

# Modelling Body Mass Index Distribution using Maximum Entropy Density

F. Chan <sup>a</sup>, M. Harris <sup>a</sup> and R. Singh <sup>a</sup>

<sup>a</sup>*School of Economics and Finance, Curtin University, GPO BOX U1987, Perth, Western Australia, 6845*  
Email: [ranjodh.singh@curtin.edu.au](mailto:ranjodh.singh@curtin.edu.au)

**Abstract:** The objective of this paper is to model the distribution of Body Mass Index (BMI) for a given set of covariates. BMI is one of the leading indicators of health and has been studied by health professionals for many years. As such, there have been various approaches to model the distribution of BMI. Furthermore, there are numerous studies which investigate the association between an individual's physical and socio-economic attributes (covariates) to their BMI levels. This paper proposes the use of Maximum Entropy Density (MED) to model the distribution of BMI using information from covariates. The paper shows how covariates can be incorporated into the MED framework. This framework is then applied to an Australian data set. The results show how different covariates affect different moments of the estimated BMI distribution.

**Keywords:** *Maximum entropy density, body mass index*

## 1 INTRODUCTION

The objective of this paper is to model the distribution of Body Mass Index (BMI) using a set of covariates. BMI is one of the leading indicators of an individual's health. Specifically, it estimates the amount of body fat of an individual. This is done by dividing an individual's mass (kg) by the square of their height (m). Differences in BMI across adults are generally due to the amount of body fat. As such, this metric is used as a comparison tool across individuals. As per the Australian Institute of Health and Welfare (AIHW), a BMI value under 18 is classified as underweight, values from 18 to 25 (inclusive) are considered normal, values from 25 to 30 are considered overweight and values over 30 are classified obese. According to the AIHW in 2012, 63% of Australian adults and 25% of children were overweight or obese. The AIHW claimed that being overweight and obese is the second highest contributor to the burden of disease<sup>1</sup> and obesity rates have doubled since the 1980's in Australia. This is consistent with the other developed nations around the globe according to the World Health Organization (WHO). Aside from the health implications of being obese, there are also economic consequences such as loss of productivity arising due to employee absenteeism.

Given these reasons, it is not surprising that many government agencies and academics are investigating obesity rates using measures such as BMI. As such, there exists a vast amount of academic literature on BMI. However, for the purpose of this study, the focus is on a particular subset of this literature. This is done in order to highlight the gap in the literature and thus appropriately place the paper's contribution. This paper classifies BMI research into two broad categories. The first category consists of studies which attempt to fit a density function to an empirical distribution of BMI. An example of this study is Flegal and Troiano (2000) which uses graphical methods (mean difference plots) to describe changes in the distribution of BMI for both adults and children in the US. Another paper by Penman and Johnson (2006) proposes the log-normal distribution to estimate BMI for a given population. A comprehensive paper by Lin *et al.* (2007) estimates the BMI distribution using a finite mixture model of Normal, skew Normal, Student *t* and skew Student *t* distributions. This study found that a finite mixture of skew student *t* distribution provided a better fit compared to normal mixtures. The paper by Contoyannis and Wildman (2007) estimates the BMI distribution of two different countries using non-parametric techniques. Once these distributions are constructed, a range of measures are used to examine the differences in the modelled distributions. Lastly, Houle (2010) uses similar methods as Contoyannis and Wildman (2007) to study differences in BMI distributions for Gender and Education.

The second category consists of studies which attempt to model the mean or median of BMI using a set of covariates. An example of this kind of study is Beyerlein *et al.* (2008) where three different regression approaches- Generalized Linear Models (GLMs), Quantile Regression and Generalized Additive Models for Location, Scale and Shape (GAMLSS) were employed to model childhood BMI. The major finding of their paper was that GAMLSS and Quantile regression provided a much better fit compared to GLMs for a given set of risk factors. Another more recent paper by Bottai *et al.* (2014) examined associations among age, physical activity and birth cohort on BMI percentiles in men using Quantile regression. The paper concluded that Quantile regression allow one to examine how various covariates affected BMI at different percentiles of the estimated BMI distribution.

Based on this classification, this paper attempts to combine the objectives from both categories. In other words, this paper attempts to model the distribution of BMI using covariates (risk factors, attributes). This framework will allow different covariates to influence different aspects (moments) of the estimated BMI distribution. This is a point that studies based on Quantile regression do claim. However, the impact of the covariate is measured at specific percentiles such as the 90% or 95% percentiles.

