Newcastle University

ePrints

Yu K, Liu X, Alhamzawi R, Becker F, Lord J.

Statistical methods for body mass index: a selective review.

***Statistical Methods in Medical Research* 2016**

DOI: http://dx.doi.org/10.1177/0962280216643117

Copyright:

This is the Authors' accepted manuscript of an article that will be published in its final definitive form by Sage, 2016.

DOI link to article:

http://dx.doi.org/10.1177/0962280216643117

Date deposited:

14/07/2016

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License

Newcastle University ePrints - eprint.ncl.ac.uk

Statistical methods for body mass index: a selective review

Keming Yu[1], Xi Liu[1], Rahim Alhamzawi[1], Frauke Becker[2] and Joanne Lord[2]

[1] Department of Mathematics, Brunel University London, Uxbridge, UK
[2] Institute of Environment, Health and Societies, Brunel University London, Uxbridge, UK

Corresponding author:
Keming Yu, Department of Mathematics, Brunel University London,
Uxbridge, UB8 3PH, UK.
Email: keming.yu@brunel.ac.uk

Abstract
Obesity rates have been increasing over recent decades, causing significant concern among policy makers. Excess body fat, commonly measured by body mass index (BMI), is a major risk factor for several common disorders including diabetes and cardiovascular disease, placing a substantial burden on health care systems. To guide effective public health action, we need to understand the complex system of intercorrelated influences on BMI. This paper, based on all eligible articles searched from Global health, Medline and Web of Science databases, reviews both classical and modern statistical methods for BMI analysis. We give a description of each of these methods, exploring the classification, links and differences between them and the reasons for choosing one over the others in different settings.
We aim to provide a key resource and statistical library for researchers in public health and medicine to deal with obesity and BMI data analysis.

Keywords
Body mass index (BMI), obesity, regression model, risk factors, statistical analysis

Introduction

Obesity is considered as one of the most important medical and public health problems of our time1. Excess body fat has been identified as a major risk factor for several common disorders including diabetes and cardiovascular diseases and imposes a substantial burden on health care systems. Recently, the American Medical Association classified obesity itself as a disease (http://www.forbes.com/sites/brucejapsen/2013/06/18/ama-backs-disease-classification-for-obesity/), increasing the focus on its importance as a public health concern. Obesity can be measured in various ways. Body mass index (BMI) is the most commonly used measure of relative weight. It can be used both at individual level to assess body weight in a clinical setting and at population level where it would be impractical or too expensive to measure (excess) body fat accurately and consistently. Based on individual height and weight, BMI is defined as body weight measured in kilograms divided by the square of height in meters: BMI = weight(kg) / height(m)^2.

Applications of BMI often rely on the classification of ‘healthy’ and ‘unhealthy’ segments of the BMI distribution2. The contribution of statistical methods to understanding BMI is in collection, organization, analysis and presentation of data as well as the interpretation of results and prediction. The use of statistical methods in BMI-related research may provide support to:

1. Investigate factors associated with BMI, and identify the relationship between BMI and correlated (potentially causal) factors. Although BMI itself is easy to calculate, the system of underlying contributing factors and their intercorrelation is multifaceted. At the individual level, obesity is caused by a continuously positive energy balance, when more calories are consumed than expended. However, the influences driving individual choices which affect the energy balance are highly complex. Within the UK Government's Foresight Programme, a system map was developed that describes the obesogenic environment of interacting influences on weight gain, without identifying any single dominating factor3. In addition to food and physical activity choices, these influences include biological and medical traits, social and psychological components, as well as effects from the built environment and infrastructure.
2. Investigate how BMI may contribute to morbidity and mortality from various related diseases. In a public health context, BMI can be used as a predictor of unhealthy body weight and related disease at a population level. Applying statistical methods to analyse BMI data may help to identify the correlations between obesity, health characteristics and influencing factors. For example, obesity and inactivity are known as risk factors for type 2 diabetes and evidence has been published on the associations between BMI and (i) different types of cancer4;5 and (ii) risk of gestational diabetes mellitus (GDM)6.
3. Explore the classification of BMI and address uncertainty. BMI is a sufficiently good proxy to capture obesity on population level, but does not provide a direct indication of the distribution of body fat. Statistical methods can provide a way to investigate the distribution of body fat based on BMI datasets. According to Dinsdale et al.,7 the British 1990 growth reference (UK90) is recommended for population monitoring and clinical assessment in children aged four years and over providing centile curves for BMI as a norm against which changes in individual measurements can be monitored.
4. Inform policy making process by evaluation and prediction. A better understanding of how individual characteristics, choices and influences affect body weight and how excessive body fat is associated with increased risk of disease and mortality is essential for the identification of cost-effective interventions. Various statistical designs and methods can be used in the analysis of BMI for improving the understanding of (i) how to target influencing factors in order to change BMI, and (ii) how to predict variation in BMI based on specific factors. Overview of statistical methods Statistical methods to identify patterns and trends within large datasets are now integrated to the development of scientific research in biological and process modelling, personalized healthcare, pharmacology, health economics, and public health policy. In addition to general concerns such as environmental influences, genetics and disease prevalence, statistical issues that necessitate more advanced models for the analysis of BMI data include skewness, outliers and non-response values.

