

# The Wages of BMI: Bayesian Analysis of a Skewed Treatment-Response Model with Nonparametric Endogeneity

Brendan Kline  
Northwestern University  
b-kline@northwestern.edu

Justin L. Tobias<sup>1</sup>  
Department of Economics  
Iowa State University  
tobiasj@iastate.edu

September 2007

---

## Abstract

We generalize the specifications used in previous studies of the effect of body mass index (BMI) on earnings by allowing the potentially endogenous BMI variable to enter the log wage equation nonparametrically. We introduce a Bayesian posterior simulator for fitting our model that permits a nonparametric treatment of the endogenous BMI variable, flexibly accommodates skew in the BMI distribution, and whose implementation requires only Gibbs steps. Using data from the 1970 British Cohort Study, our results indicate the presence of nonlinearities in the relationships between BMI and log wages that differ across men and women, and also suggest the importance of unobserved confounding for our sample of males.

---

<sup>1</sup>We thank and acknowledge the UK Data Archive (University of Essex, Colchester, UK) for use of data from the 1970 British Cohort Study. They bear no responsibility for the analysis or interpretation of this data. All errors are, of course, our own.

# 1 Introduction

In this paper we investigate the role of body mass index (henceforth, BMI)<sup>2</sup> in the production of log wages. Given sufficient data, the simplest approach to study this issue would proceed by running a regression of the outcome, log wages, on the BMI variable and potentially a variety of controls. The resulting coefficient on the BMI variable, however, could not be convincingly argued to capture any type of “causal” impact, since it fails to control for the potential endogeneity of BMI. By “endogeneity” we mean a failure of the methodology to successfully account for factors unobserved by the econometrician which simultaneously correlate with BMI and log wages. One such confounding variable could be preferences for long-term investments, which we mean to represent characteristics that simultaneously impact decisions affecting both health and human capital accumulation. The existence of such unobserved confounding variables produces a bias and inconsistency in the simple regression-based estimator described above, requiring us to employ a more elaborate procedure that accounts for potential confounding on unobserved characteristics. To this end, we adopt in this paper a standard two equation triangular treatment-response model in which the outcome of interest is log wages and the endogenous treatment variable is BMI.

The primary methodological innovation of our approach is that we permit the BMI variable to enter the wage equation nonparametrically. The default specification in applied treatment-response modeling assumes a linear relationship between the treatment and the outcome, and defines the slope of this line as the causal effect of interest. This linearity assumption, however, is not credible in all situations. In our study, for example, it is plausible that wages are relatively unresponsive to marginal changes in BMI in the “underweight” or “normal” ranges, and relatively more responsive to marginal changes in BMI in the “overweight” or “obese” ranges. Alternatively, underweight individuals may experience a wage penalty similar to overweight or obese individuals, producing an inverted *U*-shaped relationship between BMI and log wages. Standard linear treatment-response models can not capture these features of the data if they are present. Although a linear relationship might be found

---

<sup>2</sup>BMI is computed as weight (in kilograms) divided by height (in meters) squared. The equivalent calculation in terms of pounds and inches is  $704.5 \times \text{weight (in pounds)} / \text{height (in inches)}^2$ . Standard clinical classifications of BMI are “underweight” ( $\text{BMI} < 18.5$ ), “normal” ( $18.5 \leq \text{BMI} < 25$ ), “overweight” ( $25 \leq \text{BMI} < 30$ ), and “obese” ( $30 \geq \text{BMI}$ ). It is worth mentioning that BMI may not be an ideal measure of health for all people since, for example, some very muscular people have high BMI values that reflect differences in the weight of muscle tissue and lean tissue, and not health. Finding the “best” measure of body fatness is something that the entire literature must contend with, as there is a tradeoff between methods that may provide a more accurate measure (e.g. underwater weighing or dual-energy x-ray absorptiometry) and those that can be applied to a large sample, cheaply (e.g. BMI). Prentice and Jebb (2001) point out some of the flaws of body mass index, while Deurenberg, Weststrate and Seidell (1991) in their Table 1 report that the correlation between body mass index and body fat, measured by underwater weighing, for adults aged 26-35 is .92 (men) and .89 (women).

*ex post* by “testing down” from more flexible specifications, it does not seem desirable to impose these restrictions *ex ante*. To our knowledge, no applied studies in this area have investigated this possibility extensively. Therefore, we hope that use of a more flexible empirical specification which can either refute or provide support for the specifications used in past work offers a useful contribution to this literature.

A second, yet more minor, methodological contribution is that we introduce a model capable of accommodating skew in the distribution of the continuous endogenous treatment variable. This particular modeling assumption is made with an eye toward our empirical application, where the unconditional and conditional BMI distributions have a pronounced right-skew. Our approach for handling this skew is to generalize the error distribution associated with the treatment variable to the class of skew-normal type distributions [e.g., Azzalini and Dalla Valle (1996), Branco and Dey (2002)]. We find that use of the skew-normal distribution provides an adequate fit to our data, outperforms the standard semi-log specification, and is parsimonious as it introduces only one additional parameter. In addition, the methods described here can be easily generalized to handle cases where there is skew in the outcome variable, skew in both the outcome and the endogenous treatment variable, and when the variables being modeled are censored, binary or ordered.<sup>3</sup> In all cases, however, careful diagnostic checking should be performed to assess if the maintained distributional assumptions are sufficiently flexible. We fit the model from a Bayesian perspective and present a computationally attractive posterior simulator which handles the endogeneity problem, skew, and nonparametric component simultaneously, and involves only standard Gibbs steps.

Studies like this one on the effect of “weight-for-height”<sup>4</sup> on economic outcomes must contend with observational data. One approach that has been used to identify a causal effect with such data is differencing, within twins or siblings [e.g., Averett and Korenman (1993), Behrman and Rosenzweig (2001), and Baum and Ford (2004)] or over time with panel data [e.g., Baum and Ford (2004), and Cawley (2004)], to purge the model of unobserved confounding variables. These approaches require within-sibling variation in weight-for-height or temporal variation in weight-for-height to identify the effects of interest. Since such variation may not be substantial, particularly in longitudinal data studies with a short time dimension, it may often prove difficult to precisely estimate effects of interest with such an approach. Alternatively, some studies have made use of instrumental variables [e.g., Behrman and Rosenzweig (2001), and Cawley (2004)] to identify a causal effect.

We follow the instrumental variables approach and use parent BMI as a source of exogenous

---

<sup>3</sup>See, for example, Koop, Poirier and Tobias (2007, chapter 14).

<sup>4</sup>That is, BMI as a continuous variable or obesity as a binary variable.

variation. The validity of this IV strategy hinges on the assumptions that parent BMI is conditionally correlated with the child's adult BMI, and that parent BMI can be excluded from the child's adult log wage equation. The first assumption is not controversial (and is empirically testable), but the second assumption may be met with more skepticism. In particular, one could argue that the correlation between parent BMI and child BMI is due partly to learned preferences about the importance of long-run investments in health capital, which may be correlated with learned preferences regarding investments in human capital, which in turn are known to have a direct effect on earnings. In this case, parental BMI may embody unobserved factors that have a structural effect on the child's adult earnings, potentially undermining the validity of our identification strategy.

Despite this example of why our instrumental variable strategy may fail, we offer three reasons why such concerns should not necessarily invalidate the use of parental BMI as valid instruments. First, and perhaps most importantly, the validity of the instruments relies on their *conditional* uncorrelation with the errors, and thus the inclusion of a rich set of proxies for family preferences and background characteristics will surely bolster the case for the use of parental BMI in practice. Specifically, the set of variables we include in the wage equation should reduce the problem a "shared family environment" mechanism may present, as they serve as proxies for family preferences, even if this problem is likely to exist unconditionally (i.e., without any controls). To this end, we include in the wage equation covariates capturing aspects of environment and background including indicators for parental education, parental occupation, and parental income when the worker was ten years old.

Second, we appeal to studies in the health literature which suggest that the "learning" mechanism in determining adult BMI outcomes is relatively weak. Stunkard et al (1986), for example, obtain a sample of information from Danish adoptees, and find a strong correlation between the BMIs of adoptive children and their biological parents, yet find no correlation between weight classifications of adoptees and their adoptive parents. In a review of several studies of the determinants of BMI, Maes, Neale and Daves (1997) similarly conclude that "... it is unlikely that environmental factors shared with family members contribute substantially to variance in BMI" (pp. 329-330). In another review, Grilo and Pogue-Geile (1991) write "... this suggests little environmental effect of parental weight on offspring weight through modeling" (pg. 534). Thus, these studies provide suggestive evidence that the correlation between parent and child BMI may have little to do with shared environmental factors, and therefore, parental BMI can potentially be excluded from the wage equation, especially given adequate controls for family characteristics and background characteristics of the child.

Finally, since we have two instruments available<sup>5</sup> we can conduct the standard exercise of assuming that one parent’s BMI is a valid instrument, including the other in the log wage equation, and calculating the Bayes factor in favor of the hypothesis that the other parent’s BMI coefficient in the log wage equation is equal to zero. From a theoretical perspective, however, it seems reasonable that the BMIs of the parents are either jointly valid as instruments, or jointly invalid, thus potentially calling into question what is actually learned from this procedure. On the empirical side, however, the correlation between parental BMI was found to be reasonably small (around .16), suggesting that something can be learned from this exercise, and that its implementation is not obviously redundant or “circular.” For both men and women, these Bayes factors are found to support the restricted model, providing evidence that parent BMI can be excluded from the log wage equation, and informal evidence of the validity of our identification strategy.

Using data from the 1970 British Cohort Study, we apply our estimation algorithm and find strong evidence that BMI affects log wages. We find that log wages are decreasing throughout the BMI support, and that this result holds for both men and women. For men, the wage penalty to marginal increases in BMI is modest provided the individual is in the “normal” BMI range, whereas penalties are comparably large for overweight and obese men. For women the results are essentially reversed. The largest penalties for a marginal increase in a woman’s BMI are found over the “normal” BMI range. In addition to these nonlinearities within gender groups, we find evidence of differential BMI wage penalties across these groups. We find some evidence that for individuals with BMIs in the alternatively defined normal range of 20-25 the wage penalty for a marginal increase in BMI is smaller for men than it is for women. Overweight and obese men, however, receive substantially larger wage penalties to marginal increases in BMI than comparably overweight and obese women. To our knowledge, these results have not been documented before in the literature. These findings also raise questions about the credibility of the assumption of linearity between log wages and BMI that has been made in past work.

The outline of this paper is as follows. In the following section we describe our model, strategy for handling the nonparametric component, skew, and endogeneity issues, and the associated Bayesian posterior simulator. Section 3 describes our data, while section 4 presents our empirical results. The paper concludes with a summary in Section 5.

---

<sup>5</sup>We have two instruments: mother’s BMI and father’s BMI.