To the authors knowledge, there is another study which has combined the objectives of both categories. The paper by Brown *et al.* (2014) proposes a statistical model (Normal distribution) to model the BMI distribution of an unobserved (latent) class of individuals within a population. It is expected that a finite mixture of these models will provide a good fit for the overall BMI distribution. The weight of the each model is determined using the covariates (individual attributes) in the class and these covariates are same for each class. As a result of this differing values of the same covariate determine the weights for distribution of each class. Hence, the paper has been able to model the distribution of BMI using information from the covariates.

There are however, a number of factors that one needs to consider when implementing such an approach. One such factor is the number of distributions/classes one should use. Especially since this choice affects the level of complexity in the estimation *i.e.* as the number of distributions/classes increases the estimation procedure may result in non-convergence. Secondly, each model (normal distribution) as well as the resulting

---

<sup>1</sup> after dietary risks and before smoking

final mixture have infinite support. Whilst this may be desirable for certain applications, it is not the case for BMI. Negative or zero BMI values are nonsensical. Lastly, interpretation of the results can be complex. The weights for the each class specify the probability of an individual (based on covariates) falling into a that class. Hence, the covariates affect the weight assigned to the distribution rather than drive any changes in the distribution itself.

The approach proposed by this paper circumvents these issues. Using the MED approach with covariate information, a single density is produced for a given set of covariates. The estimated density is constructed over a closed interval. In this case, a set of plausible BMI values. For a given set of optimal parameters, the covariate values determine the shape, scale and location of the estimated density. This provides an intuitive explanation in that different covariates affect different moments of the estimated density. Although the idea of using measures pertaining to entropy is not entirely new to BMI studies (Contoyannis and Wildman (2007) and Houle (2010)), the application of MED to model the distribution of BMI using covariates is indeed novel.

Despite the vast amount of literature on BMI, it is important to address the limitations that some health professionals have identified. Given the definition of BMI, it is easier to interpret a change in BMI when only the mass of an individual changes. In most cases, this is associated with an increase in body fat. This is generally the case with adults. However, with children both height and mass can vary and as a result it is more difficult to interpret the change in BMI levels. This is also the case in adults who may have increased their muscle mass i.e. the extra mass does not consist entirely of body fat. Given both these cases, it is possible to exclude children (under 16) from the study and if possible also athletes provided they can be identified. There are nonetheless studies which solely focus on studying the BMI levels in both these groups (Walsh *et al.* (2011), Ortlepp *et al.* (2003) and Beyerlein *et al.* (2008)). Lastly, health professionals have introduced a new measure in 2012 appropriately named Body Shape Index (BSI). It is claimed that this measure is a better indicator of health risks compared to BMI. The definition of BSI contains the waist circumference, height and BMI itself. Given that the latest measure is a function of BMI, it is safe to assume that it still has value as a health indicator. Lastly, national and global health institutions as well as medical personnel continue to use and report BMI statistics for the general population.

The rest of the paper is organized as follows: Section 2 of the paper provides a brief introduction to MED as well as outlines the conditions required for its application. Section 3 contains the details of the model specification and estimation. Most importantly, it proposes a method of incorporating covariates into the MED framework and contains the estimation methodology for the proposed model. Section 4 provides a brief introduction to the data set used in this study. Section 5 contains the estimated model along with some discussion of the results. Lastly, section 6 summarizes the major results of the paper with some points on the future direction of the study.

## 2 MAXIMUM ENTROPY DENSITY

The Maximum Entropy Density (MED) is obtained by maximizing Shannon's information entropy (Shannon (1948)) subject to a set of moment constraints. Jaynes (1957) termed this process as the Principle of Maximum Entropy and provided a continuous version of Shannon's entropy ( $E$ ) which is defined as

$$E = - \int_{\mathbf{A}} f(y) \log f(y) dy. \quad (1)$$

Here  $f(y)$  is a probability density function and  $\mathbf{A}$  represents the set in which the integration occurs. The moment constraints (conditions) used in the optimization are:

$$\begin{aligned} \int_{\mathbf{A}} f(y) dy &= 1. \\ \int_{\mathbf{A}} y^j f(y) dx &= \mu_j \text{ where } j = 1, 2, \dots, k. \end{aligned} \quad (2)$$

By definition,  $\mu_j$  represents the  $j^{th}$  moment of the distribution. Solving this non-linear optimization problem yields the following solution for  $f(y)$ :

$$f(y) = Q^{-1} \exp \left( \sum_{j=1}^k \lambda_j y^j \right) \quad (3)$$

where  $Q = \int_{\mathbf{A}} \exp \left( \sum_{j=1}^k \lambda_j y^j \right) dy$  denotes the normalizing constant. Refer to Chan (2009) for more details on the derivation. From the above derivation, one can see that the resulting density (equation 3) is a special case of the Generalized Exponential (GE) distribution. The  $\lambda_j$  values represent parameters of the MED. These are essentially non-linear functions of the moments (Proposition 2 of Chan (2009)) and are responsible for controlling the shape, scale and location of the distribution.

