Title
Bayesian Model Checking with Applications to Hierarchical Models

Author
R. E. Weiss

Publication Date
2011-10-25

Bayesian Model Checking with Applications to Hierarchical Models

Robert E. Weiss*

August 13, 1996

Abstract

In a Bayesian model with proper prior, all functions of the parameters and data are known. After observing the data, the joint prior specification of data and parameters can be checked by comparing the posterior of any function of the parameters to its assumed prior. This paper gives checks for missing predictors, goodness-of-fit, and over-diffuseness of the prior. The approach is illustrated in a hierarchical random effects model.

Key Words: Bayesian Data Analysis, Diagnostics, Goodness-of-Fit, Longitudinal Data, Outlier, Quantile-Quantile Plots.

1 Introduction.

This paper introduces a general approach to Bayesian model checking. Like previous authors (Box, 1981; Chaloner and Brant 1988; Dey, Gelfand, Vlachos and Schwarz 1994; Gelman, Meng and Stern 1996; Meng 1994; Rubin 1984), we may consider a model suspect when some residual or checking function g, a function of the data Y and/or parameters θ, is far from an appropriate measure of center,

*Robert E. Weiss is Assistant Professor, Department of Biostatistics, Box 177220; UCLA School of Public Health; Los Angeles CA 90095-1772 U.S.A.; email rob@rem.ph.ucla.edu. This work was supported by grant #GM50011 from the NIGM. The author thanks Charlie Zhang and Meehyung Cho for help with the calculations and graphs and M. Cho and E. Bradlow for comments.

or out in the tails of some distribution. The art and science of Bayesian model checking currently lies in (i) picking the diagnostic function g; (ii) the choice of relevant diagnostic distribution(s) for g; and (iii) the definition of when the measures indicate a lack of fit.

Zellner (1975) first proposed looking at the posterior distribution of the residuals and functions of the residuals in linear regression. Chaloner and Brant (1988), and Chaloner (1991, 1994), discuss outlier checking in various models and Albert and Chib (1996) extend these methods to discrete data. Box (1980) proposed a special case of the methodology to be proposed here; he marginalizes the parameters out of the prior and permits g to be a function of the data only; Hodges (1994) proposes similar analyses in hierarchical models. Meng (1994) and Gelman et al (1996) permit g to be a function of both data Y and parameters θ. Gelman et al (1996) and Dey et al (1994) give relatively complete presentations of competing methodologies, but their approaches are much more complicated than the approach here; the major differences are in the choice of distributions for comparison. My approach is given in the next section; it includes the methodology of Chaloner and Brant (1988) as a special case and is more formal than the methodology of Zellner (1975). Section 3 gives examples of diagnostics; I propose novel Bayesian checks for missing predictors, goodness-of-fit, and over-diffuseness of the prior in the linear model. The methods and specific checks are applied in section 4 to a random effects model. The paper finishes with a short discussion.

2 Full Prior Predictive Model Checking

A Bayesian model specifies a joint prior distribution $p_0(Y, \theta)$ for the data Y and parameters θ. This also implies a prior distribution $p_0(g)$ for any univariate function $g = g(Y, \theta)$. After observing the data Y, we calculate a posterior $p(\theta|Y) = p_0(Y, \theta)/p_0(Y)$, where $p_0(Y) = \int p_0(Y, \theta) d\theta$. This induces a posterior

$p(g|Y)$ for g. The prior completely specifies the distributional assumptions involved in the analysis; if these assumptions are violated, inference from the model is suspect. Suppose for the moment that $g(Y, \theta)$ is fully observed at g_{obs} and define $P_0(g)$ to be the cumulative distribution function corresponding to the prior density $p_0(g)$. If g_{obs} is in the tails of $P_0(g)$, then doubt is cast on the model $p_0(Y, \theta)$. We can formalize this. Consider the classical test of the hypothesis H_0 that g_{obs} comes from the density $p_0(g)$. We might reject H_0 if $P_0(g_{\text{obs}}) < \delta$, the left tail test where $0 < \delta < 1$ is an appropriately chosen constant. We can similarly handle a right tailed or two tailed test. The smallest δ with which we reject H_0 is the p-value associated with the test. When g is a function of θ as well as Y, it is not fully observed. In this case, I propose to calculate the posterior probability of rejecting H_0.