## 2 The Model

In our empirical investigation we are primarily interested in determining if the impact of BMI on log wages is linear, and seek to do so within a framework that permits the potential endogeneity of BMI. A useful two-equation system which serves as a starting point for handling these issues is given as follows:<sup>6</sup>

$$y_i = f(s_i) + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \quad (1)$$

$$s_i = \mathbf{z}_i\boldsymbol{\theta} + u_i, \quad (2)$$

where

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \Big| \mathbf{x}_i, \mathbf{z}_i \stackrel{iid}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{pmatrix} \right], \quad i = 1, 2, \dots, n.$$

In the above,  $s$  is a potentially endogenous variable, and the endogeneity (or unobserved confounding) problem is handled by introducing a possible correlation between  $u$  and  $\epsilon$ , denoted  $\rho_{\epsilon u} \equiv \sigma_{\epsilon u} / (\sqrt{\sigma_u^2 \sigma_\epsilon^2})$ . That is, unobservables which are correlated with the process generating BMI (denoted by  $s$  above) may also be correlated with unobservables affecting the production of log wages (denoted as  $y$  above).

Again, it is important to recognize that many applied studies in the treatment-response literature, and to our knowledge all of those that have been conducted on this specific topic, *assume* the relationship between the treatment variable and the outcome variable is linear (i.e.,  $f(s) = \alpha_0 + \alpha_1 s$ ), and define the slope of this function as the causal effect of interest. The assumption of linearity is likely made on computational considerations, as IV is simple to use in this context. In our study, and upon reflection, in many studies in the treatment-response literature, the relationship between the treatment variable and the outcome may not be linear, and may not be easily characterized *ex ante* by a particular parametric form. This motivates the value of specifications like (1) and (2) that allow for a flexible treatment of the endogenous variable.

Despite this added flexibility and generalization of existing empirical work on this topic, it is important to acknowledge that equations (1) and (2) are still restrictive in many aspects. First, we maintain the textbook assumption of joint normality, the credibility of which must largely be determined on an application-specific basis. This assumption, however, can be generalized using, among other possibilities, finite Gaussian mixtures to flexibly accommodate the distributions associated with the treatment and outcome variables. Second, although we allow treatment effects to be quite heterogeneous in the sense that  $f'(s)$

---

<sup>6</sup>We adopt the conventions of using bold script to denote vector or matrix quantities, and capital letters to denote matrices.

can vary at each point in the BMI support,<sup>7</sup> we still maintain treatment effect homogeneity within a given BMI cell. Thus, our model is one that allows for a degree of treatment effect heterogeneity through specific observables, but does not explicitly account for other aspects of heterogeneity, both unobserved and (potentially) observed.

If such treatment effect heterogeneity is present and the treatment variable is binary or discrete, Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996) and Heckman and Vytlacil (2006), for example, show that care must be given in the interpretation of the IV estimate. Specifically, in the binary treatment, binary outcome case, linear IV can be rationalized as a treatment effect for “compliers,” i.e., those individuals whose behavior can be manipulated by the availability of the instrument [Imbens and Angrist (1994)]. In more general settings, linear IV can be shown to converge to a weighted average of treatment effects [e.g., Heckman and Vytlacil (2006)]. In cases where the outcome and endogenous variables are continuous, however, (as in our application), and treatment effect heterogeneity is present, Heckman and Vytlacil (1998) and Wooldridge (2003) provide conditions under which standard linear IV / 2SLS can still consistently estimate the average treatment effect (ATE) in the population.<sup>8</sup> To map our model in (1) and (2) into a specification where the results of these studies could potentially be applied seems to require the creation of dummy variables for each specific BMI value. This construction, however, would lead us back to the case of a continuous outcome model with multiple *binary* treatment variables, whence the assumptions in Wooldridge (2003) would not directly apply, and therefore we can not appeal to these results to interpret our estimates as capturing ATE within each cell.

In the current paper we maintain the assumption of treatment effect homogeneity within our narrow cells, though, admittedly, this assumption may not be perfectly satisfactory in the context of our empirical application. For example, it is possible that one physically fit male in our sample possesses a BMI of 28 (and is thus inappropriately categorized as “overweight”), while another male in our sample has an identical BMI and is correctly classified as overweight. We should not necessarily expect the slope of the log wage- BMI relationship to be the same for these two individuals, yet our model, owing primarily to data constraints and limitations with the construction and interpretation of BMI, imposes equal treatment effects for these agents. Other sources of treatment effect heterogeneity, such as differential impacts across industries, are also possible, but are not explicitly captured in (1) and (2). Our view is that the assumption of homogenous treatment effects within age, gender and

---

<sup>7</sup>Actually, in our empirical work, we also allow these functions to differ across men and women, and given the nature of our sample, these are also conditioned on a cohort of a particular age.

<sup>8</sup>The additional assumption imposes that the covariance between unobserved components of treatment effect heterogeneity and the treatment variable itself does not depend on the exogenous variables of the model.

BMI cells, though still somewhat restrictive, substantially generalizes those made in earlier work on this topic, and thus offers a big step in a more general direction. What we offer in this paper is a framework for flexibly exploring the relationship between an endogenous treatment variable and an outcome of interest within the assumption of (conditional) treatment effect homogeneity, and seek to apply this methodology to characterize the relationship between BMI and log wages.<sup>9</sup> In the case where our model is misspecified and treatment effect heterogeneity persists within these cells, care must be taken with respect to the interpretation of our results; they should not necessarily be interpreted as capturing a population average treatment effect within the cell, but rather, a weighted average of treatment effects whose value may or may not be of direct policy-relevance.<sup>10</sup>

## 2.1 Bayesian Implementation

Though the model described in (1) and (2) offers a useful starting point, the textbook joint normality assumption made there turns out to be inappropriate for our empirical investigation. That is, we find that the unconditional distribution of BMI in our sample is quite skewed in general, and that this skew is particularly pronounced for females. Although the normality assumption made for the error terms posits the shape of the *conditional* rather than the *unconditional* distribution, diagnostic checks revealed that normality was not appropriate, and moreover, that such skew persists even upon taking the log transformation of BMI and defining  $s$  accordingly.<sup>11</sup> To this end, we introduce a generalized assumption regarding the error structure that, though tailored to accommodate specific features of our application, will also be useful for other empirical analyses where the variables of interest are skewed.<sup>12</sup> Specifically, we generalize (1) and (2) by writing

$$\begin{aligned} y_i &= f(s_i) + \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i \\ s_i &= \mathbf{z}_i\boldsymbol{\theta} + \delta\tilde{h}_i^* + u_i, \end{aligned}$$

where

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \Big| \mathbf{x}_i, \mathbf{z}_i, \tilde{h}_i^* \stackrel{iid}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{pmatrix} \right] \quad i = 1, 2, \dots, n,$$

---

<sup>9</sup>Our study is certainly not the first to generalize the standard treatment-response model in a Bayesian framework and apply this methodology to applications of substantive interest. For example, in a pair of interesting papers with applications to health economics, Deb, Munkin and Trivedi (2006) derive a posterior simulator for a two-part model with a multinomial endogenous variable, while Munkin and Trivedi (2006) derive a related simulator with an ordered outcome. We continue in this tradition of generalizing the standard framework by considering a nonparametric treatment of the endogenous variable, and allow for possible skew.

<sup>10</sup>See Heckman and Vytlacil (2006) for more on this issue. Manning (2004) is a particularly readable reference on these general issues, as is Angrist (2004).

<sup>11</sup>Evidence of this is presented in the analysis of section 4.

<sup>12</sup>For example,  $s$  could be modeled in logarithmic form and  $\delta\tilde{h}_i^*$  simultaneously added to the disturbance term to accommodate excess skew.



$$\begin{aligned}\tilde{h}_i^*|\lambda_i &\stackrel{ind}{\sim} TN_{[0,\infty)}(0, \lambda_i), \quad i = 1, 2, \dots, n, \\ \lambda_i &\stackrel{iid}{\sim} IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right), \quad i = 1, 2, \dots, n,\end{aligned}$$

$TN_{(a,b)}(\mu, \sigma^2)$  denotes a normal distribution with mean  $\mu$  and variance  $\sigma^2$  truncated to the interval  $(a, b)$ , and  $IG(\cdot, \cdot)$  denotes an inverse gamma distribution [see Koop, Poirier and Tobias (2007, pp. 335-341) for parameterizations]. The specification above looks like the textbook Gaussian model of (1) and (2) apart from the appearance of the terms  $\tilde{h}_i^*$  in the equation generating the endogenous variable  $s$ . Indeed, it is the inclusion of the latent  $\tilde{h}_i^*$  that enables our model to accommodate the skew of the BMI variable, as we describe in more detail below.

The construction of the equation generating the endogenous variable  $s$  follows similarly to the formulation of the binary choice model of Chen, Dey and Shao (1999). In that paper the authors choose  $\tilde{h}_i^*$  to be *half-normal* (i.e., they set  $\lambda_i = 1$ ) and show that when integrating the conditional distribution of  $s|\tilde{h}$  over the half-normal prior for  $\tilde{h}$ , the resulting marginal distribution of  $s$  follows a *skew-normal* distribution [e.g., Azzalini and Dalla-Valle (1996)] with the parameter  $\delta$  governing the direction of the skew. Specifically, when  $\delta = 0$  we revert back to the textbook Gaussian model, while  $\delta > 0$  yields a marginal density for  $s$  with a right-skew, and  $\delta < 0$  yields a marginal density with a left-skew.

Our version of the model is similar, and  $\delta$  continues to retain the interpretation as a “skewness parameter,” but is introduced at a slightly more general level. Specifically, we do not restrict the variance of the truncated normal for  $\tilde{h}_i^*$  to unity (which produces the half-normal specification), but instead add a mixing variable  $\lambda_i$  in the variance function. A particular inverse gamma prior for  $\lambda_i$  is then employed for the mixing variables  $\lambda_i$ . It is well-known and easily demonstrated that, when integrating the mixing variables  $\lambda_i$  out of the conditional distribution  $\tilde{h}_i^*|\lambda_i$ , the the resulting marginal distribution will be (truncated) Student-t and not half-normal. Thus, our formulation of the model is one that can accommodate skew through different values of  $\delta$ , and can also accommodate variation in tail thickness through different choices of the degrees of freedom parameter  $\nu$  in the hierarchical prior for  $\lambda_i$ . While  $\nu$  could be added as a parameter of the posterior distribution in principle, and thus one could learn simultaneously about skew and tail thickness, in practice, we have found that the addition of  $\nu$  to the sampler tends to slow convergence. Thus we simply select  $\nu = 8$  *a priori*, which seems to perform well in the context of our empirical application. In other applications, alternate values of  $\nu$  could be explored, and the value which is deemed to provide the best fit to the data could be employed in practice.<sup>13</sup>

---

<sup>13</sup>Alternatively, marginal likelihoods could be calculated at alternate  $\nu$  values to either average over various specifications or to select a particular model.