Table 1 provides a summary of a range of statistical methods and models that can be applied to address BMI related issues. Most of these statistical models could be summarized by a single equation:

{EQUATION1}

where Y is the response variable such as BMI or disease incidence, m(Y) is a characteristic of Y such as the mean E(Y), probability related to Y or quantile Q(Y), and h(.) defines a link function such as a logarithm function or a logistic function. bx specifies a linear combination of explanatory variables x, γZ is a collection of random factors taking grouping of data into account, ε is the model error or random noise. Each of f1(.), f2(.) and f3(.) could be a linear function such as f(bx) = bx, a nonlinear function such as $f2(\gamma Z) = I[\gamma Z > 30]$, a

nonparametric regression function such as an unknown smooth function f3(t) over time t, or even a time series such as an autoregressive model.

{ADD TABLE1}

Review of the literature

We conducted a literature review to identify statistical methods that could be applied to BMI based research. The search was conducted in Global health, Medicine and Web of Science databases using the Web of Knowledge interface, searching for the following terms: Topic=(health OR nutrition OR medic*) AND Topic=(statistics analysis* OR regression model* OR BMI data analysis* OR overweight* OR obesity* OR adiposity* OR body composition* OR zBMI*). The search was restricted to research articles in English. No date restrictions were applied.

The search identified more than 100 relevant papers. We classified these according to research questions and modelling methods. Rather than enumerate all papers, we identified the range of methods and selected papers accordingly to illustrate the methods, as described below. Where possible, we also mention software packages for R to ease the application of statistical models. The review aims to produce a representative review of statistical methods used in analyses that involve BMI in some way and provide a useful resource for obesity-related research.

Mean methods: what are they and where to use?

Mean-based methods mainly include correlation analysis and regression models for the mean of the dependent variables. The former is used to measure dependency of BMI on other factors and the latter is modelling the relationship between BMI and other factors. These methods are often the first choices for BMI-based investigation and data analysis and are implemented in all statistical packages.

Correlation analysis

The population correlation coefficient corr(X; Y) between two random variables X and Y with expected values E(X) and E(Y) can be applied for the simple dependency measurement of BMI and other factors. While measurement needs to take into account potential time-variation and repeated measurement of BMI, component analysis based dependency measurers can be used for this purpose. For example, Tangugsorn et al.8 used canonical-correlation analysis (CCA) to demonstrate the relationship of cervicocraniofacial skeletal and upper airway soft tissue morphology to comprehend the complicated pathogenic components in obese (BMI >= 30kg=m2) and non-obese (BMI <30kg=m2) patients. Hu et al.9 used CCA to examine the relationship between obesity, body fat distribution and lipoprotein profiles. Even though canonical variables are artificial, they can often be identified in terms of the original variables.

To identify the variables, one must inspect the standardized coefficients of the canonical variables and the correlations between the canonical variables and their original variables. R-packages CCA can be used in numerical analysis. Other modern dependency measurements such as the Functional Singular Component Analysis developed by Yang et al.10 has been used in quantifying the dependency between BMI and systolic blood pressure (SBP) as an indicator of the general individual health status.

Linear mixed model for mean: Examples

Based on longitudinal data from a population-based mammography screening program introduced between 1987 and 1990 in central Sweden, Newby et al.11 used a linear regression model

{EQUATION2}

with assumption of $E(\varepsilon) = 0$ to estimate the associations between a change in BMI and changes in food patterns. Below is an example of linear mixed model. While a genetic contribution to obesity susceptibility has been identified12, the correlation between longer duration of breastfeeding and the fat mass- and obesity-associated (FTO) gene has been subject to further analysis. Based on cohort data for children who were followed up from birth to 14 years of age13, first set the cut off to be age 1.5 years for all individuals to model BMI denoted as Y in terms of xr representing time-independent covariates, including two FTO genotypes:

{EQUATION3}

where each of the models is a fixed effect model. BMI is continuous across all ages, so that the model with enforced continuity at the cut off may best match the BMI process and then have an equation as a mixed effect model

{EQUATION4}

where the indicator function I[.] indicates age group the child belongs to. The intercept term b0 here represents a random effect intercept due to the heterogeneity of age groups or the difference in BMI growth trajectories across individuals. Accordingly, both b1 and b2 are the random effect coefficients. Let γ = (b0; b1; b(infant2) ; b(child2)) be a collection of these random coefficients then these first four terms of the model can be written as 0Z. Hence, the model above is a standard linear mixed model and also a special case of model (1):

