

Statistical methods for body mass index: a selective review

Keming Yu^{*1}, Xi Liu¹, Rahim Alhamzawi¹, Frauke Becker², and Joanne Lord²

¹*Department of Mathematics, Brunel University London*

²*Institute of Environment, Health and Societies, Brunel University London, Uxbridge, UB8 3PH, UK*

Summary

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.

Keywords:

Body mass index (BMI); obesity; regression model; risk factors; statistical 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).

^{*} *Address for correspondence:* Department of Mathematics, Brunel University London, UK.
E-mail: keming.yu@brunel.ac.uk

Conflict of interest statement

None of the authors have declared a conflict of interest.

1 Introduction

Obesity is considered one of the most important medical and public health problems of our time (Ells and Cavill, 2009). 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:

$$BMI = \frac{mass(kg)}{height(m)^2}.$$

Applications of BMI often rely on the classification of ‘healthy’ and ‘unhealthy’ segments of the BMI distribution (W. H. Organization, 2000). 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 factor (Vandenbroeck et al., 2007). 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 risk factors for type 2 diabetes and evidence has been published on the associations between BMI and (i) different types of cancer (Renehan et al., 2008; Larsson and Wolk, 2008) and (ii) risk of gestational diabetes mellitus (GDM) (Torloni et al., 2009).

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. (2011), 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 for the analysis of BMI 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.

1.1 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 health-care, 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:

$$h(m(Y)) = f_1(\boldsymbol{\beta}' \mathbf{x}) + f_2(\boldsymbol{\gamma}' \mathbf{Z}) + f_3(t) + \epsilon, \quad (1)$$

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 $h(\cdot)$ defines a link function such as a logarithm function or logistic function. $\boldsymbol{\beta}' \mathbf{x}$ specifies a linear combination of explanatory variables \mathbf{x} , $\boldsymbol{\gamma}' \mathbf{Z}$ is a collection of random factors taking grouping of data into account. ϵ is the model error or random noise. Each of $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ could be a linear function such as $f(\boldsymbol{\beta}' \mathbf{x}) = \boldsymbol{\beta}' \mathbf{x}$, a nonlinear function such as $f_2(\boldsymbol{\gamma}' \mathbf{Z}) = I[\boldsymbol{\gamma}' \mathbf{Z} > 30]$, a nonparametric regression function such as an unknown smooth function $f_3(t)$ over time t , or even a time series such as an autoregressive model.

1.2 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, Medline 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 obesity*). 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. The detailed results of the BMI studies via different statistical methods and models are summarized in Table 2.

2 Mean methods: what are they and where to use?

Mean-based methods mainly includes correlation analysis and linear regression modelling. The former is used to measure dependency of BMI and other factors and the latter is modelling the relationship between BMI and other factors. Both approachers are in terms of expected value or mean. These methods are often the first choices for BMI-based investigation and data analysis and are implemented in all statistical packages.

2.1 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. (2001) used canonical-correlation analysis (CCA) to demonstrate the relationship of cervico-craniofacial skeletal and upper airway soft tissue morphology to comprehend the complicated pathogenic components in obese ($BMI \geq 30kg/m^2$) and non-obese ($BMI < 30kg/m^2$) patients. Hu et al. (2000) used CCA to examine the relationship between obesity, body fat distribution and lipoprotein profiles. 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. (2011) have been used in quantifying the dependency between BMI and systolic blood pressure (SBP) as an indicator of the general individual health

status.

2.2 Linear regression modelling : Examples

Based on longitudinal data from a population-based mammography screening program introduced between 1987 and 1990 in central Sweden, Newby et al. (2006) used a linear regression model

$$Y = \boldsymbol{\beta}'\mathbf{x} + \epsilon,$$

with assumption of $E(\epsilon) = 0$ 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 identified (Yang et al., 2007), 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 age, Abarin et al. (2012) first set the cut off to be age 1.5 years for all individuals to model BMI denoted as Y in terms of \mathbf{x}_r representing time-independent covariates, including two FTO genotypes:

$$Y = \begin{cases} \beta_{0,0} + \beta_{1,0}age + \mathbf{x}_r\boldsymbol{\beta}'_r & : age < 1.5 \\ \beta_{0,1} + \beta_{1,1}age + \mathbf{x}_r\boldsymbol{\beta}'_r & : age \geq 1.5, \end{cases}$$