Maximizing equation (1) using the only the first condition (equation (2)) produces a uniform distribution as a MED. This is expected since no other information is used in constructing the density function. As additional information is added i.e. additional moment conditions, the resulting distribution moves away from the uniform distribution. Furthermore, carrying out the optimization under different moment conditions produces MEDs such as the Exponential, Normal, Log-normal, Pareto, Gamma and Beta distributions. Given this flexibility, the MED is quite useful for approximating empirical densities. As a result, it has been applied to different econometric problems. Some examples include Rockinger and Jondeau (2002), Wu (2003), Park and Bera (2009) and Chan (2009).

As indicated by the Principle of Maximal Entropy, the construction of MED requires use of moments. Specifically, the existence of these population moments is required. In empirical studies, one only has the luxury of a sample. A natural question to ask is that can the sample provide any information about the existence of the population moments. A result by Hill (1975) can be used to verify the existence of the highest population moment available for any given *iid* sample. Subsequently, if the moment does exist then one can argue that the sample moment is a consistent estimator of the population moment. As such, under this condition sample moments can be substituted in place of population moments in the MED derivation process.

Given the non-linear nature of the optimization process, one is required to verify the existence of the solution (MED). Furthermore, given that a solution exists, there is a possibility that it may not be unique. Frontini and Tagliani (1997) showed that a positive determinant of the Hankel matrix (consisting of moments) was a necessary condition for the existence of a MED. The Hankel matrix ( $H_k$ ) is expressed as With regard to uniqueness of solution, the paper by Mead and Papanicolaou (1984) provides the necessary and sufficient conditions that ensure the resulting MED is unique. Additionally, Zellner and Highfield (1988) showed that if the solution is equivalent to equation 3 then it is unique.

### 3 MODEL SPECIFICATION AND ESTIMATION

This paper aims to estimate the distribution of BMI using information from covariates. The MED as expressed in equation 3 contains no covariates. One possible way to incorporate covariates is via conditioning. Hence, estimate the density of BMI for a given set of covariates. The conditioning enables one to focus on the impact a particular set of covariates has on BMI. Let  $y$  denote an individual's BMI and let  $\mathbf{x}$  denote a matrix ( $m \times k$ ) containing  $m$  covariates of the individual. Hence, the conditional density of BMI for an individual can be expressed as

$$f(y|\mathbf{x}) = Q^{-1} \exp \left[ \sum_{j=1}^k \beta^T \mathbf{x} y^j \right] \quad (4)$$

where

$$Q = \int_{\mathbf{A}} \exp \left( \sum_{j=1}^k \beta^T \mathbf{x} y^j \right) dy. \quad (5)$$

is the normalizing constant of the density calculated by integrating the density over a set of all possible BMI values (set  $\mathbf{A}$ ). Here,  $\beta$  denotes the vector of  $p$  parameters of the conditional density. Comparing this specification with equation 3, one can see that the  $\lambda_j$  values are a function of the new parameters ( $\beta$ ) and a set of covariates. As such,  $\beta$  and the covariates govern the shape, scale and location of the BMI distribution. More specifically for fixed  $\beta$  values, changing the covariates will result in changes in the BMI distribution. This specification offers flexibility with regard to how the covariates affect different moments of the BMI distribution. For example, covariates may be transformed and/or combined with other covariates or with intercept terms to model their impact.

The resulting conditional density is a generalized exponential distribution which has a supports negative and positive values. One could rightly argue that BMI cannot be negative and hence this specification may not be

accurate. However, the normalizing constant (equation 5) is obtained by integrating the density over of a set of plausible BMI values. As such, there is a zero probability of obtaining a BMI value outside this set of plausible BMI values.<sup>2</sup> Given the specification for the conditional distribution of BMI (equation 4), the next step is to estimate the parameters of this distribution i.e. the values of  $\beta$ . The method of Maximum Likelihood can be used to carry out this estimation. In this section, the log-likelihood function of the conditional distribution is derived. The values of parameters that maximize this function are considered to be optimal estimates.