{EQUATION5}

where b= br and x = xr stand for the fixed effects term. We assume that $E(\varepsilon) = 0$ and variance-covariance matrix $Var(\varepsilon) = G$ is known. Model (2) can be extended into a semiparametric model

{EQUATION6}

and a nonparametric model

{EQUATION7}

respectively where f1(.) and f2(.) are nonparametric and non-random functions and f3(_) specifies a random function. Warrington et al.14 used a linear mixed-effects approach to model BMI trajectories in children for genetic association studies by comparing four different mixed-effect models for their data from the Western Australian Pregnancy Cohort. They found that the semiparametric linear mixed model was the most efficient for modelling childhood growth to detect modest genetic effects in this cohort. R packages such as lme4 are available to implement linear mixed-effects models.

Probability models: what are they and where to use?

Probability models for BMI mainly include logistic regression and probabilistic index models. The former is used to model the probability of binary outcomes from BMI and the latter can be used to analyse continuous outcomes on a ratio scale from BMI. Instead of modelling conditional mean directly in mean-based models, the probability models focus on the investigation of the probability of an event such as different types of cancer related to BMI, although both mean and probability of a binary variable means same thing.

Logistic regression

Logistic regression is often used to check the probability of risk with diseases due to BMI, where BMI is an independent variable or predictor. For example, logistic regression can be used to explore the association of BMI with diabetes risk, setting the categorical dependent variable Y as binary, Y = 1 for diabetes, and Y = 0 otherwise. Boffetta et al.15 examined the association in logistic regression models

{EQUATION8}

by employing BMI as an independent categorical variable: ten BMI categories were established. The categories were chosen to improve the ability to investigate the association between BMI and diabetes, in particular at the extremes of the BMI distribution. Razak et al.16 examined the change in BMI across all segments of the BMI distribution in 96 countries, and assessed whether the BMI distribution is changing between cross-sectional surveys conducted at different time points. As the number of survey cycles per country varied between two and five, they used multilevel regression models, between countries and within countries over survey cycles. A logistic multilevel regression such as

{EQUATION9}

which employs the same logistic transformation as binary logistic regression above, can be used to analyze data for participants that are organized at more than one level, where one considers a level-1 outcome, Yij , taking on a value of 1 with conditional probability pij , and uj is a random effect across level 2 units (within a country).

Probabilistic index models

Individual BMI may be affected by several risk factors. The location, skewness and shape of BMI distributions may change with covariate patterns17;18. Probabilistic index models (PIMs) have been proposed19 as a semiparametric framework for modelling the probabilistic index (PI) as a function of covariates. PIMs summarize the covariate effects on the shape of the response distribution, while providing sufficient information on the covariate effect sizes. For the model, Y and Y* are independent random response variables associated with covariate patterns X and X*, respectively, where (X; Y) and (X*; Y*) denote two independent and identically distributed random vectors. The PIM is denoted as

{EQUATION10}

where g is a link function defining the relationship between the PI and a linear predictor. Z is a vector that contains elements from X and X*, Z = X - X*, where X and X* are 0/1 dummies coding for two distinct groups of the population.

Quantile methods: what are they and where to use?

In contrast to the mean methods in Section 2, the quantile methods here model the dependency of quantiles of BMI on covariates. Often the distribution of BMI self or its related variable is typically skewed, so that obesity or other extreme events cannot be represented by mean, but quantile20. For example, if one fits quantile regression to BMI with age as one of covariates, then Figure 1 displays the typically age coefficient and its 95% confidence bands against the BMI quantile α. It provides a clear way to illustrate the effect of age on BMI.

{ADD FIGURE1}

A good introduction to quantile regression and its application can be found in Koenker and Hallock21 and Yu et al.22. Assume a random variable Y has a cumulative distribution function F and F is continuous. Given a probability level $\alpha \in [0; 1]$, the αth quantile of Y, denoted as $Q(\alpha)$, is defined as the inverse function of F(.): $Q(\alpha) = F^{-1}(\alpha)$. And $Q(\alpha)$ can be estimated via the following optimization problem with observations on Y:

{EQUATION11}

where $rho(u) = u(\alpha - I[u \leq 0])$ is a check function and I[.] is an indicator function. The R package quantreg can be used for the optimization above. Applying mean-based regression on the studies of the relationship between sleep duration and BMI has shown inconsistent results, because only the long or short sleep duration has significant impact on BMI. Chen et al.23 re-examined the relationship by quantile regression to account for the potential heterogeneous effect of sleep duration on BMI in different BMI categories and compared estimation results from different types of models. Beyerlein et al. 17 employed different regression approaches to predict childhood BMI by using parental socio-demographic and lifestyle information as well as child data on sex and age. In a similar study, Yang et al. 24 used quantile regression to analyze the relationships between sleep, stress, and obesity by gender. They found that the relationships between BMI and covariates were not constant across the BMI distribution and between women and men. Quantile models mainly include parametric, nonparametric, semiparametric models and density regression approach. All of the models below can be regarded as specific cases of equation (1) with m(Y) = Q(Y). Parametric quantile regression. Under a parametric regression model, _th conditional quantile of Y, such as BMI, given X is modelled as