As pointed out by Pewsey (2000) and others, a “problem” associated with this parameterization of the model is that the error terms  $\tilde{h}_i^*$  are not mean zero. Thus, alternate values of  $\delta$  affect both the skew of the distribution of  $s$  and at the same time alter the mean of  $s$  (and therefore affect the intercept’s value). In practice, this redundancy can lead to slow mixing of the posterior simulator and also confounds the interpretation of the model parameters. To resolve this issue, we re-center the latent variables  $\tilde{h}_i^*$  to have mean zero. To this end, first note that the truncated normal specification for  $\tilde{h}_i^*$  implies

$$E(\tilde{h}_i^* | \lambda_i) = \sqrt{\frac{2\lambda_i}{\pi}}$$

and thus, by iterated expectations,<sup>14</sup>

$$E(\tilde{h}_i^*) = \sqrt{\frac{2}{\pi}} E(\sqrt{\lambda_i}) = \sqrt{\frac{\nu}{\pi} \frac{\Gamma[(\nu - 1)/2]}{\Gamma[\nu/2]}} \equiv c_\nu.$$

We can then define a new truncated latent variable  $h_i^* \equiv \tilde{h}_i^* - c_\nu$ . Performing a change of variable, we can then re-formulate the model as follows:

$$y_i = f(s_i) + \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \quad (3)$$

$$s_i = \mathbf{z}_i \boldsymbol{\theta} + \delta h_i^* + u_i, \quad (4)$$

where

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \left| \mathbf{x}_i, \mathbf{z}_i, h_i^* \right. \stackrel{iid}{\sim} N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{pmatrix} \right] \quad i = 1, 2, \dots, n,$$

$$h_i^* | \lambda_i \stackrel{ind}{\sim} TN_{[-c_\nu, \infty)}(-c_\nu, \lambda_i), \quad i = 1, 2, \dots, n, \quad (5)$$

$$\lambda_i \stackrel{iid}{\sim} IG \left( \frac{\nu}{2}, \frac{2}{\nu} \right), \quad i = 1, 2, \dots, n, \quad (6)$$

so that the latent  $h_i^*$  have mean-zero and thus  $\delta$  does not enter the (unconditional) mean function of  $s$ , fixing the interpretation of  $\delta$ .

It is also important to mention that the model given in (3)-(4) is not the only alternative that one could take. Our motivation for employing the model in (3)-(4) is that it is simple, intuitive, and parsimonious, and offers a useful generalization of the textbook model to account for skew. However, we could account for such skew via the use of finite Gaussian mixtures, among other possibilities. Indeed, Gaussian mixtures are far more flexible than the distributional assumptions entertained here and could capture, among other features, multimodality in the error distributions. However, the flexibility afforded by Gaussian mixtures could be argued to come at a cost; they can typically be parameter-rich, though one

---

<sup>14</sup>The mean  $E(\sqrt{\lambda})$  can be obtained by performing the necessary integration with respect to the inverse gamma density, and making a change of variable to represent the integral in terms of the gamma function.

can get around this problem by restricting all parameters other than the intercepts to be equal across mixture components. In addition, the use of Gaussian mixtures typically requires model selection methods or the adoption of reversible jump samplers to determine the number of mixture components, and often requires restrictions to separately identify the individual mixture components. These restrictions are not innocuous in practice, and this problem of component identification (or label-switching) has received considerable attention in the literature. Stressing the importance of this issue, Celeux, Hurn and Robert (2000, pg. 957) write that “almost the entirety of MCMC samplers implemented for mixture models has failed to converge” owing to the label-switching problem. Geweke (2007), however, provides a more moderate view regarding the use of Gaussian mixtures, and stresses that some parameters of interest, such as predictive densities, are not subject to the label-switching issue.

The model in (3)-(4) is not subject to many of these concerns; there are no analogous identification issues to worry about and model selection can be easily carried out.<sup>15</sup> Furthermore, and much like a Gaussian mixture model with a known number of mixture components, a posterior simulator for fitting this model is rather simple to implement in practice. However, this model surely does not enjoy the flexibility of Gaussian mixtures or other robust alternatives, but is a simpler and seemingly valuable alternative when the error distributions are unimodal and accounting for skew is the salient problem. What is required in practice, of course, is an application-specific verification that the specification in (3)-(4), or straightforward generalizations of it, is adequately flexible for the problem at hand.<sup>16</sup> We will perform such diagnostic checking in the context of our specific application in section 4.1. Our view is that the model in (3)-(4), though making no claims to be fully general, offers a simple and intuitive generalization of the textbook Gaussian model which can go a long way in accommodating skew in one, or potentially both of these equations. The use of this method, however, will not be appropriate in all situations, and in these cases, more flexible alternatives such as Gaussian mixtures or Dirichlet processes could be implemented.

Aside from the appearance of the truncated latent variables in (4), the biggest departure of the model in (3)-(6) relative to the “textbook” treatment- response model is the nonparametric specification of the endogenous treatment variable in (3). Though several approaches could be employed to handle this issue, our approach to a nonparametric treatment of  $f(s)$  follows that described in Koop and Poirier (2004).

---

<sup>15</sup>For example, a test of symmetry can be conducted by testing  $\delta = 0$  which can easily be implemented via the Savage-Dickey density ratio.

<sup>16</sup>It is also worth noting that we suppose independence between  $h^*$  and  $\epsilon$ , which may or may not be appropriate. The adequacy of this assumption can also be assessed with diagnostic checks via posterior predictive simulation (see section 4).

First, we sort the data by values of  $s$ , so that  $s_1$  refers to the lowest value in the sample,  $s_j < s_{j+1}$ ,  $j = 1, 2, \dots, k_\gamma - 1$ , and  $s_{k_\gamma}$  denotes the largest value. (In practice, different observations may have the same values of  $s$ , and thus  $k_\gamma \leq n$ . In our data, for example, we have information on  $n = 1,782$  females, which produces  $k_\gamma = 672$  unique values of BMI). Stacking (3) and (4) over  $i$ , we can then obtain

$$\mathbf{y} = \mathbf{D}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (7)$$

$$\mathbf{s} = \mathbf{Z}\boldsymbol{\theta} + \delta\mathbf{h}^* + \mathbf{u}, \quad (8)$$

$$[\boldsymbol{\epsilon}' \ \mathbf{u}']' | \mathbf{x}, \mathbf{z}, \mathbf{h}^* \sim N(\mathbf{0}_{2n}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n),$$

where

$$\boldsymbol{\gamma} \equiv \begin{bmatrix} f(s_1) \\ f(s_2) \\ \vdots \\ f(s_{k_\gamma}) \end{bmatrix},$$

$\mathbf{D}$  is a  $n \times k_\gamma$  matrix with  $i$ th row  $\mathbf{d}_i$ , which is constructed to select off the appropriate element of  $\boldsymbol{\gamma}$  for that observation, and the quantities  $\mathbf{y}$ ,  $\mathbf{X}$ ,  $\boldsymbol{\epsilon}$ ,  $\mathbf{s}$ ,  $\mathbf{Z}$ ,  $\mathbf{h}^*$  and  $\mathbf{u}$  have been stacked over  $i$  in the obvious fashion.

As described in Koop and Poirier (2004), an informative prior on  $\boldsymbol{\gamma}$  can be used to surmount the problem of insufficient observations when  $k_\gamma = n$ , and also to introduce the possibility of smoothing the regression curve. To this end, we reparameterize the model in terms of the quantity  $\boldsymbol{\psi} \equiv \mathbf{H}\boldsymbol{\gamma}$ , where

$$\mathbf{H} \equiv \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \Delta_2^{-1} & -\Delta_2^{-1} - \Delta_3^{-1} & \Delta_3^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \Delta_3^{-1} & -\Delta_3^{-1} - \Delta_4^{-1} & \Delta_4^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \Delta_{k_\gamma-1}^{-1} & -\Delta_{k_\gamma-1}^{-1} - \Delta_{k_\gamma}^{-1} & \Delta_{k_\gamma}^{-1} \end{bmatrix},$$

and  $\Delta_j \equiv s_j - s_{j-1}$ . The elements of  $\boldsymbol{\psi} = \mathbf{H}\boldsymbol{\gamma}$  thus consist of a pair of ‘‘initial conditions’’  $\psi_1 = \gamma_1$  and  $\psi_2 = \gamma_2$  (i.e., the first two points on the regression curve) as well as differences of the form

$$\begin{aligned} \psi_j &= \frac{\gamma_j - \gamma_{j-1}}{s_j - s_{j-1}} - \frac{\gamma_{j-1} - \gamma_{j-2}}{s_{j-1} - s_{j-2}}, \quad j = 3, 4, \dots, k_\gamma \\ &\approx f'(s_{j-1}) - f'(s_{j-2}), \end{aligned}$$

so that the final  $k_\gamma - 2$  elements of  $\boldsymbol{\psi}$  are first-differences of the pointwise slopes.

To introduce the potential of smoothing the regression curves, we place an informative prior on the vector  $\boldsymbol{\psi}$ . Specifically, we will place a reasonably flat (but proper) prior on the initial

conditions  $\psi_1$  and  $\psi_2$  of the form

$$[\psi_1 \ \psi_2]' \sim N[\mathbf{0}_2, 10\mathbf{I}_2], \quad (9)$$

with  $\mathbf{I}_2$  denoting the  $2 \times 2$  identity matrix. As for the remaining elements of  $\psi$ , we specify a prior of the form

$$[\psi_3 \ \psi_4 \ \cdots \ \psi_{k_\gamma}]' | \eta \sim N[\mathbf{0}_{k_\gamma-2}, \eta \mathbf{I}_{k_\gamma-2}], \quad (10)$$

with  $\eta$  acting as a *smoothing parameter*, similar in spirit to the frequentist bandwidth parameter in local polynomial regression. In the limiting case where  $\eta \rightarrow 0$ , the first-differences of pointwise slopes are imposed to be equal, resulting in  $f$  being linear (with  $\psi_1$  and  $\psi_2$  defining the intercept and slope of the line). Moderate values of  $\eta$  will result in smoothed regression curves, while choosing  $\eta$  to be too large will produce regression functions that are excessively erratic (see Koop and Poirier (2004) for more details). Putting (9) and (10) together, we obtain a prior for  $\boldsymbol{\psi}$  of the form

$$\boldsymbol{\psi} | \eta \sim N[0, \mathbf{V}_\psi(\eta)],$$

where  $\mathbf{V}_\psi(\eta)$  is a  $k_\gamma \times k_\gamma$  block diagonal matrix with  $10\mathbf{I}_2$  on the upper block and  $\eta\mathbf{I}_{k_\gamma-2}$  on the lower block. The following priors complete the specification of the model:

$$\boldsymbol{\beta} \sim N(0, \mathbf{V}_\beta) \quad (11)$$

$$\boldsymbol{\theta} \sim N(0, \mathbf{V}_\theta) \quad (12)$$

$$\boldsymbol{\Sigma}^{-1} \sim W(\mathbf{R}, \rho) \quad (13)$$

$$\eta \sim IG(a, b) \quad (14)$$

$$\delta \sim N(0, V_\delta), \quad (15)$$

with  $W(\cdot, \cdot)$  denoting the Wishart distribution [see, e.g., Koop, Poirier and Tobias (2007, pg. 339)].