One practical advantage of the approach just sketched is that automatically, classical residual diagnostics for the linear model have potential use for Bayesian model checking. Consider the linear model $Y = X\beta + \sigma\epsilon$, with X a known $n \times p$ matrix; regression coefficients β; $\epsilon \sim N_n(0, I)$, and prior $p(\beta)$, where $N_n(\mu, \Sigma)$ is an n-dimensional normally distributed random variable with mean μ and covariance matrix Σ. Take σ^2 known to fix ideas; the example in section 4 takes all variance parameters unknown. Define $e = QY/\sigma$, with $Q \equiv (I - X(X^tX)^{-1}X^t)$; a priori e has a known singular $N(0, Q)$ distribution. Apparent disagreements between e and this distribution cast doubt upon the model.

A simple but key example sheds light on the difference between this Bayesian approach and classical residual analysis, and why Bayesian analysis improves on the older classical analysis. Consider the i^{th} element e_i of e. When e_i is far from its prior mean of zero, doubt is cast upon the model. This has traditionally been taken as evidence that the i^{th} observation y_i is outlying. However, since $e_i = Q_i^tY$, where Q_i is the i^{th} column of Q, is a linear combination of the y vector, actually it is evidence that the linear combination, $Q_i^t Y$ and not necessarily y_i, is outlying. This distinction is particularly important when the leverage $h_i = x_i^t (X^t X)^{-1} x_i$ of the i^{th} observation is large.

A Bayesian approach for checking the i^{th} case for outlyingness was given by Chaloner and Brant (1988). To check for outlyingness, interest actually lies in ϵ_i, not e_i. A posteriori, $\epsilon_i|\sigma \sim N(e_i, h_i)$; for high leverage points, this distribution has a large variance, and we are uncertain as to the exact value of ϵ_i and we can't tell if case i is outlying. Chaloner and Brant's outlier diagnostic $P(|\epsilon_i| > z_{1-\delta/2}|Y)$ falls directly in the current framework. It can be interpreted as the posterior probability of rejecting the null hypothesis that ϵ_i has prior mean zero at a level δ, where $z_{1-\delta/2}$ is the $1-\delta/2$ quantile of a standard normal. The Bayesian approach potentially permits classification of cases into not outlying, outlying, and can't tell classes, corresponding respectively to $|\epsilon|$ known small, $|\epsilon|$ known large, and h_i large.

3 Three Diagnostic Measures.

This section illustrates three novel Bayesian diagnostic measures which are specific applications of the methodology developed so far. Possibly due to a lack of practical experience, priors in Bayesian practice are often quite diffuse if not actually improper; subsection 3.1 presents a check for over-diffuseness. The second diagnostic is for the situation when model misspecification is suspected but details are unknown. In this situation, a lack of fit statistic may be useful, and I propose a Bayesian lack of fit check. The third is for when a known covariate has been omitted and we wish to check for the usefulness of adding the covariate to the model. For this section, the model is the linear model of the previous section; I take $p(\beta)$ as $N(\beta_0, \sigma^2 A)$ and continue to condition on σ^2 to fix ideas.

3.1 Over-diffuseness of the Prior

A common conjugate prior for the p regression coefficients is $\beta|\sigma^2 \sim N(\beta_0, \sigma^2 A)$. A priori given σ^2

$$Q_\beta = (\beta - \beta_0)^t A^{-1} (\beta - \beta_0)/\sigma^2 \sim \chi^2(p)$$

Often the eigenvalues of A are taken to be larger than is actually believed, so as to lead to "a conservative inference", or alternately, an inference dominated by the data. However when the eigenvalues of A are overly large, a posteriori, Q_β will be approximately zero, and a posteriori, $P(Q_\beta < \chi^2(p, \delta)|Y)$ will be suspiciously large, even for δ quite small, where $\chi^2(p; \delta)$ is the δ quantile of a chi-square random variable with p degrees of freedom. This suggests that the prior is too diffuse. Another possibility is that the prior mean β_0 or A may have been derived from the data; for example β_0 might be chosen to be equal to the posterior mean of β given a flat prior. In either case the prior is not a representation of true prior belief. An alternative problem is that the prior may have been derived from incorrect information; if $P(Q_\beta > \chi^2(p, 1 - \delta)|Y)$ is large again for small δ, then the prior is refuted by the data.