For a sample of size  $n$  individuals, the log-likelihood function for equation 4 can be expressed as

$$L(\beta; \mathbf{x}, y) = - \sum_{i=1}^n \log Q_i + \sum_{i=1}^n \sum_{j=1}^k \left( \beta^T \mathbf{x}_i y_i^j \right)$$

Here  $y_i$  denotes the BMI of individual  $i$ ,  $\mathbf{x}_i$  is a matrix consisting of an individual's attributes ( $m$  covariates). The normalizing constant ( $Q_i$ ) is can be written as

$$Q_i = \int_{\mathbf{A}} \exp \left( \sum_{j=1}^k \beta^T \mathbf{x}_i y^j \right) dy.$$

Note that the normalizing constant differs across individuals. In order to compute the parameter estimates, one needs to maximize equation 6 over a set of all possible parameters values. Numerical optimization and integration ( $Q_i$ ) procedures are used to achieve this since no closed form expressions exist when  $k > 2$  (Rockinger and Jondeau (2002)). For computational convenience, the first order derivatives are derived and included in the optimization routine.<sup>3</sup>

#### 4 DATA

The BMI data used in this study has been sourced from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. This is a household-based panel study which began in 2001 and collects information about all individuals in a household. This includes family attributes, economic well being, labour market information, health and subjective well-being, education status as well as a variety of other household and individual variables. Physical attributes such as height and weight have been captured since 2006. This data set is particularly suited to this study because it is the only national level panel data set available in Australia. As such, it is representative of the Australian population. This paper focuses on survey results for the year 2012. This survey had approximately 25,000 individuals from 10,000 households.

#### 5 RESULTS

In order to estimate the MED, a value for  $k$  (equation 4) is required. This specifies the number of MED parameters ( $\beta$ ) to be estimated. The value of  $k$  is chosen based on the existence of moments in the data. Hence, the first step is to estimate the number of moments available in the data. The Hill estimator (Hill (1975)) is used to estimate the tail index ( $\hat{\alpha}$ ) of the BMI distribution. This can then be used to estimate the highest moment available in the BMI distribution. The results indicate that the sixth moment exists. Given this, this study conservatively sets the value of  $k$  to 4 based on the paper by Wu (2003). The results in Wu (2003) provide an insight on the effect of sequentially updating the moment conditions i.e. iteratively including one moment condition at a time in the optimization process. The results show that there is only a marginal improvement in AIC and BIC measures when increasing the values of  $k$  from 4 to 12. Additionally, the interpretation could be an issue with moments higher than 4. Next the conditions for the existence of the MED are verified (section 2). This is done by computing the determinant of the Hankel matrix and ensuring that it is positive.

Finally, covariates are selected from the survey data. This selection process takes into account different type of covariates which may potentially impact BMI levels in individuals. These covariates include examples of physical, economic and social attributes of individuals. This consistent with the approach used in the literature. For instance, a study by Zhang and Wang (2004) examines the relationship between BMI and Gender, Socio-economic inequality, Age and Ethnicity. Similarly, Houle (2010) investigates the effect of Gender, Ethnicity and Education on BMI. A study by Bottai et al. (2014) considers the impact of physical activity on BMI.

<sup>2</sup>Plausible BMI values for this study range from 9 to 100

<sup>3</sup>All computations in this paper were carried out in R

Having attempted a number of model specifications with different covariates, the final (optimal) <sup>4</sup> model specification is:

$$\begin{aligned}\lambda_1 &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ \lambda_2 &= \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \\ \lambda_3 &= \beta_7 X_7 \\ \lambda_4 &= \beta_8 X_8.\end{aligned}$$

Table 1 contains the final covariates and their corresponding estimates ( $\beta$ )<sup>5</sup>.

**Table 1.** Final Model Estimates

Covariate	Estimate
(Log of) Age ( $X_1$ )	0.238735
Active ( $X_2$ )	0.100000
Married ( $X_3$ )	0.000012
Male ( $X_4$ )	-0.000036
(Log of) Household Income ( $X_5$ )	-0.001385
Number of Children ( $X_6$ )	-0.000557
Employment ( $X_7$ )	0.000017
University Education ( $X_8$ )	-0.000003

Inputting these estimates along with a set of covariates pertaining to an individual into equation 4 produces a BMI distribution for that individual. As expected the final model does contain covariates which are significant in other BMI studies. For example, in the study by Brown et al. (2014) all of above covariates were used in their analysis. However, in their study the household income and number of children were not significant. One possible reason for this may be that their approach tracks changes in the mean value of BMI for a given covariate. On the other hand, the approach used in this paper is able to track changes in moments higher than the mean. Hence, the results in this paper show that the household income affects more than just the mean of the distribution i.e. possibly mean and variance. Similarly, the number of children may not impact the mean of BMI distribution as shown in Brown et al. (2014), but does affect higher moments of the BMI distribution.