## 2.2 The Posterior Simulator

The complete model is given by the likelihood implied from (7) and (8) together with the distributional assumptions on  $h_i^*$  and  $\lambda_i$  in (5) and (6) and the priors in (9)-(15). Since our priors are conditionally conjugate, posterior simulation is straightforward via the Gibbs sampler,<sup>17</sup> and proceeds in five steps. Before describing these in detail, let us first define

$$\bar{\mathbf{X}} = \begin{bmatrix} \mathbf{D}\mathbf{H}^{-1} & \mathbf{X} & \mathbf{0}_{n \times k_\theta} & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{n \times k_\gamma} & \mathbf{0}_{n \times k_\beta} & \mathbf{Z} & \mathbf{h}^* \end{bmatrix}, \boldsymbol{\pi} = \begin{bmatrix} \boldsymbol{\psi} \\ \boldsymbol{\beta} \\ \boldsymbol{\theta} \\ \delta \end{bmatrix},$$

---

<sup>17</sup>Note that the triangularity of our model in (3) and (4) implies a unit Jacobian, regardless of the form of  $f(s)$ .

and note that the priors given in (9)-(12) and (15) imply that  $\boldsymbol{\pi} \sim N(0, \mathbf{V}_\pi)$ , where  $\mathbf{V}_\pi$  is the  $(k_\gamma + k_\beta + k_\theta + 1) \times (k_\gamma + k_\beta + k_\theta + 1)$  block diagonal matrix with  $\mathbf{V}_\psi(\eta)$ ,  $\mathbf{V}_\beta$ ,  $\mathbf{V}_\theta$  and  $V_\delta$  stacked along the main diagonal. The five steps of our posterior simulator are enumerated below:

**Step 1:**  $\boldsymbol{\pi} | \boldsymbol{\Sigma}^{-1}, \eta, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s}$

$$\boldsymbol{\pi} | \boldsymbol{\Sigma}^{-1}, \eta, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim N(\mathbf{D}_\pi \mathbf{d}_\pi, \mathbf{D}_\pi), \quad (16)$$

where

$$\mathbf{D}_\pi = \left( \overline{\mathbf{X}}' (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \overline{\mathbf{X}} + \mathbf{V}_\pi^{-1} \right)^{-1}, \quad \mathbf{d}_\pi = \overline{\mathbf{X}}' (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \overline{\mathbf{y}}, \quad \text{and} \quad \overline{\mathbf{y}} \equiv [\mathbf{y}' \mathbf{s}']'.$$

**Step 2:**  $\boldsymbol{\Sigma}^{-1} | \boldsymbol{\pi}, \eta, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s}$

$$\boldsymbol{\Sigma}^{-1} | \boldsymbol{\pi}, \eta, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim W \left[ \left( \mathbf{R}^{-1} + \sum_{i=1}^n \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} [\epsilon_i \ u_i] \right)^{-1}, n + \rho \right]. \quad (17)$$

Note that  $\epsilon_i$  are  $u_i$  are “known” given the data and parameters  $\boldsymbol{\pi}$ .

**Step 3:**  $\eta | \boldsymbol{\pi}, \boldsymbol{\Sigma}^{-1}, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s}$

$$\eta | \boldsymbol{\pi}, \boldsymbol{\Sigma}^{-1}, \mathbf{h}^*, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \sim IG \left( \frac{k_\gamma - 2}{2} + a, \left[ b^{-1} + \frac{1}{2} \sum_{j=3}^{k_\gamma} \psi_j^2 \right]^{-1} \right). \quad (18)$$

**Step 4:**  $\mathbf{h}^* | \boldsymbol{\pi}, \boldsymbol{\Sigma}^{-1}, \boldsymbol{\lambda}, \eta, \mathbf{y}, \mathbf{s}$

The assumptions of our model imply that each  $h_i^*$  can be sampled independently from their respective posterior conditionals. Completing the square on terms involving  $h_i^*$  in the posterior, we obtain

$$h_i^* | \boldsymbol{\Sigma}^{-1}, \boldsymbol{\pi}, \eta, \boldsymbol{\lambda}, \mathbf{y}, \mathbf{s} \stackrel{ind}{\sim} TN_{[-c_\nu, \infty)}(\mu_{h_i^*}, \sigma_{h_i^*}^2), \quad i = 1, 2, \dots, n, \quad (19)$$

where

$$\sigma_{h_i^*}^2 = \left( \lambda_i^{-1} + \frac{\delta^2}{\sigma_u^2} + \frac{\sigma_{\epsilon u}^2 \delta^2}{\sigma_u^4 \sigma_v^2} \right)^{-1} \quad \text{and} \quad \mu_{h_i^*} = \sigma_{h_i^*}^2 \left( \frac{\delta \tilde{s}_i}{\sigma_u^2} - \frac{c_\nu}{\lambda_i} - \frac{\delta \sigma_{\epsilon u} \tilde{y}_i}{\sigma_u^2 \sigma_v^2} \right),$$

with

$$\tilde{s}_i \equiv s_i - \mathbf{z}_i \boldsymbol{\theta}, \quad \tilde{y}_i \equiv y_i - (\mathbf{D}\mathbf{H}^{-1})_i \boldsymbol{\psi} - \mathbf{x}_i \boldsymbol{\beta} - \frac{\sigma_{\epsilon u}}{\sigma_u^2} \tilde{s}_i \quad \text{and} \quad \sigma_v^2 \equiv \sigma_\epsilon^2 (1 - \rho_{\epsilon u}^2).$$

### Step 5: $\lambda|\Sigma^{-1}, \boldsymbol{\pi}, \eta, \mathbf{h}^*, \mathbf{y}, \mathbf{s}$

Equations (5) and (6) imply that each mixing variable  $\lambda_i$  can be sampled independently from its posterior conditional, which is of the form

$$\lambda_i|\Sigma^{-1}, \boldsymbol{\pi}, \eta, \mathbf{h}^*, \mathbf{y}, \mathbf{s} \stackrel{ind}{\sim} IG\left(\frac{\nu+1}{2}, \left[\frac{\nu}{2} + \frac{1}{2}(h_i^* + c_\nu)^2\right]^{-1}\right), \quad i = 1, 2, \dots, n. \quad (20)$$

The Gibbs sampler is implemented by drawing from (16)-(20), updating parameters in all conditioning sets to equal their most recent values drawn from the algorithm.

## 3 The Data

In our empirical analysis we use the 1970 British Cohort Study, a longitudinal survey which tracks the cohort of all people born in Great Britain between April 5 and April 11, 1970. In this data set, cohort members and/or their parents are interviewed at the birth of the cohort and when the cohort is five, ten, sixteen, twenty-six, and twenty-nine/thirty years old<sup>18</sup>. During the first four interview waves, the parents were surveyed. Of particular interest to this study, the parents reported their height and weight during the interview wave when the cohort was ten years old. This allows for the creation of parent BMI variables. Finally, we also exclude observations not pertaining to individuals living with his or her biological parents at 10 years of age.

We observe labor market outcomes for the cohort when they were twenty-nine/thirty years old. STATA code provided by the Centre for Longitudinal Studies is used to calculate hourly wages from the raw data. Individual-level controls that we are able to extract from the data include tenure on the current job (measured in months and denoted as JobTenure), labor market experience (Experience), an indicator denoting the completion of a lower level of secondary education (Highschool), an indicator denoting the completion of a higher level of secondary education (ALevel), and an indicator for the completion of a college degree program (Degree).<sup>19</sup>

We additionally include controls denoting if the individual is married (Married) or has a

---

<sup>18</sup>This wave of interviews was conducted between October 1999 and September 2000.

<sup>19</sup>The education variables indicate highest level of educational attainment. We consider an individual to have a lower amount of secondary education if he or she has passed at least one GCSE, O-level, or CSE exam. An individual has a higher level of secondary education if he or she has passed at least one A-level exam.

union job (Union). With the exception of parent BMI, all variables enumerated to this point are employed in both the log wage and BMI equations.

For family background characteristics, we obtain the cohort members' family incomes at ten years of age (FamilyIncome), indicators denoting whether the cohort members' mothers or fathers held a college degree (MomDegree,DadDegree), and indicators denoting if the cohort members' mothers or fathers worked in a managerial or professional position (MomManProf, DadManProf). Family income is included in both equations (3) and (4), as it was found to be empirically important, while the remaining demographic controls are included in the wage equation primarily to mitigate the "learned preferences" link between parent and child BMI, which may undermine the instruments' validity.<sup>20</sup> Finally, heights and weights of the parents are used to create parent BMIs (MomBMI,DadBMI), which we then employ as our exclusion restrictions to identify our model.

We abstract from issues related to missing observations, and also focus on those individuals in the sample who are engaged in full time employment. This last sample restriction is potentially limiting, since BMI may play an important role in the decision to work and not just the level of wages given that one is employed.<sup>21</sup> We choose, however, to focus on the employed to remain consistent with the model discussed in section 2, and also because a more general model that accounts for the decision to work raises new endogeneity issues. Specifically, an elaborated model of this type would require an additional source of exogenous variation - some characteristic that influences the decision to work, but not log wages - and we are not able to credibly determine such an exclusion restriction from our data. Following the precedent set by other studies in the literature, we choose to conduct separate analyses for men ( $n = 2,561$ ) and women ( $n = 1,782$ ). This yields  $k_\gamma = 817$  and  $k_\gamma = 672$  unique values of BMI for our samples of men and women, respectively.

---

<sup>20</sup>The variables MomDegree, DadDegree, MomManProf and DadManProf were also added to the BMI equation, and the analysis was repeated. We found little role for these variables in the production of BMI (given the other set of employed controls) and moreover, their inclusion produced no meaningful changes in the remaining parameters of the model.

<sup>21</sup>To investigate this issue we fit a probit model on the decision to work, which, like the outcome equation in (3), included a nonparametric component of BMI as well as the other controls employed in this analysis. Given the lack of an exclusion restriction in this exercise, the results should not be interpreted as causal, but instead, simply summarize the association between BMI and the probability of employment across the BMI support. For both men and women, we found that the probability of employment when plotted over the BMI support has a (slight) inverse *U*-shape, with both tails of the BMI distribution associated with relatively lower probabilities of employment. For women, the maximum probability of employment was around .75, occurring at a BMI value near 28. At the 5<sup>th</sup> and 95<sup>th</sup> percentiles of the BMI distribution, the probabilities of employment were .69 and .73, respectively. For men, the maximum probability of employment was approximately .88, occurring at a BMI value near 25. At the 5<sup>th</sup> and 95<sup>th</sup> percentiles of the BMI distribution, the probabilities of employment were both approximately .85. Thus, although curvature is present, the shapes of these relationships are reasonably flat throughout a large portion of the BMI support.



## 4 Results and Diagnostic Checks

Before diving into the results of our analysis, we first provide some evidence that our elaborated treatment-response model described in equations (3) - (6) provides an adequate fit to our data. This exercise seems particularly important in light of the fact that instrumental variables, the dominant estimation technique in these types of models, does not require the distributional assumptions made in our analysis.<sup>22</sup> In the following section we provide evidence that the assumptions made here are not introduced arbitrarily, with the singular purpose of simplifying the posterior calculations, but instead, the model in (3) - (6) appears capable of adequately modeling the observed log wage and BMI data in our application.

### 4.1 Diagnostic Checking

To begin our diagnostic checks, it is useful to start with what might be considered the “default” assumption regarding the error structure and to document some deficiencies associated with this specification. We will then illustrate that our preferred model, as described in (3) - (6), can overcome these deficiencies. We thus begin with the “textbook” Gaussian selection model, as described in (1) and (2), and for the sake of brevity, we illustrate its performance using the subsample of females.<sup>23</sup>

We fit the model in (1)-(2) using a restricted version of the Gibbs sampler described in section 2, noting that the mixing variables  $\boldsymbol{\lambda}$ , latent variables  $\mathbf{h}^*$  and nonparametric component do not appear in this posterior simulator. Instead, wages are imposed to be linearly related to BMI, and joint normality is assumed. The sampler for this case (and all other cases) is run for 10,500 iterations and the first 500 of these are discarded as the burn-in period.