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

$$Y = \beta_0 + \beta_1(age - 1.5) + I_{(age < 1.5)}\beta_2^{infant}(age - 1.5)^2 + I_{(age \geq 1.5)}\beta_2^{child}(age - 1.5)^2 + \mathbf{x}'_r\boldsymbol{\beta}_r + \epsilon.$$

Where the indicator function $I[\cdot]$ indicates age group the child belongs to. The intercept term β_0 here represents a random effect intercept due to the heterogeneity of age groups or the difference in BMI growth trajectories across individuals. Accordingly, both β_1 and β_2 are the random effect coefficients. Let $\boldsymbol{\gamma} = (\beta_0, \beta_1, \beta_2^{infant}, \beta_2^{child})'$ be a collection of these random coefficients then these first four terms of the model can be written as $\boldsymbol{\gamma}'\mathbf{Z}$. Hence, the model

above is a standard linear mixed model and also a special case of model (1):

$$Y = \boldsymbol{\gamma}'\mathbf{Z} + \boldsymbol{\beta}'\mathbf{x} + \epsilon, \quad (2)$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_r$ and $\mathbf{x} = \mathbf{x}_r$ stand for the fixed effects term. We assume that $E(\epsilon) = 0$ and variance-covariance matrix $Var(\epsilon) = G$ is known.

Model (2) can be extended into a semiparametric model

$$Y = X\beta + Zu + f_3(t) + \epsilon,$$

and a nonparametric model

$$Y = f_1(X\beta) + f_2(Zu) + f_3(t) + \epsilon,$$

respectively where $f_1(\cdot)$ and $f_2(\cdot)$ are nonparametric and non-random functions and $f_3(\cdot)$ specifies a random function.

Warrington et al. (2013) 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.

3 Probability models : what are they and where to use?

Probability models for BMI mainly includes 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. In contrast to mean-based model and analysis, the probability models focus on the investigation of the probability of an event such as cancer related to BMI.

3.1 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. (2011) examined the association in logistic regression models

$$\text{logit}(P[Y = 1]) = \log\left(\frac{P[Y = 1]}{P[Y = 0]}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m,$$

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. (2013) 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 use multilevel regression models, between countries and within countries over survey cycles.

A logistic multilevel regression such as

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta x_{ij} + u_j Z_{ij},$$

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, Y_{ij} , taking on a value of 1 with conditional probability p_{ij} , and u_j is a random effect across level 2 units (within a country).

3.2 Probabilistic index models

Individual BMI may be affected by several risk factors. The location, skewness and shape of BMI distributions may change with covariate patterns (Beyerlein et al., 2008, Beyerlein et al., 2008). Probabilistic index models (PIMs) have been proposed (Thas et al., 2012) as a semi-parametric 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 $Pr[Y < Y^* | X, X^*] = g^{-1}(Z^T \beta)$ 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.

4 Quantile methods: what are they and where to use?

In contrast to mean method in Section 2, quantile methods mainly model the dependency of quantile of BMI on others. Because extreme BMI such as obesity can not be represented by mean, but quantile due to the typical skewness of the distribution of BMI (Bottai et al., 2014).

A good introduction to QR can be found in Koenker and Hallock (2001) and Yu et al. (2003). Assume a random variable Y has a cumulative distribution function F and F is continuous. Given a probability level $\alpha \in [0, 1]$, the α th quantile of Y , denoted as $Q(\alpha)$, is defined as the inverse function of $F(\cdot)$: $Q(\alpha) = F^{-1}(\alpha)$. And $Q(\alpha)$ can be estimated via the following optimization problem with observations on Y :

$$\min_Y \sum_i \rho_\alpha(Y_i - Q(\alpha)),$$

where $\rho_\alpha(u) = u(\alpha - I[u \leq 0])$ is a check function and $I[\cdot]$ 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. (2012) 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. (2008) 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. (2014) 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.

