau G_\tau(\epsilon), & \text{if } \epsilon \le 0; \\ \tau + (1-\tau) H_\tau(\epsilon), & \text{if } \epsilon > 0, \end{cases} \tag{5}$$

where $G_\tau(\epsilon)$ and $H_\tau(\epsilon)$ are (right continuous) CDFs on R^- and R^+, respectively, and they can depend on τ. AL is such a case, because the corresponding G_τ and H_τ are given by

$$G_\tau^{AL}(\epsilon) = e^{(1-\tau)\epsilon}, \quad \epsilon \le 0 \quad \text{and} \quad H_\tau^{AL}(\epsilon) = 1 - e^{-\tau\epsilon}, \quad \epsilon > 0. \tag{6}$$

Or equivalently by extending the stochastic representation (3), ϵ is a random variable with zero as its τth quantile if and only if

$$\epsilon = \xi G_\tau^- + (1-\xi) H_\tau^+, \quad \text{with } \xi \sim Ber(\tau) \perp \{G_\tau^-, H_\tau^+\}, \tag{7}$$

where, with slight abuse of notation, G_τ^- and H_τ^+ denote two arbitrary but not necessarily independent random variables on R^- and R^+, respectively.

Expression (5) is a mathematical representation of an intuitive fact: restricting the τth quantile does not in any way restrict the distributional shape above or below it; it restricts only their relative total masses to be $(1-\tau)/\tau$. Therefore, for any pair of G_τ and H_τ, (5) will lead to a working likelihood for $\beta(\tau)$ (and $\sigma(\tau)$) when we replace ϵ by $(y - \mathbf{x}^\top \beta(\tau))/\sigma(\tau)$. Evidently, we can use empirical likelihood or non-parametric methods to estimate G_τ and H_τ, as cited in YWH. But if one insists on using a parametric working model, then a natural question is if there is a more sensible choice of $\{G_\tau, H_\tau\}$ than (6) once we do not insist on recovering exactly the classical Koenker–Bassett estimator, but rather on ensuring that the resulting Bayesian inference is valid for as large a class of real likelihood functions as possible. After all, the whole reason for introducing a working likelihood is to conduct Bayesian inference, not to recover some known point estimators.

Given the popularity of the AL working model despite of its obvious flaws, I surmise the aforementioned question is not easy to answer, because otherwise the alternative model would have been in use. But fortunately for me, I promised the Editor to finish this discussion by the end of 2015, which is less than 2 h away as I type this sentence. I will therefore have to save the question for YWH and experts in quantile regressions as my 2016 present!

But regardless whether or not a satisfactory answer can be found in 2016, YWH reminded us *WHY* (a cyclical permutation of YWH!) it is important to not let mathematical or computational convenience trump statistical or scientific considerations; minimally, we should at least investigate the consequences of such convenient models, so even if we decide to adopt them, we can properly document and warn others about their potential harmful effects. We all like working models, or any other procedures for that matter, that are simple to understand and easy

to implement. But these desiderata must remain as secondary considerations–our highest priority must be on ensuring their validity, that is, guaranteeing they will lead to answers that are statistically and scientifically defensible. On that note, let me conclude with a New Year toast to YWH for making a working model actually work!

Acknowledgements

I thank Editor Marc Hallin (and Xuming He) for giving me the opportunity to discuss a topic for which I have no prior research experience and Joe Blitzstein, Radu Craiu, Alan Garber, Andrew Gelman, Keli Liu, and Neil Shephard for very helpful comments. I also thank NSF for partial financial support and my family for full moral support, which made it possible for me to spend the last quantile of 2015 on quantiles.

References

Cox, D.R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Series B Stat. Methodol.*, **34**(2), 187–220.
Hoff, P.D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.*, **100**(469), 286–295.
Hoff, P.D., Raftery, A.E. & Handcock, M.S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, **97**(460), 1090–1098.
Liu, K. & Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annu. Rev. Stat. Appl.*, **3**, 79–111.
McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Stat.*, **11**(2), 59–67.
Meng, X.-L. (2009). Decoding the H-likelihood. *Stat. Sci.*, **24**(3), 280–293.
Müller, U.K. (2013). Risk of Bayesian inference in misspecified models and the sandwich covariance matrix. *Econometrica*, **81**(5), 1805–1849.
Peat, F.D. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century.* Washington, DC: Joseph Henry Press.

[*Received January 2016, accepted February 2016*]

International Statistical Review (2016), 84, 3, 367–370 doi:10.1111/insr.12181

Rejoinder

Yunwen Yang[1], Huixia Judy Wang[2], Xuming He[3]

[1]*Google Inc., Seattle, WA, USA*
E-mail: yunweny@google.com
[2]*George Washington University, Washington, D.C., USA*
E-mail: judywang@gwu.edu
[3]*University of Michigan, Ann Arbor, MI, USA*
E-mail: xmhe@umich.edu

1 Can We Trust the Working Model?

Meng is very direct in pointing out that the asymmetric Laplace working likelihood is simply too artificial; in general, it does not provide a decent approximation to the underlying likelihood. This sentiment is shared by Smith. In fact, two such working likelihoods at two values of τ