There are numerous diagnostic checks for investigating the reasonableness of a model’s assumptions, including, for example, the use of QQ plots [e.g., Lancaster (2004, Chapter 2)], posterior predictive p-values [e.g., Gelman, Carlin, Stern and Rubin (2004, section 6.3)], and other comparisons of specific features of the model to their counterparts in the observed data

---

<sup>22</sup>Note, however, that our model is non-standard in the sense that we provide a nonparametric treatment of the endogenous treatment variable and thus linear IV can not be immediately applied. In addition, we are able to obtain posterior inference that is exact rather than relying on large-sample approximations to the sampling distribution of the IV estimator.

<sup>23</sup>The BMI distribution for males was decidedly less skewed, and the differences in the performances across models were not as stark as those given here. For this reason, we focus on the female subsample to highlight the performance of our method.

[e.g., Koop, Poirier and Tobias (2007, Chapter 11)]. We focus here on one such exercise, which we feel offers a reasonable global assessment of the adequacy of our model.

We first obtain histograms of the observed BMI and log wage data. The bin centers and observed frequencies associated with these histograms are noted and stored. Once a model is fit using the Gibbs sampler, we then obtain draws from the posterior predictive distribution. That is, we obtain a vector  $y_j^{rep}$  from  $p(y^{rep}|y)$  where

$$p(y^{rep}|y) = \int_{\Theta} p(y^{rep}|\theta, y)p(\theta|y)d\theta, \quad (21)$$

and  $y_j^{rep} \sim y^{rep}|\theta = \theta_j, y$ , with  $\theta_j$  representing the  $j^{th}$  post-convergence draw from our posterior simulator. The density  $p(y^{rep}|\theta, y)$  is simply the likelihood function assumed by the given model (which does not depend on  $y$  given  $\theta$ ). In these simulations, we obtain a vector  $y_j^{rep}$  from the conditional density  $p(y^{rep}|\theta = \theta_j, y)$  by choosing exactly the same  $x$  and  $z$  values as those that our found in our sample of data. This motivates our use of the notation “rep” to denote *replications* of the observed data from the posterior predictive, i.e., they are “a re-run of history on the assumption that the model is what generates histories” [Lancaster (2004, pp. 90-91)].

The act of generating a series of  $y^{rep}$  variates in this way can reveal how well the model fares in reproducing the actual BMI and log wage distributions in the sample. Each  $y^{rep}$  is an  $n \times 1$  vector, whose distribution should mimic the actual distribution of  $y$  and  $s$  if our model’s assumptions are reasonable. To compare the predictions of our model to those found in the actual data, we obtain a histogram of the log wage and BMI values that are replicated from the posterior predictive, using the same bin centers and bin widths that were used to obtain the histograms associated with the observed log wage and BMI data. We obtain such a histogram for each post-convergence draw  $\theta_j$ . The posterior predictive frequencies within each bin are then averaged across iterations to produce a final histogram. This final histogram is then graphed alongside the histogram of the actual data, using the same scales for the x and y axes.

In Figure 1, we compare the posterior predictive and observed BMI histograms under the textbook Gaussian treatment-response model. Not surprisingly, the Gaussian histogram appears symmetric and is clearly not capable of reproducing the skew in the observed BMI distribution. The leftmost bin of the Gaussian histogram is also quite large, revealing that as the Gaussian model matches the (conditional) error variance in the data, many replications are produced in the far left-tail of the BMI distribution. These replications are then lumped into the first (smallest) bin, resulting in the spike appearing in the graph. These

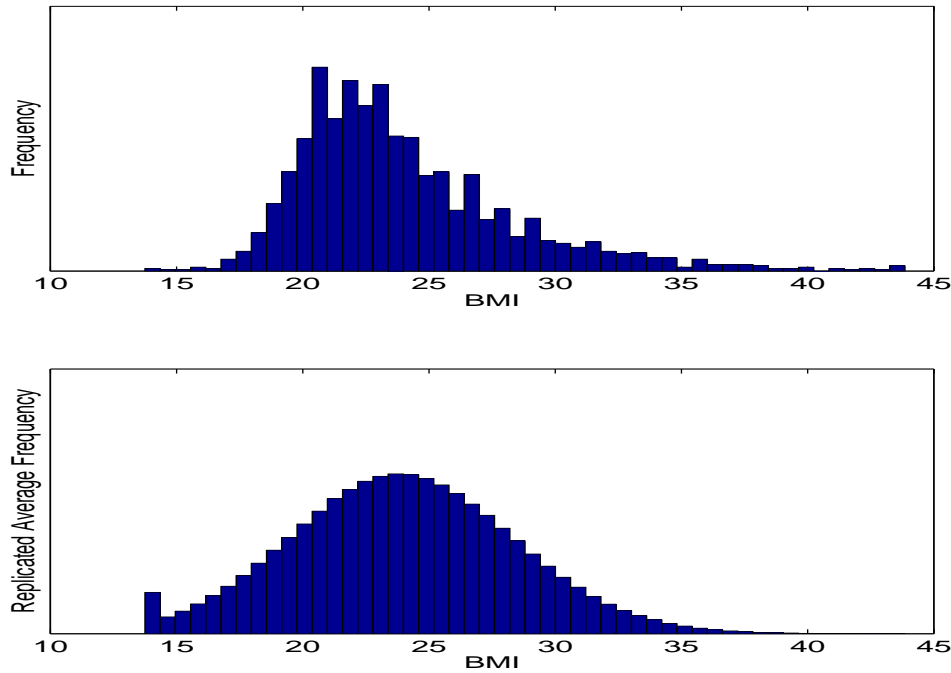


Figure 1: Actual (Top) and Replicated (Bottom) BMI Histograms obtained from Gaussian Model.

results suggest that the default treatment-response assumptions are not appropriate for this application.<sup>24</sup>

A remedy to the problem of modeling a skewed variable which is ubiquitous in empirical practice is to model its logarithm. In our application the log transformation does not eliminate the skew associated with BMI, as the natural log of BMI retains a skewness coefficient of approximately .74. To investigate the performance of the lognormal model more formally, we repeat the analysis described above, this time defining  $s$  to be log BMI, and retaining the assumption of joint normality. The posterior predictive and observed histograms associated with log BMI from this exercise are presented in Figure 2. As seen in this figure, the distribution of observed log BMI continues to have a right-skew, and, not surprisingly, the replicated log BMI distribution is approximately symmetrically distributed. Again, this suggests that the lognormal model, though an improvement over the textbook Gaussian model,<sup>25</sup> is not ideally suited to this data.

In Figure 3 we present results associated with the skew-normal model of (3)-(6), but to keep comparisons with earlier exercises consistent, we restrict the nonparametric specification

<sup>24</sup>This conclusion was not guaranteed, since other variables, such as parental BMI, are also skewed, which could induce a skew in the predictive histogram even under normality. Our results show that the predictive essentially remains symmetric despite this issue.

<sup>25</sup>Note that, in levels, the predictive BMI distribution from the lognormal model will have a right-skew, a feature that the textbook Gaussian model does not reproduce.

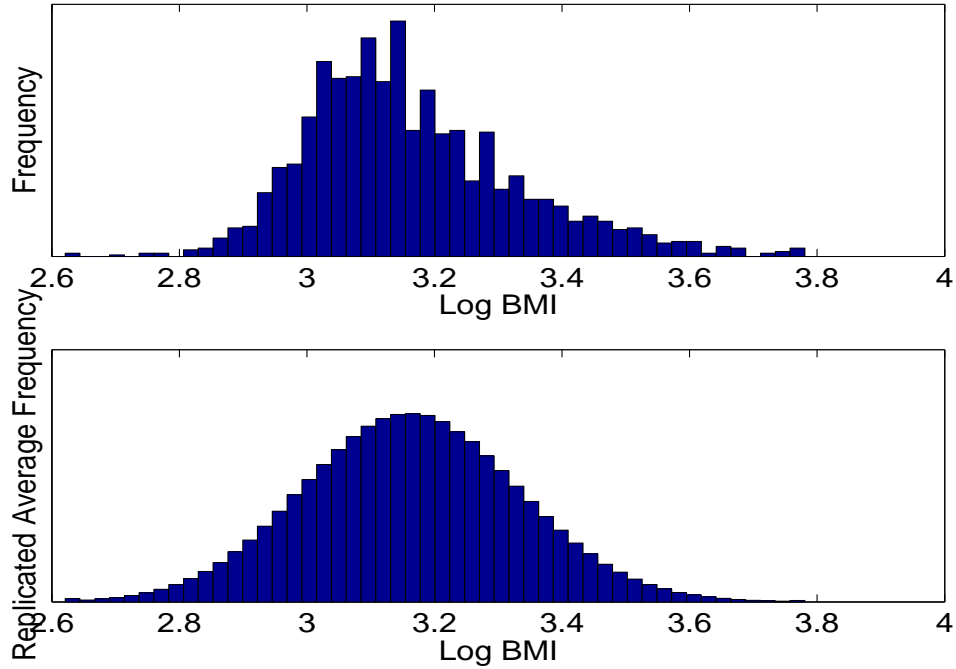


Figure 2: Actual (Top) and Replicated (Bottom) Log BMI Histograms. Gaussian Model with  $s$  defined as Log BMI.

in (3) to a linear one. As is evident from Figure 3, the predictions (replications) from our model with skew-normality fare well, and clearly perform better than the lognormal model in producing a histogram that reproduces the overall shape of the observed sample distribution of log BMI. Though the results of Figures 1-3 offer a reasonably complete picture of this comparison, we can also focus on specific features of the BMI distribution to illustrate the relative and absolute performance of the skew-normal model. These results are presented in Table 1.

In Table 1, we record the skewness, maximum, minimum, median, 15<sup>th</sup> and 85<sup>th</sup> percentiles of the actual BMI data as well as the skew of log BMI and  $\text{Corr}(y, s)$ . Posterior replications of each of these statistics are then obtained from posterior predictive distributions associated with the skew-normal, lognormal and Gaussian models. As Table 1 illustrates, the skew-normal model performs the best of those considered in all but one category (where the best in each category is marked with a “\*”), and the lognormal model does not perform well in statistics related to the skewness of the BMI variable and its natural logarithm. Finally, the skew-normal model also fares well in matching  $\text{Corr}(y, s)$  in the data, suggesting that the assumed independence between  $h^*$  and  $\epsilon$  is not obviously problematic for this application.