3.2 Goodness of Fit

Goodness-of-fit statistics assess the fit between the model and data. Poor fit should leave the model with too many outliers; an overfitted model may exhibit too few outliers. Define $1\{|\epsilon_j| > z_{1-\delta/2}\}$, the indicator function that $|\epsilon_j|$ is greater than $z_{1-\delta/2}$; traditional choices are $\delta = .05$ or $.01$. Define

$$\phi(\delta) = \sum_{j=1}^{n} 1\{|\epsilon_j| > z_{1-\delta/2}\}\,.$$

If ϵ were fully observed, ϕ_δ is the number of outliers at the $|z_{\delta/2}|$ level. A priori, $\phi(\delta)$ is distributed Binomial(n, δ). A posteriori, ϕ_δ has support on $0, \ldots, n$. If the posterior distribution of ϕ_δ is on values that had large prior support, then no lack of fit is found by the statistic. If the posterior probability is partially on implausible values a priori, then there is some probability of lack of fit. Finally, if the posterior is entirely on implausible values, then lack of fit is definitely identified. For example, if the posterior probability is high that $\phi(\delta) > n\delta + z_{1-\delta_2}(n\delta(1-\delta))^{1/2}$, for suitable choice of δ_2 such as .05 or .01, the model does not fit the data. If a posteriori, $\phi(\delta) < n\delta - z_{1-\delta_2}(n\delta(1-\delta))^{1/2}$, we might say that the model over-fits the data. In practice we can investigate the entire posterior $p(\phi(\delta)|Y)$, as a simple table can display the entire prior and posterior.

The statistic $\phi(\delta)$ is an omnibus statistic capable of responding to many different potential model failures; it is non-specific and may therefore presumably exhibit moderate ability to detect any one of a wide range of problems. In contrast, if a specific model failure is suspected, then a targeted diagnostic will likely have much greater ability to identify such problems. Examples are the previous prior diffuseness diagnostic or the diagnostics in the next subsection for omitted predictors.

3.3 Omitted Predictors

Let W be a known n by r matrix with columns W_j; W represents a set of covariates not in the regression model. To see if W_j could be a useful addition to the model, it would be helpful to plot W_j directly against ϵ. Since we cannot, we plot W_j against samples from the posterior distribution of ϵ. Dynamic graphics makes this relatively easy. We then summarize the plots qualitatively after viewing many plots or numerically through use of a summary statistic.

A numerical summary of the plot is $\gamma_j = (W_j^t W_j)^{-1/2} W_j^t \epsilon$, a priori distributed $N(0,1)$. Then γ_j is a function of Y, β, and σ suitable for testing whether ϵ and W_j are linearly correlated. Since γ_j is a function of the parameters as well as Y, it has a posterior distribution $\gamma_j|Y,\sigma^2 \sim N(m_j,V_j)$ based on the prior $\beta|\sigma^2 \sim N(\beta_0,\sigma^2 A)$ from subsection 3.1 where

$$m_j = \frac{W_j^t(Y - X\bar{\beta})}{(W_j^t W_j)^{1/2}\sigma}$$

with $\bar{\beta} = E[\beta|Y]$ and

$$V_j = \frac{W_j^t X(X^t X + A)^{-1} X^t W_j}{W_j^t W_j}.$$

We can investigate $p(\gamma_j|Y)$ through appropriate posterior summaries. One summary borrowed from Chaloner and Brant (1988) is the probability $q(W_j, z_{1-\delta/2}) = P(|\gamma_j| > z_{1-\delta/2}|Y)$. The cutoff value $z_{1-\delta/2}$ comes from the $N(0,1)$ prior distribution of γ_j. This is the posterior probability of rejecting the two sided test for $H_0 : \gamma_j = 0$; alternatively, $q(W_j, z_{1-\delta/2})$ is the posterior probability that the vector ϵ is outlying in Euclidian n-space in the direction of W_j. When $W_j = a_i$, the coordinate indicator vector of the i^{th} case, $q(W_j, z_{1-\delta/2})$ is the Chaloner and Brant (1988) posterior probability $E[1\{|\epsilon_j| > z_{1-\delta/2}\}|Y]$.