{EQUATION12}

where $b(\alpha)$ is the regression coefficient. $b(\alpha)$ can be estimated by solving

{EQUATION13}

Examples of parametric quantile modelling include Costa-Font et al.25, who studied the cross-country gap in BMI between Italy and Spain in 2003 by applying a decomposition methodology to the entire BMI distribution, by

{EQUATION14}

where f is vector of food consumption measures, e represents covariates of physical activity, x is a vector of individual characteristics, and e represents residuals due to unobserved

effects. Fenske et al.26 extended the QR framework to an additive model to include k additive effects, which allowed the inclusion of nonlinear effects

{EQUATION15}

where X considers six effects: child's age, duration of breastfeeding, maternal BMI, maternal age, years of maternal education and years of education of the mother's partner, Z is a vector of additional variables, including education, marital status, work stress, behaviours such as smoking and breakfast, diet, etc, and the nonlinear terms, f(α,i) for i = 1, …, k, denote generic functions of Z with nonlinear relationship. Sturm and Datar27 examined the association of BMI among US elementary school children with food price and restaurant density. Stifel and Averett28 used a quantile regression approach to explore the correlates of childhood overweight and ethnicity, gender and other influences in the United States. Popkin29 found a parametric QR with actual BMI data regressed against age and age squared to be the best fit to provide estimates for relations between age and BMI for upper extremes of the BMI distribution.

Nonparametric quantile regression

If the relationship between BMI and other factors or variables is complicated beyond of a parametric model such as a polynomial, then a parametric model is possibly misspecified in this case, nonparametric regression models offer a more flexible way of modelling a relationship than parametric models, but may require more sophisticated methods and large sample sizes. For example, a nonparametric model for BMI would imply that the relationship between BMI and other covariates is unknown but assumed to follow an estimated smooth function. Fitting a smoothing quantile function could be done using a spline according to Koenker et al. 30 and R function rqss. Alternatively, Li et al.31 considered a nonparametric model for age-specific BMI that used a double-kernel-based method and an automatic bandwidth selection procedure. The method employed the basic settings of a double-kernel estimator from Yu and Jones32, which used two local-linear kernels to smooth both variables Y = BMI and T = age with some adaptation for the survey data. Different from spline smoothing, kernel smoothing is a weighted average of all data points around a local area, where the weights are specified using a standard probability function as the kernel function and using a bandwidth to control the local area.

Semiparametric quantile regression

In order to deal with non-normal and non-homogeneous distributions of BMI among different age groups, one of the most successful and most widely applied methods is the LMS (lambda-mu-sigma) model introduced by Cole and his collaborators33–35. The LMS model uses an age-specific Box-Cox power transformation to yield normality. Let μ(T) and σ(T) be the age-specific mean and standard deviation, and λ(T) be the Box-Cox power, then the age-dependent αth regression quantile of the BMI distribution is given by

{EQUATION15}

where Q(α|T) is the conditional αth quantile given T; Φ^-1(.) is the inverse of standard normal distribution. Then smoothing functions μ(T), σ(T) and λ(T) are fitted nonparametrically. Cole et al. have implemented the LMS method in several fundamental BMI studies36–38. The LMS method has been extensively used in weight-related research, and has become a 'standard' framework for studying age-specific BMI or other growth references, along with an application in plotting charts for BMI or other growth references.

For example, Ogden et al.39 compared the U.S. Centers for Disease Control and Prevention (CDC) growth data from 2000 with historical data from 1977 using the LMS method. Onis et al.40 considered the LMS method for the development of BMI cut-offs for both children and adults. Heagerty and Pepe41 considered a semiparametric model for age-specific BMI. Their model was based on the following linear representation of BMI:

{EQUATION16}

where T could be age or time or a continuous variable, μ(T) and σ(T) are the location and scale functions, and ε(T) is the function of the noise term that depends on T. Under this model the αth conditional quantile of T is defined as

{EQUATION17}

where z(α|T) is the αth quantile of ε(T). This model is specified as a parametric model in terms of μ(T) and σ(T), but nonparametric smoothing methods mentioned in Section 4.3 are used to fit both μ(T) and σ(T). An alternative and popular semiparametric quantile regression model uses a normal transformation-based approach.