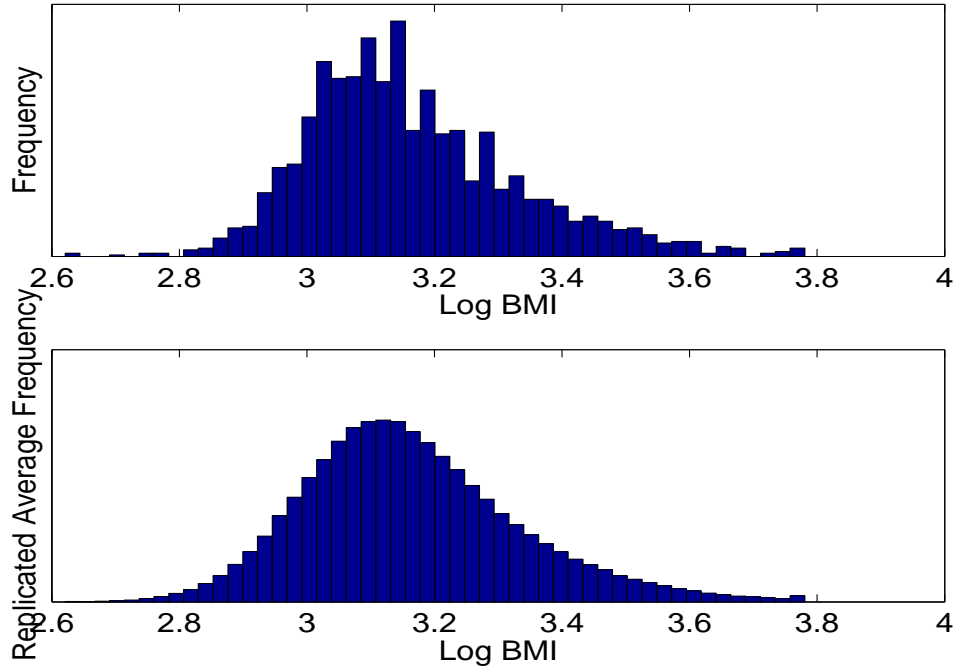


Figure 3: Actual (Top) and Replicated (Bottom) BMI Histograms obtained from Skewed Treatment Response Model

## 4.2 Empirical Results

We are interested in applying the algorithm of section 2 to the 1970 British Cohort Study data to address the following questions: (1) Does BMI play a role in the production of wages for men or women? (2) If so, is there evidence of a differential impact of BMI across gender? (3) Is there evidence of unobserved confounding, or that BMI needs to be treated as an endogenous variable? (4) Are there any nonlinearities in the relationships between BMI and log wages? (5) If so, do these nonlinearities align with clinical classifications of the body mass index? That is, are the wage penalties different for individuals in the “normal” BMI range (18.5-25), in the “overweight” range (25-30) and in the “obese” range (30+)? <sup>26</sup>

For both the male and female samples, we fit the model using the Gibbs sampler, as described

<sup>26</sup>Sargent and Blanchflower (1996) study workers in Great Britain and find evidence that weight-for-height has an effect on wages of females, but do not find evidence that weight-for-height has an effect on the wages of males. Specifically, they find that females in the top 10% of the BMI distribution (which they report to be a BMI of at least 26.1) earn roughly 5% less than females in the bottom 90%. Females in the top 1% (a BMI of at least 33.3) earn roughly 14% less than females in the bottom 99%. These results are similar to those of Harper (2000), another study of workers in Great Britain. The results of Sargent and Blanchflower (1996) are also in general agreement with what is reported by studies of the same relationships in samples of U.S. workers. Cawley (2004) studies the sample of U.S. workers in the 1979 National Longitudinal Survey of Youth and finds that a one point increase in BMI decreases wages by about 1% for white females, and does not affect wages for white males.

in section 2, setting  $a = 3$ ,  $b = 1.0 \times 10^5$ ,  $\rho = 5$ ,  $\mathbf{R} = \mathbf{I}_2$ ,  $\mathbf{V}_\beta = \mathbf{I}_{k_x}$ ,  $\mathbf{V}_\theta = \mathbf{I}_{k_z}$  and  $V_\delta = 10$ . Without question, the most influential of these choices on our posterior results is the choice of hyperparameters  $a$  and  $b$ , which govern the smoothness of our regression function. This choice of  $a$  and  $b$  sets the prior mean and prior standard deviation of the smoothing parameter  $\eta$  equal to  $5.0 \times 10^{-6}$ . These hyperparameters were settled upon after some experimentation and model checking, and were found to produce posterior results that were smooth (in accord with our prior beliefs), yet did not constrain the regression function to be necessarily linear.

To expedite convergence of the sampler, we started the parameter chain at reasonable values based on separate single-equation analyses of (3) and (4). Specifically, we first employed a posterior simulator to fit equation (3) only, and from the Gibbs output, we obtained posterior means of  $\beta$ ,  $f(s)$  and  $\sigma_\epsilon^2$  to use as starting values. Similarly, a posterior simulator for equation (4) was also employed, which produced posterior means of  $\theta$ ,  $\delta$ , and  $\sigma_u^2$  to use as starting values. Since these single-equation analyses are appropriate in the absence of unobserved confounding, we set the initial value of the covariance parameter  $\sigma_{\epsilon u}$  to zero. With these starting values, standard diagnostic checks suggested convergence within a hundred iterations. In practice, we obtained 10,500 posterior simulations and discarded the first 500 as the burn-in.<sup>27</sup>

#### 4.2.1 Results for the Females Sample

Coefficient posterior means, standard deviations, and probabilities of being positive for the female sample are presented in Table 2 of the appendix. For the log wage equation [i.e., equation (3)], the results presented in the table are generally consistent with our prior expectations. Specifically, we see moderate evidence of a quadratic relationship in both tenure on the current job and labor market experience. Family income also plays a significant role in the wage equation, as does the education level of the worker. The posterior distributions associated with the parental education and occupation parameters, however, generally placed a reasonable mass near zero, though having a father employed in a management or professional position does seem to be strongly associated with higher wages.

In terms of the reduced form BMI equation [i.e., equation (4)], we see clear evidence that parental BMI plays a strong role in the production of BMI of the child. Specifically, a unit

---

<sup>27</sup>We also performed generated data experiments in attempt to find errors in the posterior simulator and to investigate the performance of the algorithm. These experiments suggested that the program performed adequately in recovering the parameters used in the data generating process.

increase in the BMI of the mother or the father tends to increase the BMI of the child by approximately .3 points. The posterior standard deviations associated with these coefficients are also quite small relative to their means, and our posterior simulator produced no draws of these regression parameters whose values were negative.

In addition, as an informal check of the instruments validity, we performed a version of the standard overidentification-type test. That is, we first assumed that MomBMI was a valid instrument and then included DadBMI in both the BMI and log wage equations. We then calculated the relevant Bayes factor, which under equal prior odds gives the posterior odds in favor of imposing the restriction that the coefficient associated with DadBMI in the wage equation is zero. Performing these calculations via the Savage-Dickey density ratio, we obtained a Bayes factor of 290.27, indicating strong support that DadBMI can be excluded from the log wage equation. Repeating this exercise, but this time assuming that DadBMI was a valid instrument, we obtained a Bayes factor of 115.95 associated with the hypothesis that the MomBMI coefficient was equal to zero. These results provide intuitive support (though certainly not formal proof) for the validity of the instruments.

For the remaining parameters of our model, we note that the posterior mean and standard deviation of the skewness parameter  $\delta$  were 3.91 and .141, respectively, providing strong evidence of a pronounced right-skew in the BMI distribution. The correlation parameter  $\rho_{eu}$ , which quantifies the degree of unobserved confounding, had a posterior mean of .082 and a posterior standard deviation of nearly equal magnitude. A formal Bayes factor in favor of the restriction  $\rho_{eu} = 0$ , calculated via the Savage-Dickey density ratio,<sup>28</sup> yielded a value of 5.26. This suggests that, for our sample of females, concerns regarding the endogeneity of BMI seem rather minimal, as the restricted model with no confounding is favored by a factor of approximately 5.26 to 1 under the employed priors.

Finally, in Figure 4, we present a point estimate (i.e., posterior mean) of the function  $f(s)$  in equation (3). For the sake of comparison, we also include a plot of the same function under the restricted model in (1) and (2) where the relationship between BMI and log wages is assumed to be linear. In terms of the point estimates, the linear model is found to overstate the magnitude of the BMI penalty (i.e., the slope of the regression curve) over both very small and reasonably large values of BMI, and understate this magnitude for moderate BMI values. These results are somewhat consistent with our prior intuition, where we speculated that marginal increases in BMI would not be penalized severely, if at all, for an individual with low values of BMI, but would become increasingly penalized as the individual moves

---

<sup>28</sup>See, for example, Koop, Poirier and Tobias (2007, page 69) for more on the implementation of this test.

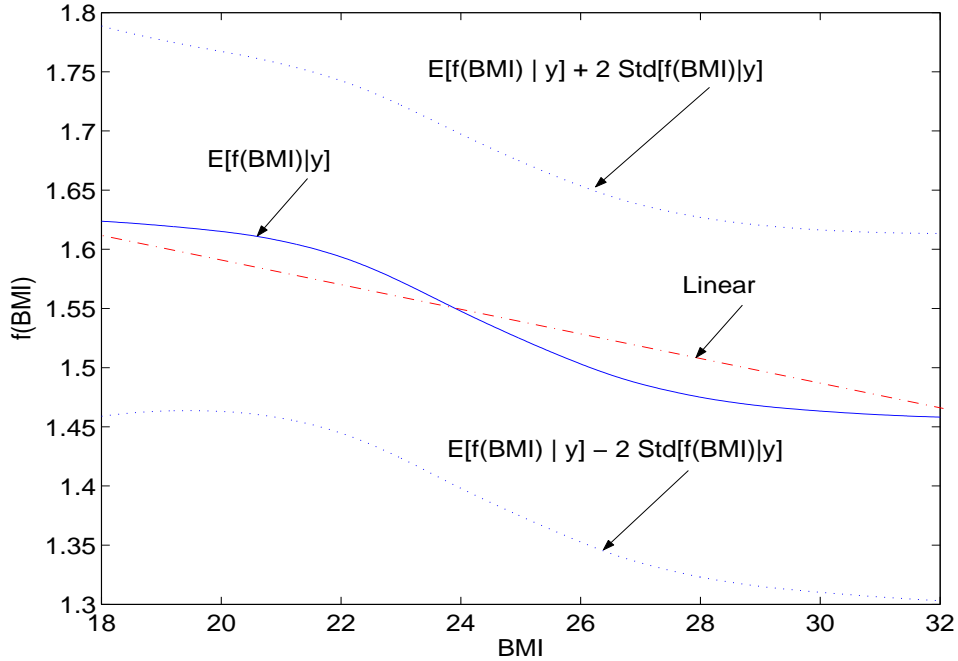


Figure 4: Posterior Means of Functions Relating BMI to Log Wages from Nonparametric and Linear Models. Females Sample.

toward the right of the BMI distribution. We find empirical support for this hypothesis until we reach a BMI value of approximately 28 (roughly the 85th percentile of the BMI distribution), where, somewhat surprisingly, the BMI penalty tends to level off and continues to do so even for obese females with BMI values exceeding 30. These results are also robust to moderate changes in the prior (i.e., the general shape of Figure 4 is unchanged when choosing hyperparameters similar to those currently employed), though, of course, strong (or weak) priors on  $\eta$  can essentially force linearity (or yield an estimated function that is unreasonably erratic).

It is also relevant to question whether the data support the nonlinear specification relative to the more parsimonious linear model. We do not report a formal marginal likelihood calculation in this regard, as Bayes factors are highly sensitive to the prior specification, particularly in this situation where the prior plays an important role in governing the shape (smoothness) of  $f(s)$ .<sup>29</sup> However, we note that the posterior mean of the smoothing parameter  $\eta$  is approximately one-half of its prior mean, and the posterior standard deviation is approximately one-third the magnitude of the prior standard deviation. This suggests

<sup>29</sup>For example, if the prior for  $\eta$  essentially imposes linearity, yet the data pull our prior toward slightly larger values of  $\eta$ , we will tend to reject the hypothesis of linearity, even if our point estimates are indistinguishable from a linear model. Conversely, if the prior for  $\eta$  concentrates over “large” values, and the data pull our prior slightly toward zero, marginal likelihood calculations will tend to support linearity, even if the resulting posterior estimates seem quite nonlinear.



that the data have moved our prior toward smaller values of the smoothing parameter, thus leading us to favor linearity, even though our point estimates under this prior suggest some nonlinearities in the regression function.