All of the models below can be regarded as specific cases of equation (1) with $m(Y) = Q(Y)$. Quantile models mainly include parametric, nonparametric, semiparametric models and density regression approach.

4.1 Parametric quantile regression

Under a parametric regression model, α th conditional quantile of Y , such as BMI, given X is modelled as

$$Q_Y(\alpha|X) = X'\beta(\alpha),$$

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

$$\hat{\beta}(\alpha) = \operatorname{argmin}_{\beta} \sum_i \rho_{\alpha}(Y_i - X_i'\beta). \quad (3)$$

Examples of parametric quantile modelling include Costa-Font et al. (2009), 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

$$\text{BMI} = x'\beta + f'\gamma + e'\delta + \epsilon,$$

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

Fenske et al. (2011) extended the QR framework to an additive model to include k additive effects, which allows the inclusion of nonlinear effects

$$Q_Y(\alpha|X, Z) = X'\beta + \sum_{i=1}^k f_{\alpha,i}(Z),$$

where X considers six effects: child's age, duration of breastfeeding, maternal BMI, maternal age, years of maternal education, 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, $f_{\alpha,i}$, for $i = 1, \dots, k$, denote generic functions of

Z with nonlinear relationship. Sturm and Datar (2005) examined the association of BMI among US elementary school children with food price and restaurant density. Stifel and Averett (2009) used a quantile regression approach to explore the correlates of childhood overweight and ethnicity, gender and other influences in the United States. Popkin (2010) 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.

4.2 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 (1994) and R function *rqss*. Alternatively, Li et al. (2010) considered a nonparametric model for age-specific BMI that uses a double-kernel-based method and an automatic bandwidth selection procedure. The method employs the basic settings of a double-kernel estimator from Yu and Jones (1998), which uses 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.

4.3 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 collaborators (Cole, 1988, 1990; Cole and Green, 1992). The LMS model uses an age-specific Box-Cox power transformation to yield normality. Let $\mu(T)$ and $\sigma(T)$ be the age-specific mean and standard deviation, and $\lambda(T)$ the Box-Cox

power, then the age-dependent α th regression quantile of the BMI distribution is given by

$$Q(\alpha|T) = \mu(T) \left(1 + \Phi^{-1}(\alpha)\lambda(T)\sigma(T)\right)^{1/\lambda(T)}, \quad (4)$$

where $Q(\alpha|T)$ is the conditional α th quantile given T ; $\Phi^{-1}(\cdot)$ is the inverse of standard normal distribution. Then smoothing functions $\mu(T)$, $\sigma(T)$ and $\lambda(T)$ are fitted nonparametrically.

Cole et al. have implemented the LMS method in several fundamental BMI studies (Cole et al, 1995, Cole et al, 2000 and Cole et al, 2007). 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 (2002) compare 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 (2007) consider the LMS method for the development of BMI cut-offs for both children and adults.

Heagerty and Pepe (1999) considered a semiparametric model for age-specific BMI. Their model is based on the following linear representation of BMI:

$$Y = \mu(T) + \sigma(T)\epsilon(T),$$

where T could be age or time or a continuous variable, $\mu(T)$ and $\sigma(T)$ are the location and scale functions, and $\epsilon(T)$ is the function of the noise term that depends on T . Under this model the α th conditional quantile of T is defined as

$$Q(\alpha|T) = \mu(T) + \sigma(T)z(\alpha|T),$$

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

4.4 Density regression approach

Dunson et al (2007) proposed a density regression to study the association between Luteinizing

hormone (LH) and BMI in randomly selected women defined by:

$$f(y|\mathbf{x}) = \sum_{k=1}^{\infty} w_k(\mathbf{x})N(\mathbf{x}'\boldsymbol{\beta}, \sigma^2).$$

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 w_k could or could not depend on \mathbf{x} , and could be inference by classical methods or Bayesian methods, although Dunson et al (2007) assumed their dependency of \mathbf{x} and used nonparametric Bayesian inference.