Density regression approach

Dunson et al.42 proposed a density regression to study the association between Luteinizing hormone (LH) and BMI in randomly selected women defined by:

{EQUATION17}

where BMI is an independent variable. The aim of this study was to identify how changes in LH may affect the BMI distribution while adjusting for the potentially confounding effect of age. The conditional BMI distribution was not assumed to be normally distributed, but could be regarded as a mixture of conditional normal densities because mixtures of a sufficiently large number of normal densities can be used to approximate any smooth density accurately. Where the weights wk could or could not depend on x, and could be inference by classical methods or Bayesian methods, although Dunson et al.42 assumed their dependency of x and used nonparametric Bayesian inference.

Data quality issues: what are they and how to address?

Data quality issues include missing values, measurement error and non-response. Our review found that the methods used to deal with BMI related missing values are multiple imputation (MI) methods and ignorable likelihood (IL) method; the methods dealing with BMI related measurement error are instrumental variable approach and auxiliary data, and the methods dealing with BMI related non-response are two-stage clustering and Bayesian inference.

Multiple imputation methods for missing values

Elliott and Stettler43 used survey data and a mixture model based multiple imputation to obtain the BMI distribution for a paediatric population in the presence of clerical errors. The mixture model is defined by latent classes that have common means, conditional on age and health centre to accommodate the disproportional sample design, but differing covariances. The clerical error class is the class with the largest covariance matrix determinant. The general approach for multiple imputation is as follows. For the ith individual, i = 1, …, n let Zi be a q-dimensional outcome of interest. Each individual value is assumed to depend on a set of p covariates xi. The associated covariance is given by the individual's latent variance

class membership, which is denoted by the unobserved latent variable Ci, where Ci = K indicates that the ith individual belongs to the clerical error class with the largest variability. The Ci can be regarded as missing for all subjects belonging to the clerical error class. The complete-data mixture model considered is

{EQUATION18}

Once priors are postulated for model parameters β, Σ(k) and others such as (p(1), …, p(K)) ~DIRICHLET(1, 1, …, 1), proper Bayesian inference such as Gibbs sampler can be applied. All missing values are then replaced by imputed values from the Gibbs sampling procedure.

Ignorable likelihood method for missing values

In a study by Little and Zhang44, some variables of interest, including BMI, were subject to missing data. They applied an ignorable likelihood (IL) method to the subsample of observations that are complete on one set of variables, but possibly incomplete on others multiply impute the full sample, and then used it for regression analyses. Results for different imputing methods were compared and yielded similar estimates for the effect of household income and education on blood pressure.

In general, the IL approach requires a model for the distribution of covariates W and an outcome variable Y, both with missing values, given a fully observed covariate set Z, indexed by parameters θ, for example p(wi; yi|zi; θ), where fully observed covariates can be treated as fixed45. Integrating the missing variables out of the joint distribution and treating θ as the argument of the resulting density yields the IL

{EQUATION19}

where ($w_{obs;i}$; $y_{obs;i}$) are the observed components of (wi; yi).

Instrumental variable approach for measurement error

Many survey data sets rely on self-reported measures of BMI. It is suggested that people often underreport their weight and overstate their height. So BMI is measured with error. Although there are many instrument variable estimation procedures dealing with measurement errors, the typical instrument in BMI-based analysis is to use the BMI of a sibling or other relatives to instrument for the respondents BMI on the assumption that this should pick up genetic and environmental factors, which is easy to implement and particularly suitable for BMI-based big data analysis. See Bound et al.46 for the details.

Auxiliary data for measurement error

Auxiliary data, which contain information about the conditional distribution of the true variables given the mismeasured variable, are often used in dealing with measurement errors. This type of auxiliary data is easy to get for BMI-based analysis, so that researchers in BMI usually adopt simple methods. For instance, Cawley47 corrected for measurement error in reported BMI by predicting true height and weight by using information on the relationship between true and reported values in the Third National Health and Nutrition Examination survey. Here he treated BMI as an independent variable.

Two-stage clustering for non-response Data may be missing due to non-response. Rubin48 and Little49 classified non-response mechanisms into three types: missingness completely at random (MCAR), when the probability of the non-response does not depend on clusters or survey variables; missingness at random (MAR), when the probability of response depends only on the observed values; and non-ignorable non-response, when the probability of non-

response depends on unobserved values. In their analysis, Yuan and Little50 dealt with a unit nonresponse rate about 40% when households failed to answer questions in a questionnaire. To assess the relationship between cluster response rates and cluster means, they plotted cluster sample response rates against cluster sample means of log(BMI), which displayed a slightly linear trend with a correlation coefficient of 0.32, suggesting that the non-response mechanism is not missingness completely at random (MCAR), and a cluster-specific non-ignorable (CSNI) non-response mechanism may be indicated. Based on a logarithm transformation of BMI measurements, they proposed several model-based estimates of the finite population mean for two-stage samples with unit nonresponse and compared them with existing methods by a simulation study. These models include: (1) applying standard two-stage mean estimators from complete response observations51 to non-response observations; (2) discarding non-respondents and basing estimates on predictions from a random-effects model fitted to respondents; (3) adding noninformative priors for the fixed parameters of a randomeffects model and simulating draws from the posterior distribution of the parameters.