A seemingly useful exercise here involves changing the prior hyperparameters for the smoothing parameter, and keeping track of when the data move our prior toward larger values of  $\eta$  (and thus away from linearity). If we can find a value of  $b$  (keeping  $a = 3$ ) such that the posterior becomes shifted to the right relative to the prior, yet the results in Figure 2 still suggest the same general pattern of nonlinearities, this suggests data-provided evidence in favor of the existence of these nonlinearities. Conversely, if such a movement to the right is difficult to document under any reasonable choice of  $b$ , or only happens under a  $b$  where the prior and posterior means of  $f$  appear linear, then little support seems to be offered against the linear model.<sup>30</sup> With this exercise in mind, we find that when setting  $b = 1.0 \times 10^6$  (which implies a prior mean and standard deviation for  $\eta$  equal to  $5.0 \times 10^{-7}$ ), the posterior mean of  $\eta$  becomes larger than the prior mean and the posterior distribution becomes more concentrated than the prior. This suggests that our data do not support such a large degree of smoothing. This trend of right-shifting persists for values of  $b$  larger than  $b = 1.0 \times 10^6$ , though, again, if  $b$  is sufficiently large, both the prior and posterior mean of  $f(s)$  will appear linear. Interestingly, under this prior with  $b = 1.0 \times 10^6$ , the posterior mean of  $f$  retains the same general shape as presented in Figure 4, suggesting that such departures from linearity appear to be supported by the data. We will also revisit this issue of nonlinearities and their empirical importance in a formal comparison of average derivatives in the following section.

#### 4.2.2 Results for the Males Sample

Coefficient posterior means, standard deviations and probabilities of being positive for the males sample are provided in Table 3. Figure 5, like Figure 4, plots the posterior mean of the function  $f(s)$  for our sample of males, as well as the mean obtained from the restricted linear model in (1) and (2).

The results of Table 3 are, for the most part, very similar to those obtained for the females analysis. In particular, we see strong evidence of a quadratic profile in job tenure, modest evidence of a quadratic profile in labor market experience, family income and educational

---

<sup>30</sup>Note that this technique, unlike traditional Bayes factors calculations, explicitly makes use of different priors rather than a single prior in the model comparison exercises. The spirit of this approach, as we search for a prior that is “inline” with the posterior, is somewhat similar to an empirical Bayes procedure, where  $\eta$  would be chosen to maximize the marginal likelihood.

attainment are clearly important explanatory variables in the log wage outcome equation, the parental BMI instruments play strong roles in the production of child BMI, and the posterior mean of  $\delta$  again provides evidence of a right skew in the BMI distribution. This skew, however, is not as pronounced as the skew found in the females sample. We also obtain similar results when performing the overidentification-type tests. When assuming that DadBMI is a valid instrument and including MomBMI in the log wage equation, we calculated a Bayes factor equal to 36.41 in favor of the model that excludes MomBMI. Similarly, when assuming that MomBMI is a valid instrument and including DadBMI in the log wage equation, we calculated a Bayes factor equal to 60.56 in favor of the model that excludes DadBMI. These results, like those of the female sample, provide suggestive evidence supporting the validity of our identification strategy.

One important difference relative to our females sample, however, is that we find evidence of an endogeneity problem for males. Specifically, the posterior mean of the correlation parameter  $\rho_{eu}$  is .26, and all post-convergence simulations associated with this parameter were positive. These posterior statistics clearly illustrate the importance of controlling for unobserved confounding for our males sample. In our view, the positive correlation can also be provided with a reasonable economic interpretation. To this end, consider the case of a male worker who is (unobservably) dedicated to his job, and thus earns a higher wage than would otherwise be predicted based solely on observables. This “dedication” would, at least in part, likely require the worker to spend less time engaged in other activities, such as exercise, or home preparation of meals, which we might associate with lower values of BMI. Such a story is consistent with the finding of a positive correlation among the unobservables, which is also consistent with our prior expectations. The fact that such a correlation was not found among our females sample, however, leaves something of a quandary, although the point estimate for females was also positive (.082), with a reasonably high posterior probability of being positive (.897), which is broadly consistent with our finding for the males sample. We have little explanation, however, for the remaining differential impact across genders, other than to suggest that similarly dedicated females may be less willing, on average, to substitute away from activities which affect health and lower BMI.

The graphs in Figure 5, like those of Figure 4, again suggest some evidence of nonlinearities in the relationship between BMI and log wages. Interestingly, for the males subsample, the function is reasonably flat (though clearly downward sloping) over BMI values in the “normal” range. However, unlike our results for the females subsample, we see comparably large wage penalties for males who are overweight or obese.

Table 4 offers a clearer picture of the nonlinearities found within the male and female sam-

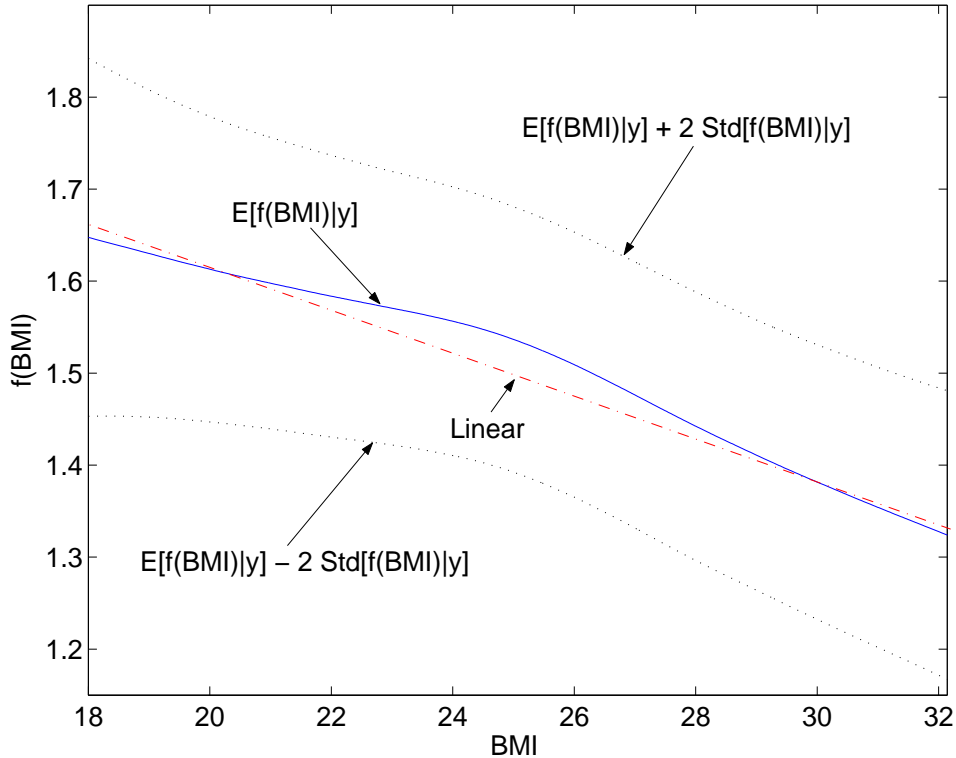


Figure 5: Posterior Means of Functions Relating BMI to Log Wages from Nonparametric and Linear Models. Males Sample.

ples and a more careful comparison of slopes across gender groups. We focus in particular on how BMI wage penalties change across common clinical divisions of the body mass index. To this end, we focus on obtaining distributions associated with the *average derivative* within a given region. The average derivative is simply the expected rate of change of the function  $f$ , where the averaging is performed with respect to the density of the data, which in our application, would be the density of BMI.<sup>31</sup> Focusing on the average derivative, unlike nonparametric estimation of the function of itself, can prove to be insightful since  $\sqrt{n}$  asymptotics are applicable regarding its estimation,<sup>32</sup> leading to more precise calculations regarding the possible existence of nonlinearities. Our idea is to make use of the average derivative to investigate the possible existence nonlinearities across various “clinical” classifications of BMI, and to see if these average rates of change within a given region vary across gender groups. Formally, within a given BMI region  $\mathcal{C}$ , the average derivative can be

<sup>31</sup>Formally, the average derivative is obtained as  $\int f'(s)g(s)ds$ , where  $g(s)$  denotes the density of BMI.

<sup>32</sup>See, for example, Banerjee (2007).

calculated as follows:

$$\begin{aligned}
\text{AvgDer}_{\mathcal{C}} &\equiv E[f'(s)|s \in \mathcal{C}] \\
&\approx \sum_{i=1}^n f'(s_i) \Pr(s = s_i | s \in \mathcal{C}) \\
&= \sum_{i=1}^n f'(s_i) \left[ \frac{\Pr(s = s_i)}{\Pr(s \in \mathcal{C})} I(s_i \in \mathcal{C}) \right] \\
&= \frac{1}{\Pr(s \in \mathcal{C})} \sum_{i:s_i \in \mathcal{C}} f'(s_i) \Pr(s = s_i) \\
&\approx \frac{\sum_{i:s_i \in \mathcal{C}} w_i f'(s_i)}{\sum_{i:s_i \in \mathcal{C}} w_i}.
\end{aligned}$$

The second line above notes that the actual distribution of BMI in our sample is discrete, with multiplicities occurring at “common” BMI values. The final line notes that the average derivative for the region  $\mathcal{C}$  can be calculated as a weighted average of pointwise derivatives within that region (which are produced from our posterior simulator), with weights denoted as  $w_i$ . These weights can be easily obtained from our sample as the fraction of observations with the given BMI value:  $w_i = [1/n] \sum_{j=1}^n I(s_j = s_i)$ . Since  $\text{AvgDer}_{\mathcal{C}}$  is a function of the parameters  $\psi$ , we can use our posterior simulations to approximate the posterior distribution of this average derivative. The results of these calculations are provided in Table 4.

First, we focus on within-gender comparisons. For females, the posterior means of the average derivatives are -.016, -.014, and -.002 for our three BMI classifications. These results indicate that the average percentage decreases to wages resulting from a one-point increase in BMI are 1.6, 1.4 and .2 for normal weight, overweight, and obese females, respectively. The final row in this top section of the table calculates the likelihood that the normal weight average derivative exceeds the average derivatives in the overweight and obese ranges. Although the probability that the average slope for normal weight women exceeds (i.e., is less negative than) the average slope for overweight women is rather modest (.314), we do see strong evidence that the slope is less negative for obese women than it is for normal weight women. (The posterior probability associated with this statement is  $1 - .031 = .969$ ). For men, the results are somewhat reversed, as the average slopes in the overweight and obese ranges are -.03 and -.027, respectively, which are far more negative than the average slope in the normal weight range (-.014). In fact, the probability that the average slope in the normal range exceeds (i.e., is less negative than) the average slopes in the overweight and obese ranges are .986 and .899, respectively.