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

Data quality issues include missing values, measurement error and non-response. The methods dealing with missing values are Multiple imputation (MI) methods and ignorable likelihood (IL) method; the methods dealing with measurement error are instrumental variable approach and auxiliary data and the methods dealing with non-response are two-stage clustering and Bayesian inference.

5.1 Multiple imputation methods for missing values

Elliott and Stettler (2007) 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 i th individual, $i = 1, \dots, n$, let \mathbf{Z}_i be a q -dimensional outcome of interest. Each individual value is assumed to

depend on a set of p covariates \mathbf{x}_i . The associated covariance is given by the individual's latent variance class membership, which is denoted by the unobserved latent variable C_i , where $C_i = K$ indicates that the i th individual belongs to the clerical error class with the largest variability. The C_i can be regarded as missing for all subjects belonging to the clerical error class. The complete-data mixture model considered is

$$\begin{aligned} \mathbf{Z}_i | C_i = k &\sim N_q(\mu_i, \Sigma_k), \\ C_i &\sim \text{MULTI}(1, p_1, \dots, p_k) \end{aligned}$$

Once priors are postulated for model parameters β , Σ_k and others such as $(p_1, \dots, p_K) \sim \text{DIRICHLET}(1, 1, \dots, 1)$, proper Bayesian inference such as Gibbs sampler can be applied. All missing values are then replaced by imputed values from the Gibbs sampling procedure.

5.2 Ignorable likelihood method for missing values

In a study by Little (2011), some variables of interest, including BMI, were subject to missing data. They applied an ignorable likelihood (IL) method to 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 $p(w_i, y_i | z_i, \theta)$, where fully observed covariates can be treated as fixed (Little and Rubin, 2002). Integrating the missing variables out of the joint distribution and treating θ as the argument of the resulting density yields the IL

$$L_{ign}(\theta) \propto \prod_{i=1}^n p(w_{obs,i}, y_{obs,i} | z_i, \theta),$$

where $(w_{obs,i}, y_{obs,i})$ are the observed components of (w_i, y_i) .

5.3 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. The

typical instrument in BMI-based analysis is to use the BMI of a sibling or other relative to instrument for the respondents BMI on the assumption that this should pick up genetic and environmental factors. see Bound (2001) for the details.

5.4 Auxiliary data for measurement error

Cawley (2004) corrects 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.

5.5 Two-stage clustering for non-response

Data may be missing due to non-response. Rubin (1976) and Little (1994) 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 (2007) dealt with a unit non-response 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 $\log(\text{BMI})$, 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 non-response and compared them with existing methods by a simulation study. These models include: (1) applying standard two-stage mean estimators from complete response observations (Horvitz, 1952) to non-response observations; (2) discarding non-respondents and basing estimates on predictions from a random-effects model fitted to respondents; (3) adding non-informative priors for the fixed parameters of a random-effects model and simulating draws from the posterior

distribution of the parameters.

5.6 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 (2010) used differential probabilities for selection of these individuals. A non-ignorable 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 N_i individuals for i countries. The authors assumed that the response indicator r_{ij} for j th individual within the i th country related to BMI value x_{ij} via

$$r_{ij}|x_{ij}, \boldsymbol{\beta} \sim^{iid} \text{Bernoulli}\{\exp(\beta_{0i} + \beta_{1i}x_{ij})/(1 + \exp(\beta_{0i} + \beta_{1i}x_{ij}))\},$$

and employed a hierarchical setting for the regression parameters above. These x_{ij} 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.

6 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 modeling due to a high number of factors. The following sections describe typical approaches used in

these contexts, which includes Generalized estimating equations (GEE), Generalized method of moments (GMM) and Generalized additive model (GAM).

6.1 Generalized estimating equations (GEE)

Remmers et al (2013) examined longitudinal relationships between PA and BMI z-scores by using GEE for the analysis. Remmers et al (2013) and Branum et al (2011) applied GEE in the analysis of BMI data by taking account of individual information being stratified by gender and baseline weight status.