When the prior for β is flat, $A^{-1} = 0$, then $E[\gamma_j|Y] = W_j^t QY(\sigma^2 W_j^t W_j)^{-1/2}$ is proportional to the classical test statistic $(\sigma^2 W_j^t Q W_j)^{-1/2}(W_j^t QY)$ for testing the coefficient of W_j equal to zero in the regression $Y = X\beta + W_j\alpha + \epsilon$. If $X(X^tX)^{-1}X^tW_j = W_j$, we have $\gamma_j|Y \sim N(0,1)$, the posterior is the same as the prior, and the data do not tell us about whether W_j is correlated with ϵ. When $W_j^t X = 0$, then $\gamma_j = E[\gamma_j|Y,\sigma^2] = (\sigma^2 W_j^t W_j)^{-1/2} W_j^t Y$, there is no uncertainty in our posterior estimate of the test statistic, and we reject H_0 at the level δ that a priori γ_j is $N(0,1)$ if $|\gamma_j| = |(\sigma^2 W_j^t W_j)^{-1/2} W_j^t Y| > z_{1-\delta/2}$.

Sometimes we have more than one predictor W_j we wish to explore for adding to the model. For example, W_2 might be the element-wise square of W_1; or W could include all interactions amongst variables already in the model. In this case we can explore the posterior distribution of $\gamma = (W^tW)^{-1}W^t\epsilon$. A simple summary of this distribution is $\gamma^t(W^tW)\gamma$ which is distributed a priori as a chi-square random variable with r degrees of freedom. We can summarize further using $P(\epsilon^t W(W^tW)^{-1}W^t\epsilon > \chi^2(r; 1-\delta)|Y)$. If the rows of W are a permutation of a r by r identity matrix and an $n-r$ by r matrix of zeros, then we are checking for a r-variate outlier.

4 Weight Loss Data.

Here I illustrate the proposed diagnostics in a hierarchical random effects model (REM) of a repeated measures (RM) weight loss data set. Four different models will be considered and compared.

4.1 The Model and Notation

The basic RM REM is

$$\begin{aligned} Y_i &= X_i\alpha + Z_i\beta_i + \epsilon_i \\ \beta_i &\sim N(0, D), \\ \epsilon_i &\sim N(0, \sigma^2 I), \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$; where $Y_i = (y_{i1}, \ldots, y_{in_i})^t$ is the n_i by 1 vector of repeated measurements on subject i taken at times $t_i = (t_{i1}, \ldots, t_{in_i})^t$; X_i, n_i by p, and Z_i, n_i by q are matrices of known covariates; α is a p by 1 parameter vector of fixed effects; β_i is a q by 1 parameter vector of random effects with $q \leq p$. Except to illustrate the prior diffuseness diagnostic, a flat prior $p(\alpha, \sigma^2, D) \propto 1$ is used.

The Gibbs sampler (Gelfand, Hills, Racine-Poon, and Smith 1990; Zeger and Karim 1991; Gilks, Wang, Yvonnet, and Coursaget 1993) permits straightforward Markov chain Monte Carlo sampling from the posterior of the parameters $\theta = (\alpha, \beta_1, \ldots, \beta_n, D, \sigma^2)$ given the data. I assume that samples $\theta^{(\ell)}, \ell = 1, \ldots, L$ are available from $p(\theta|Y)$. Define $\epsilon = (\epsilon_1^t, \ldots, \epsilon_n^t)^t$ the vector of residuals; then $\epsilon^{(\ell)}$ is a single sample from $p(\epsilon|Y)$. Calculations are based on Gibbs samples of sizes 1000 or 2000.