Bayesian method for non-response

In order to deal with serious non-response and selection bias due to missing BMI values for a considerable number of individuals, Nandram and Choi52 used differential probabilities for selection of these individuals A nonignorable non-response model was proposed to estimate the finite population means of covariates where the log(BMI) values were used to obtain more normally distributed data. The model included a spline regression of log(BMI) on age, adjusted for several individual characteristics. Their data contained information on Ni individuals for i countries. The authors assumed that the response indicator rij for jth individual within the ith country related to BMI value xij via

{EQUATION20}

and employed a hierarchical setting for the regression parameters above. These xij are a regression function of other covariates such as age, ethnicity and sex. Within the Bayesian framework, the sampled non-respondent BMI values are then obtained from their conditional posterior densities in the Metropolis-Hastings algorithm, and the non-sampled BMI values are drawn from their conditional posterior densities.

Data complexity issues: what are they and how to address?

Many scientific questions related to BMI can be answered by the analysis of cross-sectional data. However, BMI and associated factors may vary over time, resulting in repeated observations of the same variable at different time points, or anthropometric measurements required for BMI-related research are naturally longitudinal observations correlated to other factors. If longitudinal data are available, specific statistical methods or models are needed to take account of the nature of data and fully use the information provided. In general, complex data issues may arise due to the functional nature of the data, dimensional problems in modelling due to a high number of factors. The linear mixed models described in Section 2.2 are useful in longitudinal data analysis. This section provides additional approaches used in these contexts of BMI-based data analysis, which include Generalized estimating equations (GEE), Generalized method of moments (GMM) and Generalized additive model (GAM).

Generalized estimating equations (GEE)

Remmers et al.53 examined longitudinal relationships between PA and BMI z-scores by using GEE for the analysis. Remmers et al.53 and Branum et al.54 applied GEE in the analysis of BMI data by taking account of individual information being stratified by gender

and baseline weight status. GEE55 can be considered as a method for combining certain estimating equations in presence of time-dependent covariates. Given a mean model, mij subject to unknown parameter vector β, and a variance structure, Vi, the estimating equation is described as:

{EQUATION21}

The parameter estimates solve U(β) = 0. The GEE algorithm has been incorporated into many major statistical software packages, including SAS, STATA, SPSS and R.

Generalized method of moments (GMM)

Lai and Small56 analysed the relationship between BMI and future morbidity among children using longitudinal data with time-dependent covariates. They found that some of the estimating equations combined by GEEs with an independent correlation structure are not valid. The authors distinguished between three types of time-dependent covariates and provided a test for whether a time-dependent covariate is of a certain type. Results indicated that when a covariate is of type I or II, valid estimating equations are available that are not exploited by GEEs assuming an independent correlation structure. As a likelihood analysis is impossible or extremely difficult in this case, and to make optimal use of the valid estimating equations, they use the generalized method of moments (GMM)57.

Generalized additive model (GAM)

Generalized additive models (GAM) combine properties of generalized linear models with additive models58. Each additive term is typically modelled as nonparametric function. Gregory59 used GAM to analyse how wages are affected by BMI and age. They modelled both BMI and wages nonparametrically by use of an oracle estimator. This study, however, focused on young workers and did not examine whether the effect of obesity changes as people age.

Variable selection

Modern variable selection methods including ridge regression, bridge regression60, least absolute shrinkage and selection operator (LASSO)61, elastic net62 and clipped absolute deviation method63 can be used for factor selection when covariates available for the analysis of BMI and the extent of their correlation may vary substantially between groups and settings64.

The LASSO technique is particularly suitable when the number of variables p exceeds the number of observations n, i.e. where p > n poses a problem for the regression analysis. However, although Lasso is popular for its mathematical performance, it is not robust to skewed distributions. For highly skewed distributions, such as diabetes prevalence in the population, robust regression methods such as least absolute deviation (LAD) and quantile regression methods have received considerable attention recently in variable selection methods65;66.

Conclusions

Obesity rates have been increasing over recent decades, causing significant concern among policy makers. Understanding which factors influence individual body weight when BMI is taken as the dependent variable, and how exactly excess body fat is contributing to increased risk for disease, when BMI is regarded as one of independent variables involved, may help to reduce the increased prevalence of several common disorders associated with obesity, thereby lessening the burden placed on health care systems. Use of appropriate statistical methods is essential to produce high quality research that can inform public health policy. Depending on

the specific policy concern, research question or data available for analysis, both classical and modern methods can be used to improve the understanding of the complex system of intercorrelated influences on BMI. Since the choice of a specific method and its implementation may be challenging, this paper aimed to give an overview of available methods and provide a key resource and statistical library for researchers in public health and medicine to deal with obesity and BMI data analysis.