The final row of the table helps quantify how different the average BMI wage penalties are across men and women over different regions of the BMI support. For the normal BMI region,

there is little evidence of a differential penalty across men and women, though, when defining the normal BMI region as  $[20, 25]$ , (as occasionally done in the literature), we do see modest evidence that men receive a smaller average penalty to marginal increases in BMI than women receive. (The posterior probability that this statement is true was approximately .7). However, overweight and obese men experience a significantly steeper average BMI penalty than comparably overweight and obese women. (The posterior probability that overweight women had a smaller average BMI penalty than overweight men was approximately  $1 - .024 = .976$  and the posterior probability that obese women had a smaller average BMI penalty than obese men was approximately  $1 - .006 = .994$ ). As a whole, the results of Table 4 attest to the importance of nonlinearities in the BMI-log wage relationships since, for a given gender group, we find strong evidence that average slopes differ across BMI regions. In addition, the results suggest that men and women are not penalized in the same way for increases to BMI, and perhaps surprisingly, men tend to receive the largest penalties for being overweight or obese. The pattern of these results are quite interesting, and illustrate the value of the methodology of section 2, which enables the researcher to flexibly explore the relationship between the endogenous BMI variable and the log wage outcome.

## 5 Conclusion

In this study we extended earlier analyses of the effect of BMI on wages by providing a nonparametric treatment of the function relating these variables. We derived and employed a Bayesian posterior simulator for fitting this model, which not only allowed for a nonparametric treatment of a potentially endogenous BMI variable within a treatment-response framework, but also flexibly allowed for possible skew in the conditional BMI distribution.

We found some evidence of nonlinearities in the relationships between BMI and log wages, and that the shapes of the estimated regression functions were different for men and women. We found that males receive relatively small penalties to increases in BMI provided their BMI falls in the “normal” range, while overweight or obese males receive comparably large wage penalties for further increases to BMI. Conversely, women were found to receive the largest wage penalty at relatively low BMI, and smaller penalties as BMI increased. Our finding in the males sample that BMI affects wages at all is somewhat novel in the literature. It is our hope that the results presented here not only add value to this specific field of inquiry, but that the general methodology described in section 2 will be useful to other researchers whose empirical models share a similar structure.

## References

- [1] Angrist, JD. 2004. Treatment effect heterogeneity in theory and practice” *Economic Journal* **114**: C52-C83.
- [2] Angrist, JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables” *Journal of the American Statistical Association* **91**: 444-455.
- [3] Averett S, Korenman S. 1996. The economic reality of the beauty myth. *Journal of Human Resources*, **31**:304-330.
- [4] Azzalini A, Dalla Valle A. 1996. The multivariate skew-normal distribution, *Biometrika* **83**: 715-726.
- [5] Banerjee, A. 2007. A method of estimating the average derivative. *Journal of Econometrics* **136**: 65-88.
- [6] Baum II CL, Ford WF. 2004. The wage effects of obesity: a longitudinal study. *Health Economics*. **13**: 885-899.
- [7] Behrman JR, and Rosenzweig MR. 2001. The returns to increasing body weight. Penn Institute for Economic Research Working Paper. No. 01-052.
- [8] Branco MD, Dey D.K. 2002. Regression model under skew elliptical error distribution. *Journal of Mathematical Sciences* **1**, 151-169.
- [9] Chen M-H, Dey D, Shao Q-M. 1999. A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association* **94**, 1172-1186.
- [10] Cawley J. 2004. The impact of obesity on wages. *The Journal of Human Resources*. **39**: 451-474.
- [11] Celeux G, Hurn M, Robert CP. 2000. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957-970.
- [12] Deb P, Munkin MK, Trivedi P. 2006. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics* **21**: 1081-1099.
- [13] Deurenberg P, Weststrate JA, Seidell JC. 1991. Body mass index as a measure of body fatness: Age- and sex- specific prediction formulas. *British Journal of Nutrition* **65**., 105-114.
- [14] Gelman A, Carlin JB, Stern HS, Rubin DB. 2007. *Bayesian Data Analysis*, Chapman & Hall / CRC: Boca Raton, FL.
- [15] Geweke J. 2007. Interpretation and inference in mixture models: simple MCMC works. *Computational Statistics & Data Analysis* **51**: 3529-2550.
- [16] Grilo CM, Pogue-Geile MF. 1991. The nature of environmental influences on weight and obesity: a behavior genetic analysis. *Psychological Bulletin* **110**:520-37.
- [17] Harper B. 2000. Beauty, stature, and the labour market: a British cohort study. *Oxford Bulletin of Economics and Statistics* **62**:771-800.
- [18] Heckman JJ, Vytlacil E. 1998. Instrumental variables methods for the correlated random coefficient model. *Journal of Human Resources* **33**: 974-987.
- [19] Heckman JJ, Vytlacil E. 2006. Structural equations, treatment effects and econometric policy evaluation. *Econometrica* **73**: 669-738.
- [20] Imbens GW, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* **62**: 467-476.
- [21] Koop G, Poirier DJ. 2004. Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics* **123**: 259-282.
- [22] Koop G, Poirier DJ, Tobias JL. 2007. *Bayesian Econometric Methods*. Cambridge University Press: Cambridge.
- [23] Lancaster T. 2004. *An Introduction to Modern Bayesian Econometrics*. Blackwell publishing.
- [24] Maes HH., Neale MC, Eaves LJ. 1997. Genetic and environmental factors in relative body weight and human adiposity. *Behavior Genetics* **27**, 325-351.

- [25] Manning A. 2004. Instrumental variables for binary treatments with heterogeneous treatment effects: A simple exposition. *Contributions to Economic Analysis & Policy*, **3**, Article 9.
- [26] Munkin M, Trivedi P. 2006. Bayesian analysis of the endogenous ordered probit model. working paper, Indiana University.
- [27] Pewsey A. 2000. Problems of inference for Azzalini's skew-normal distribution. *Journal of Applied Statistics* **27**: 859-870.
- [28] Prentice AM, Jebb SA. 2001. Beyond body mass index. *Obesity Reviews* **2**: 141-147.
- [29] Sargent JD, Blanchflower TD. 1994. Obesity and stature in adolescence and earnings in young adulthood. *Archives of Pediatrics and Adolescent Medicine* **148**: 681-687.
- [30] Stunkard AJ, Sorensen TI, Hanis C, Teasdale TW, Chakraborty R, Schull WJ, Schulsinger F. 1986. An adoption study of human obesity. *New England Journal of Medicine* **314**: 193-198.
- [31] Wooldridge JM. 2003. Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters* **79**: 185-191.

Table 1: Replication Summaries Across Alternate Models

Model		Statistic/Percentile							
Raw Data		Corr( $s, y$ )	Skew( $s$ )	Skew[log( $s$ )]	Min( $s$ )	15 <sup>th</sup>	50 <sup>th</sup>	85 <sup>th</sup>	Max( $s$ )
		-.155	1.33	.74	13.75	20.34	23.03	28.16	43.85
Skew-Normal	$E(\cdot y)$	-.119(*)	1.24(*)	.60(*)	13.47(*)	20.17(*)	23.36(*)	27.99	44.81(*)
	$Std(\cdot y)$	.029	.078	.075	.952	.112	.101	.135	1.79
Log-Normal	$E(\cdot y)$	-.015	.55	.029	13.30	19.85	23.65	28.22(*)	42.74
	$Std(\cdot y)$	.050	.079	.060	.770	.150	.143	.212	2.70
Normal	$E(\cdot y)$	.014	.078	-.57	9.19	19.39	23.93	28.59	40.51
	$Std(\cdot y)$	.047	.061	.099	1.43	.190	.157	.198	1.95

Table 2: Parameter Posterior Means, Standard Deviations and Probabilities of Being Positive. Females Subsample ( $n = 1,782$ )

Wage Equation			
Variable	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
JobTenure	.0210	.0074	.998
JobTenure <sup>2</sup>	-.0009	.0006	.044
Experience	.0247	.0148	.954
Experience <sup>2</sup>	-.001	.0008	.109
FamilyIncome	.0007	.0002	1.00
HighSchool	.0694	.0227	1.00
ALevel	.2764	.0323	1.00
Degree	.3414	.0307	1.00
Union	.0346	.0191	.967
Married	-.0139	.0174	.213
MomDegree	.0559	.0501	.868
MomManProf	-.0024	.0252	.462
DadDegree	-.0149	.0278	.297
DadManProf	.0556	.0224	.993
BMI Equation			
Variable	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
Constant	7.80	.656	1.00
FamilyIncome	.002	.001	.977
HighSchool	.304	.193	.945
ALevel	.391	.291	.905
Degree	-.379	.220	.045
Union	.158	.162	.839
Married	.278	.154	.968
MomBMI	.326	.021	1.00
DadBMI	.309	.023	1.00
Other Parameters			
Parameter	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
$\delta$	3.91	.141	1.00
$\eta$	$2.98 \times 10^{-6}$	$1.77 \times 10^{-6}$	1.00
$\sigma_\epsilon^2$	.125	.004	1.00
$\sigma_u^2$	4.19	.464	1.00
$\rho_{eu}$	.082	.063	.897



Table 3: Parameter Posterior Means, Standard Deviations and Probabilities of Being Positive. Males Subsample ( $n = 2,561$ )

Wage Equation			
Variable	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
JobTenure	.0234	.0074	.999
JobTenure <sup>2</sup>	-.0011	.0005	.011
Experience	.0325	.0139	.991
Experience <sup>2</sup>	-.0008	.0007	.142
FamilyIncome	.0008	.0002	1.00
HighSchool	.0674	.0199	1.00
ALevel	.3045	.0369	1.00
Degree	.377	.0294	1.00
Union	.0318	.0196	.947
Married	.1106	.0174	1.00
MomDegree	.0605	.0490	.891
MomManProf	-.0019	.0297	.480
DadDegree	.1388	.0288	.9998
DadManProf	-.0148	.0241	.266
BMI Equation			
Variable	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
Constant	11.08	.563	1.00
FamilyIncome	.003	.001	.997
HighSchool	.118	.134	.807
ALevel	.030	.263	.540
Degree	-.423	.189	.000
Union	.199	.143	.922
Married	.706	.127	1.00
MomBMI	.224	.017	1.00
DadBMI	.334	.020	1.00
Other Parameters			
Parameter	$E(\cdot y)$	$\sqrt{\text{Var}(\cdot y)}$	$\text{Pr}(\cdot > 0 y)$
$\delta$	2.49	.139	1.00
$\eta$	$3.45 \times 10^{-6}$	$1.85 \times 10^{-6}$	1.00
$\sigma_\epsilon^2$	.169	.006	1.00
$\sigma_u^2$	7.19	.526	1.00
$\rho_{eu}$	.261	.052	1.00

Table 4: Average Derivative Statistics Across BMI Regions

		BMI Region		
		Normal BMI $\in [18.5, 25)$	Overweight BMI $\in [25, 30)$	Obese BMI $\geq 30$
Women	$E(\cdot y)$	-.0164	-.0137	-.0020
	Std( $\cdot y$ )	.0062	.0045	.0051
	$\text{Pr}(\cdot > 0 y)$	.0057	.0000	.3320
	$\text{Pr}(\text{AvgDer. Normal} > \cdot y)$	—	.314	.031
Men	$E(\cdot y)$	-.0140	-.0306	-.0264
	Std( $\cdot y$ )	.0082	.0071	.0083
	$\text{Pr}(\cdot > 0 y)$	.0500	.0000	.0000
	$\text{Pr}(\text{AvgDer. Normal} > \cdot y)$	—	.986	.899
$\text{Pr}(\text{AvgDer. Men} > \text{AvgDer. Women} y)$		.581	.0235	.0057