4.2 Data Description.

The data set contains up to 8 weekly observations per person at times $t_{i1} = 1$ through $t_{i8} = 8$ on $n = 38$ women enrolled in a diet study. Some measurements are missing for a total of 265 individual observations. Initial plots (not shown) of the raw data suggested a random intercept model was appropriate. Model 1 has $X_i = Z_i$ both a vector of n_i ones. Analysis from this random intercept model showed the need for an additional fixed slope. Model 2 has Z_i as in model 1, but X_i has two columns, a column of ones and a column t_i. Analysis of model 2 suggested the need for a random slope. Model 3 has $X_i = Z_i$ with X_i the same as in model 2. Finally, analysis of model 3 shows the population mean at each time does not follow a linear trend, so the final model, model 4, has Z_i as in model 3, but there are 8 parameters in α, so that each week has a different population mean. To summarize, model 1 has $p = 1$, models 2 and 3 have $p = 2$, and model 4 has $p = 8$; models 1 and 2 have $q = 1$, and models 3 and 4 have $q = 2$.

4.3 Over-Diffuseness of the Prior

Consider a prior for the fixed effects α where $\alpha|\sigma^2, D$ is distributed $N(\mu_0, \sigma^2 V)$, with V possibly a function of D. Our test statistic for over-diffuseness of the prior is $Q_\alpha = \sigma^{-2}(\alpha - \mu_0)^t V^{-1}(\alpha - \mu_0)$, with prior distribution, given σ^2 and D, that is χ^2 with p degrees of freedom. Since the prior conditional distribution doesn't depend upon σ^2 or D, the prior distribution unconditional on σ^2 and D is also $\chi^2(p)$. A possible prior for α in model 4 is $\alpha \sim N(\mu_0, \sigma^2 V)$ with $\mu_0^t = (200, 0, 0, 0, 0, 0, 0, 0)$ and V is diagonal with initial element 10000 and remaining

7 diagonal elements 1000. The parameterization of $\alpha = (\alpha_j)$ has α_1 the population mean at time 1, and for $j = 2, \ldots, 8$, α_j is the difference in population mean value at time j minus that at time 1. The posterior mean of α from model 4, based on the flat prior, rounded to the nearest pound is $(193, 1, -2, -5, -6, -5, -6, -8)^t$, with posterior standard deviations ranging from .4 to 1.3 pounds. The posterior probability that $Q_\alpha < \chi^2(8, .01)$ is 1.0, suggesting that either the prior is overly diffuse, or that the prior was chosen using the data. Changing the prior mean to 0 for α_1, and increasing the variance of the first term to 100000 also gives $P(Q_\alpha < \chi^2(8, .01)|Y) = 1.0$, again suggesting an unreasonable prior.

4.4 Goodness of Fit Checks

There are several ways to extend the goodness-of-fit check to multivariate hierarchical data, because there are several different ways to identify outliers. We can consider goodness-of-fit based on $R_{ij} = Y_{ij} - X_i\alpha$ based on a marginal model derived from (1) by integrating out the β_i from the model or we can investigate the hierarchies separately by investigating the ϵ_{ij} residuals and the β_{ik} residuals. It seems preferable to consider the ϵ's and β's separately to permit targeted remediation in case of discovered problems, so I don't consider the R_{ij} further. We can treat the residuals as either univariate or multivariate residuals. Define the following sums of outlier indicator statistics

$$\begin{aligned}
\phi_\epsilon(\delta) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} 1\{|\epsilon_{ij}| > \sigma z_{1-\delta/2}\} \\
\phi_{\beta,k}(\delta) &= \sum_{i=1}^{n} 1\{|\beta_{ik}| > D_{kk}^{1/2} z_{1-\delta/2}\} \\
\Psi_\epsilon(\delta) &= \sum_{i=1}^{n} 1\{\epsilon_i^t \epsilon_i > \sigma^2 \chi^2(n_i, 1-\delta)\} \\
\Psi_\beta(\delta) &= \sum_{i=1}^{n} 1\{\beta_i^t D^{-1} \beta_i > \chi^2(q, 1-\delta)\},
\end{aligned}$$