## 6 CONCLUSION

This paper has proposed a method to model the distribution of BMI using information from the covariates. The application of the MED framework as well the incorporation of covariates into this framework presents a novel approach with regard to BMI modelling. The results clearly show how different covariates affect different aspects (moments) of the BMI distribution. Furthermore, the results also show the shift in the broad BMI categories caused by the marginal changes in the covariates. In terms future direction, a simulation study is currently being carried out to test the finite sample properties of the model. Additionally, the asymptotic properties of the estimator are to be investigated. Finally it is expected that this methodology can be extended to accommodate a panel data setup. This would allow one to assess the change in the distribution of BMI over time.

## ACKNOWLEDGMENTS

The authors would like to thank the two anonymous referees for their helpful suggestions. The authors are grateful for the financial assistance provided by the Australian Research Council.

## REFERENCES

- Beyerlein, A., L. Fahrmeir, U. Mansmann, and A. Toschke (2008). Alternative regression models to assess increase in childhood BMI. *BMC Medical Research Methodology* 8(1).
- Bottai, M., E. A. Frongillo, X. Sui, J. R. O'Neill, R. E. McKeown, T. L. Burns, A. D. Liese, S. N. Blair, and R. R. Pate (2014). Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. *Obesity* 22(5), E149–E156.

<sup>4</sup>based on IC measures

<sup>5</sup>Estimates are significant at 5% level

- Brown, S., W. Greene, and M. N. Harris (2014, March). A New Formulation for Latent Class Models. Working Paper.
- Chan, F. (2009). Modelling time-varying higher moments with maximum entropy density. *Mathematics and Computers in Simulation* 79(9), 2767–2778.
- Contoyannis, P. and J. Wildman (2007). Using relative distributions to investigate the body mass index in England and Canada. *Health Economics* 16(9), 929–944.
- Flegal, K. and R. Troiano (2000). Changes in the distribution of body mass index of adults and children in the US population. *International Journal of Obesity* 24, 807–818.
- Frontini, M. and A. Tagliani (1997). Entropy-convergence in Stieltjes and Hamburger moment problem. *Applied Mathematics and Computation* 88(1), 39–51.
- Hill, B. M. (1975, 09). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* 3(5), 1163–1174.
- Houle, B. (2010). Measuring distributional inequality: Relative body mass index distributions by gender, race/ethnicity, and education in the United Statesx (19992006). *Journal of Obesity* 2010.
- Jaynes, E. T. (1957, May). Information Theory and Statistical Mechanics. *Phys. Rev.* 106(4), 620–630.
- Lin, T., J. Lee, and W. Hsieh (2007). Robust mixture modeling using the skew t distribution. *Statistics and Computing* 17(2), 81–92.
- Mead, L. R. and N. Papanicolaou (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* 25(8), 2404–2417.
- Ortlepp, J., J. Metrikat, M. Albrecht, P. Maya-Pelzar, H. Pongratz, and R. Hoffman (2003). Relation of body mass index, physical fitness, and the cardiovascular risk profile in 3127 young normal weight men with an apparently optimal lifestyle. *International Journal of Obesity* 27, 979–982.
- Park, S. Y. and A. K. Bera (2009). Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics* 150(2), 219–230.
- Penman, A. and W. Johnson (2006). The changing shape of the body mass index distribution curve in the population: Implications for public health policy to reduce the prevalence of adult obesity. *Preventing Chronic Disease* 3.
- Rockinger, M. and E. Jondeau (2002). Entropy densities with an application to autoregressive conditional skewness and kurtosis. *Journal of Econometrics* 106(1), 119–142.
- Shannon, C. (1948). The mathematical theory of communication. *Bell Systems Technical Journal* 27, 349–423.
- Walsh, J., M. Chimstein, I. T. Heazlewood, S. Burke, J. Kettunen, K. Adams, and M. DeBeliso (2011). The loess regression relationship between age and BMI for both Sydney world masters games and the Australian national population. *International Journal of Biological and Medical Sciences* 1, 33–36.
- Wu, X. (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics* 115(2), 347 – 354.
- Zellner, A. and R. A. Highfield (1988, February). Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *Journal of Econometrics* 37(2), 195–209.
- Zhang, Q. and Y. Wang (2004). Socioeconomic inequality of obesity in the United States: Do gender, age, and ethnicity matter? *Social Science and Medicine* 58(6), 1171 – 1180.