Acknowledgements
We thank Dr Zhuo Sheng and Mr Nicola Attard-Montalto for collecting some references of this review. This work has been supported in part by the National Institute for Health Research Method Grant (NIHR-RMOFS-2013-03-09) and the National Natural Science Foundation of China (Grant No. 71490725, 11261048, 11371322).

References
1. Ells L and Cavill N. Preventing childhood obesity through lifestyle change interventions. A briefing paper for commissioners. Oxford:National Obesity Observatory, 2009.
2. Organization WH. Obesity : preventing and managing the global epidemic : report of a who consultation. Technical report, World Health Organization, 2000.
3. Vandenbroeck I, Goossens I and Clemens M. Foresight tackling obesities: Future choices - building the obesity system map. Technical report, Government Office for Science, UK Government's Foresight Programme, 2007.
4. Renehan A, Tyson M and Egger Mea. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. The Lancet 2008; 371: 569-578.
5. Larsson S and Wolk A. Excess body fatness: an important cause of most cancers. The Lancet 2008; 371: 536–537.
6. Torloni M, Betr´an A and Horta Bea. Prepregnancy bmi and the risk of gestational diabetes: a systematic review of the literature with meta-analysis. Obesity Reviews 2009; 10: 194–203.
7. Dinsdale H, Ridler C and Ells L. A simple guide to classifying body mass index in children. Technical report, National Obesity Observatory, 2011.
8. Tangugsorn V, Krogstad O and Espeland Lea. Obstructive sleep apnea: a canonical correlation of cephalometric and selected demographic variables in obese and nonobese patients. The Angle Orthodontist 2001; 71: 23–35.
9. Hu D, Hannah J and Gray Rea. Effects of obesity and body fat distribution on lipids and lipoproteins in nondiabetic American indians: The strong heart study. Obesity research 2000; 8: 411-421.
10. Yang W, M¨uller H and Stadtm¨uller U. Functional singular component analysis. J R Statist Soc B 2011; 73: 303–324.
11. Newby P, Weismayer C and A° kesson Aea. Longitudinal changes in food patterns predict changes in weight and body mass index and the effects are greatest in obese women. The Journal of nutrition 2006; 136: 2580–2587.
12. Frayling T, Timpson N andWeedon Mea. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. American Association for the Advancement of Science 2007; 316: 889–894.
13. Abarin T, Wu Y and Warrington Nea. The impact of breastfeeding on fto-related bmi growth trajectories: an application to the raine pregnancy cohort study. International journal of epidemiology 2012; 41: 1650–1660.
14. Warrington N, Wu Y and Pennell Cea. Modelling bmi trajectories in children for genetic association studies. PloS one 2013; 8: e53897.

15. Boffetta P, McLerran D and Chen Yea. Body mass index and diabetes in asia: a cross-sectional pooled analysis of 900,000 individuals in the asia cohort consortium. PloS one 2011; 6:e19930.
16. Razak F, Corsi D and Subramanian S. Change in the body mass index distribution for women: analysis of surveys from 37 low-and middle-income countries. PLoS medicine 2013; 10: e1001367.
17. Beyerlein A, Fahrmeir L and Mansmann Uea. Alternative regression models to assess increase in childhood bmi. BMC medical research methodology 2008; 8: 59.
18. Beyerlein A, Toschke A and von Kries R. Breastfeeding and childhood obesity: Shift of the entire bmi distribution or only the upper parts&quest. Obesity 2008; 16: 2730–2733.
19. Thas O, Neve J and Clement Lea. Probabilistic index models. J R Statist Soc B 2012; 74: 623–671.
20. Bottai M, Frongillo E and Sui Xea. Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. Obesity 2014; 22: E149–CE156.
21. Koenker R and Hallock K. Quantile regression: An introduction. Obesity 2001; 15: 43-56.
22. Yu K, Lu Z and Stander J. Quantile regression: applications and current research areas. J R Statist Soc D 2003; 52: 331–350.
23. Chen C, Chang C and Yeh C. A quantile regression approach to re-investigate the relationship between sleep duration and body mass index in taiwan. International journal of public health 2012; 57: 485–493.
24. Yang T, Matthews S and Chen V. Stochastic variability in stress, sleep duration, and sleep quality across the distribution of body mass index: Insights from quantile regression. International journal of behavioral medicine 2014; 21: 282–291.
25. Costa-Font J, Fabbri D and Gil J. Decomposing body mass index gaps between mediterranean countries: A counterfactual quantile regression analysis. Economics & Human Biology 2009; 7: 351–365.
26. Fenske N, Kneib T and Hothorn T. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. J Am Statist Assoc 2011; 106: 494–510.
27. Sturm R and Datar A. Body mass index in elementary school children, metropolitan area food prices and food outlet density. Public health 2005; 119: 1059–1068.
28. Stifel D and Averett S. Childhood overweight in the united states: A quantile regression approach. Economics & Human Biology 2009; 7: 387–397.
29. Popkin B. Recent dynamics suggest selected countries catching up to us obesity. The American journal of clinical nutrition 2010; 91: 284S–288S.
30. Koenker R, Ng P and Portnoy S. Quantile smoothing splines. Biometrika 1994; 81: 673–680.
31. Li Y, Graubard B and Korn E. Application of nonparametric quantile regression to body mass index percentile curves from survey data. Statistics in medicine 2010; 29: 558–572.
32. Yu K and Jones M. Local linear quantile regression. J Am Statist Assoc 1998; 93: 228–237.
33. Cole T. Fitting smoothed centile curves to reference data. J R Statist Soc A 1998; 151: 385–418.
34. Cole T. The lms method for constructing normalized growth standards. European journal of clinical nutrition 1990; 44: 45–60.
35. Cole T and Green P. Smoothing reference centile curves: The lms method and penalized likelihood. Statistics in medicine 1992; 11: 1305–1319.