$p(\Psi_\beta(.05)|Y)$

	0	1	2	3	4	5	6	7	8+
p	.1424	.2848	.2773	.1751	.0807	.0289	.0084	.0020	.0005
1	.0055	.7506	.1464	.0620	.0330	.0015	.0010		
2	.0045	.7521	.1299	.0695	.0360	.0050	.0030		
3	.0175	.3243	.3773	.2064	.0645	.0070	.0030		
4	.0155	.3228	.4098	.1989	.0475	.0040	.0010	.0005	

$p(\Psi_\beta(.01)|Y)$

	0	1	2	3	4	5	6	7	8+
p	.6826	.2620	.0490	.0059	.00053	3.6e-05	2.0e-06	9.2e-08	3.7e-09
1	.2464	.7521	.0015						
2	.1924	.8061	.0015						
3	.3333	.5612	.0960	.0095					
4	.3713	.5257	.0975	.0050	.0005				

$p(\Psi_\epsilon(.05)|Y)$

	0	1	2	3	4	5	6	7	8+
p	.1424	.2848	.2773	.1751	.0807	.0289	.0084	.0020	.0005
1	0	0	0	.019	.610	.309	.054	.007	.001
2	0	0	.001	.009	.093	.585	.255	.050	.007
3	0	0	.078	.285	.284	.209	.096	.034	.013
4	0	.013	.164	.302	.279	.162	.071	.009	

$p(\Psi_\epsilon(.01)|Y)$

	0	1	2	3	4	5	6	7	8+
p	.6826	.2620	.0490	.0059	.0005	3.6e-05	2.0e-06	9.2e-08	3.7e-09
1	0	.003	.036	.639	.305	.017			
2	0	.081	.263	.222	.329	.099	.006		
3	.018	.052	.650	.234	.041	.005			
4	.019	.364	.477	.124	.014	.002			

Table 1: Goodness of fit statistics. The symbol p indicates the prior distribution of the number of outliers, rows beginning 1, 2, 3, or 4 are posterior distributions of the number of outliers conditional on that model. The β and ϵ posteriors are based on Gibbs samples of size 2000 and 1000 respectively.

where D_{kk} is the k^{th} diagonal element of D, and δ is the probability content in the tail where an observation is declared an outlier. The ϕ statistics treat the ϵ's and β's univariately, and the Ψ's treat them multivariately. If $q = 1$ then $\phi_{\beta,1}(\delta) = \Psi_{\beta}(\delta)$. Each of these statistics leads to a different goodness of fit statistic. The multivariate $\Psi_{\epsilon}(\delta)$ and $\Psi_{\beta}(\delta)$ get at the multivariate relationships among the ϵ or β in a way that the univariate $\phi_{\beta}(\delta)$ and $\phi_{\epsilon}(\delta)$ do not.

All of the goodness of fit statistics were calculated for $\delta = .01$ and $\delta = .05$ for all four models. The first two sections of Table 1 check for an excess of multivariate β outliers at $\delta = .05$ and $\delta = .01$. The row labeled p gives the prior probability mass function of the number of outliers; this is calculated using the binomial($n = 38, \pi = \delta$) distribution, with $\delta = .05$ or $.01$. For example with $\delta = .01$, the prior probability of zero, one or two β outliers is .6826, .2620, and .0490 respectively. We see that models 1 and 2 have most of their probability mass on either zero or at most one β outlier. Models 3 and 4 have approximately 10% chance of having 2 outliers, but still have almost 90% of their probability on either zero or one β outliers. The number of outliers is not unusual, and we do not flag the β's as exhibiting any lack of fit. Other goodness of fit statistics involving the β's were similar in not showing any lack of fit.

The second two sections of table 1 check for an excess of multivariate ϵ outliers, also at tail areas $\delta = .05$ and $\delta = .01$. Inspection of $\Psi_{\epsilon}(.05)$ suggests that there are somewhat more ϵ outliers than expected a priori, especially for models 1 and 2, with model 2 being slightly worse than model 1. The results for $\Psi_{\epsilon}(.01)$ are clearer; the number of outliers appears to be largest with model 1 and least with model 4. For model 1, the prior probability of three or more outliers is less than 1%, while the posterior probability is over 95%. For model 2 it is still 65% while for models 3 and 4 the percentage drops to 28% and 14% respectively of three or more outliers. This suggests substantial lack of fit for model 1 that is improved as we move from model 1 to model 4, but there may still be some lack of fit even in model 4. Because the mean structure is as general as this data structure is capable of supporting, any inappropriateness in the model must be in the covariance structure or in the choice of the normal distribution. A reasonable further expansion of the model would be to a general covariance structure from the random effects model.

Each goodness of fit statistic has a quantile-quantile (QQ) plot associated with it. For example, for $\phi_\epsilon(\delta)$, we could draw a single sample $\epsilon^\ell/\sigma^{(\ell)}$ and look at a QQ plot against quantiles of the normal distribution. Since this is only a single sample, we would take several samples, and construct several QQ plots. For $\Psi_\beta(\delta)$, we can either investigate a $\chi^2(q)$ QQ plot of samples from the posterior of $(\beta_i^{(\ell)})^t(D^{(\ell)})^{-1}\beta_i^{(\ell)}$, or we can transform to normality by plotting $\Phi^{-1}(F_{\chi^2,q}((\beta_i^{(\ell)})^t(D^{(\ell)})^{-1}\beta_i^{(\ell)}))$ against quantiles of the normal distribution, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (cdf) of the standard normal distribution and $F_{\chi^2,q}$ is the cdf for the χ^2 distribution with q degrees of freedom. For $\Psi_\epsilon(\delta)$, we also map to the standard normal but replace the q degrees of freedom by n_i degrees of freedom. Thus we plot ordered values of $\Phi^{-1}(F_{\chi^2,n_i}(\epsilon_i^t\epsilon_i/\sigma^2)))$ against order statistics from a standard normal distribution.

Figure 1a shows a representative QQ plot of $\beta_i/D^{1/2}$ for model 1. Since $q=1$ for model 1 and 2. In figure 1a, we see a possible single outlier at the upper right of the plot. Figure 1b is also a single representative of several plots. We see that the transformed $\epsilon_i^t\epsilon/\sigma^2$ appear to have several very large outliers making up for a host of generally too small outliers at the bottom left. Recall that before transformation, the points at the bottom left of figure 1b corresponded to $\chi^2(n_i)$ quantiles near zero. We see that the observations are not distributed like a $N(0,1)$, and thus, that the model does not fit.

4.5 Missing Fixed Effects Predictors

Here we consider diagnostics to check for particular missing univariate predictors. Let U_j be a vector the same length as ϵ. For the weight loss study, I consider four predictors; the first $U_1 = (t_1, \ldots, t_n)^t$ is the vector of times that individual observations are taken. To get the second, third and fourth predictors, consider the elementwise square U_1^2, cube U_1^3 and fourth power U_1^4 of U_1; let $U_1^1 = U_1$ and define U^0 be the vector of ones. Regress U_1^j for $j = 1, 2, 3, 4$ on all lower order powers of U_1. Define W_j to be the residuals from each of these four regressions standardized to have length 1. Then $W_j^t W_j = 1$, and $W_{j'}^t W_j = 0$ if $j' \neq j$ and the W_j are an orthonormal basis of a four dimensional subspace of 265 dimensional Euclidean space. If ϵ/σ is a sample from a $N(0, I)$, then the $\gamma_j = W_j^t \epsilon/\sigma$ should also behave like a single draw from a $N(0, 1)$. If W_j is in the span of the columns of $X = (X_1^t, \ldots, X_n^t)^t$, then approximately,a posteriori, we might expect $W_j^t \epsilon \sigma^{-1} \sim N(0, 1)$, and the posterior expected value of $\epsilon^t W_j W_j^t \epsilon$ should be approximately one and the posterior probability that the contrast is an outlier should be equal to the prior probability.