36. Cole T, Freeman J and Preece M. Body mass index reference curves for the uk, 1990. Archives of disease in childhood 1995; 73: 25–29.
37. Cole T, Bellizzi M and Flegal Kea. Establishing a standard definition for child overweight and obesity worldwide: international survey. Bmj 2000; 320: 1240.
38. Cole T, Flegal K and Nicholls Dea. Body mass index cut offs to define thinness in children and adolescents: international survey. Bmj 2007; 335: 194.
39. Ogden C, Kuczmarski R and Flegal Kea. Centers for disease control and prevention 2000 growth charts for the united states:improvements to the 1977 national center for health statistics version. Pediatrics 2002; 109: 45–60.
40. Onis M, Onyango A and Borghi Eea. Development of a who growth reference for school-aged children and adolescents. Bulletin of the World Health Organization 2007; 85: 660–667.
41. Heagerty P and Pepe M. Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in us children. J R Statist Soc C 1999; 48:533–551.
42. Dunson D, Pillai D and Park J. Bayesian density regression. J R Statist Soc B 2007; 69: 163–183.
43. Elliott M and Stettler N. Using a mixture model for multiple imputation in the presence of outliers: the ‘healthy for life’ project. J R Statist Soc C 2007; 56: 63–78.
44. Little R and Zhang N. Subsample ignorable likelihood for regression analysis with missing data. J R Statist Soc C 2011; 60: 591–605.
45. Little R and Rubin D. Statistical analysis with missing data. USA: Wiley, 2002.
46. Bound J, Brown C and Mathiowetz N. Measurement error in survey data. Handbook of econometrics 2001; 5: 3705–3843.
47. Cawley J. The impact of obesity on wages. Journal of Human Resources 2004; 39: 451–474.
48. Rubin D. Inference and missing data. Biometrika 1976; 63:581–592.
49. Little R. A class of pattern-mixture models for normal incomplete data. Biometrika 1994; 81: 471–483.
50. Yuan Y and Little R. Model-based estimates of the finite population mean for two-stage cluster samples with unit nonresponse. J R Statist Soc C 2007; 56: 79–97.
51. Horvitz D and Thompson D. A generalization of sampling without replacement from a finite universe. J Am Statist Assoc 1952; 47: 663–685.
52. Nandram B and Choi J. A bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. J Am Statist Assoc 2010; 105: 120–135.
53. Remmers T, Sleddens E and Gubbels J. Relationship between physical activity and the development of bmi in children. Medicine and science in sports and exercise 2013; 46: 177–184.
54. Branum A, Parker J and Keim Sea. Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings. American journal of epidemiology 2011; 174: 1159–1165.
55. Liang K and Zeger S. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 13–22.
56. Lai T and Small D. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. J R Statist Soc B 2007; 69: 79–99.
57. Hansen L. Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society 1982; 50: 1029–1054.
58. Hastie T and Tibshirani R. Generalized Additive Models. London: Chapman and Hall, 1990.

59. Gregory C. Wages, bmi, and age: A generalized additive model using the oracle estimator. Available at SSRN 2011; http://ssrn.com/abstract=1975044 or http://dx.doi.org/10.2139/ssrn.1975044.
60. Frank L and Friedman J. A statistical view of some chemometrics regression tools. Technometrics 1993; 35: 109–135.
61. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1996; 58: 267–288.
62. Zou H and Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005; 67: 301–320.
63. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Statist Assoc 2001; 96: 1348–1360.
64. Hesterberg T, Choi N and Meier Lea. Least angle and penalized regression: A review. Statistics Surveys 2008; 2: 61–93.
65. Bradic J, Fan J and Wang W. Penalized composite quasilikelihood for ultrahigh dimensional variable selection. J R Statist Soc B 2008; 73: 325–349.
66. Wu Y and Liu Y. Variable selection in quantile regression. Statistica Sinica 2009; 19: 801.