I then calculated the posterior distributions of γ_j, for all four models, where $j = 1, 2, 3, 4$ represent the linear through quartic effects respectively. Effects through the quartic in time were chosen because prior information about the design and additional graphical diagnostics not shown here suggested that a quartic effect might reasonably be anticipated. Table 2 summarizes the results. For model 1, the linear contrast was an enormous outlier, with $q(W_1, 3) = 1$, and a posterior expected mean square $\mathrm{E}[\epsilon^t W (W^t W)^{-1} W^t \epsilon | Y]$ of 108.2. This indicates that model 1 is missing a linear time fixed effect and that the effect is quite large. As expected, for models 2, 3, and 4, the posterior probabilities are approximately equal to the prior probabilities that the linear effect is an outlier. In contrast, the quartic effect contrast is a strong outlier for models 1, 2, and 3. After removing the linear trends from the residual, the W_2 contrast is a moderate outlier for

effect	model	mean square	$q(W_j,2)$	$q(W_j,3)$
linear				
	1	108.2	1	1
	2	.95	.038	.002
	3	1.00	.049	.002
	4	1.05	.057	.004
quadratic				
	1	3.82	.362	0
	2	5.71	.987	0
	3	9.54	.999	.582
	4	.96	.044	.002
cubic				
	1	.64	0	0
	2	1.43	0	0
	3	2.21	.010	0
	4	.91	.031	.002
quartic				
	1	14.8	1	1
	2	28.73	1	1
	3	47.79	1	1
	4	.99	.050	.005

Table 2: Checks for needed polynomial effects. The first column indicates an effect linear, quadratic, cubic or quartic in time. The second column indicates model 1, 2, 3, or 4. The mean square is $E[(W^t\epsilon)^2/(\sigma^2)|Y]$, which should be approximately 1 if W_j has already been included in the model, and approximately $\chi^2(1)$ if the model is well specified and W_j is orthogonal to any predictors in the model. The next two columns are $q(W_j,2) = P(|W^t\epsilon| > 2\sigma|Y)$ and $q(W_j,3) = P(|W^t\epsilon| > 3\sigma|Y)$.

model 2 and more so for model 3; only for model 4 is it not an outlier. The cubic effect is apparently not an outlier for any model.

5 Discussion

With the current approach, any function of the parameters and data can be used to check the model. The challenge is to choose useful functions for model checking. With our approach, classical residual checks and hypothesis tests with asymptotic but no exact results are approximate Bayes checks and tests, but now these asymptotic results have a small sample distribution with which the asymptotic results can be compared.

References

Albert, J. H. and Chib S. (1996). Bayesian residual analysis for binary response regression models. *Biometrika*, 82, 747-759.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modeling and robustness (C/R: p404-430). *Journal of the Royal Statistical Society*, Ser. A, 143, 383-430.

Chaloner, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika*, 78, 637-644.

Chaloner, K. (1994). Residual analysis and outliers in Bayesian hierarchical models. In *Aspects of Uncertainty*, (eds) P. R. Freeman and A. F. M. Smith. New York: John Wiley & Sons, 149-157.

Chaloner, K. and Brant R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 75, 651-660.

Dey, D. K., Gelfand, A. E., Schwarz, T. B. and Vlachos, P. K. (1994). Simulation based model checking for hierarchical models. Technical Report, University of Connecticut, Department of Statistics.

Gelman, A. Meng X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, to appear.

Hodges, J. (1994). Some algebra and geometry for hierarchical models, applied to diagnostics. University of Minnesota, Division of Biostatistics, Technical Report 94-009.

Meng, X.L. (1994). Posterior predictive p-values. *Annals of Statistics*, 22, 1142-1160.

Zellner, A. (1975). Bayesian analysis of regression error terms. *Journal of the American Statistical Association*, 70, 138-144.

Figure Caption

Figure 1. (a) A QQ plot of $\beta_i/D^{1/2}$ for model 1. (b) A QQ plot of the transformed multivariate $\epsilon_i^t\epsilon/\sigma^2$ residuals.