GEE (Liang and Zeger, 1986) can be considered as a method for combining certain estimating equations in presence of time-dependent covariates. Given a mean model, m_{ij} subject to unknown parameter vector β , and a variance structure, V_i , the estimating equation is described as:

$$U(\beta) = \sum_{i=1}^N \frac{\partial \mu_{ij}}{\partial \beta_k} V_i^{-1} \{Y_i - m_i(\beta)\}.$$

The parameter estimates solve $U(\beta) = 0$.

The GEE algorithm has been incorporated into many major statistical software packages, including SAS, STATA, SPSS and R.

6.2 Generalized method of moments (GMM)

Lai and Small (2007) 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) (Hansen, 1982).

6.3 Generalized additive model (GAM)

Generalized additive models (GAM) combine properties of generalized linear models with additive models (Hastie, 1990). Each additive term is typically modelled as nonparametric function. Gregory (2011) used GAM to analyse how wages are affected by BMI and age. They model 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.

6.4 Variable selection

Modern variable selection methods including ridge regression, bridge regression (Frank, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and clipped absolute deviation method (Zou and Hastie, 2005) 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 settings (Hesterberg et al., 2008).

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 methods (Brdic et al., 2008; Wu and Liu, 2009).

7 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 guide the statistical analysis of BMI-related research.

Table 1: **Statistical Methods or Models**

Methods or models	Sections
Correlation Analysis or Effect Analysis	2, 4
Multiple linear regression	2, 3, 4
Random or mixed effect regression	2, 3
Quantile regression	4
Ignorable likelihood and multiple imputation methods	5
Longitudinal models and Methods	2, 6
Bayesian methods	4, 5
Moment methods and estimation equations	6
Dimension reduction methods	2, 6

Table 2: Summary of included studies

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Abarin et al. (2012)	Canada	2.2	Linear mixed-effects models	1,096; 0-14 y; M&F	Western	the duration of exclusive breastfeeding (EXBF)
Beyerlein et al. (2008)	Germany	3.2/4	GLMs, quantile regression and GAMLSS	4,967; 54-88 m; NA	Australia Germany	breastfeeding, parental social class and meal frequency (-); TV watching, maternal BMI and weight gain (Y)
Beyerlein et al. (2008)	Germany	3.2/4	linear, logistic, and quantile regression models	9,368; 54-88 m; NA	Germany	BF related to the response variable (continuous or binary) and the statistical method used
Boffetta et al. (2011)	USA	3.1	Logistic regression models	121,732; 35-85 y ; M&F	Asian	The positive association between BMI and diabetes prevalence was present in all cohorts
Bottai et al. (2014)	Sweden	4.1	Longitudinal analysis using quantile regression	17,759; 25-75 y; M	USA	Age (+), physical activity (-)
Bound (2001)		5.3				

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample ori- gin	Association
Bradic et al. (2008)	USA	6.4	LASSO			
Branum et al (2011)	USA	6.1	Generalized equations and fixed-effects models	estimating 2,758; 4 y; NA	USA	Gestational weight gain (+)
Cawley (2004)	USA	5.4	OLS, OLS with lagged weight and fixed-effects models	97,019; 14-22 y (Longitudinal for 20 years); M&F	USA	Females: wages having little ef- fect on contemporaneous weight, white (-); Males: education and intelligence test scores (+)
Chen et al. (2012)	Taiwan	4.1	Quantile regression	2,392; NA ; M&F	Taiwan	U-shaped relationship between sleep duration and BMI beyond the 90th percentile in men and an inverse U-shaped relationship at the 30th percentile of BMI in women

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Cole (1988)	UK	4.4	Box-Cox transformation and semi-parametric quantile regression	41,870; 3 w-19 y; NA	the Netherlands, USA	
Cole (1990)	UK	4.4		F	USA	
Cole and Green (1992)	UK	4.4		F	Gambian and USA	
Cole et al (1995)	UK	4.4	Reference curves and LMS method	30,535; 0-23 y; M&F	UK	It makes better sense to treat child BMI as an anthropometric measure and express it as an SD score or centile.
Cole et al (2000)	UK	4.4	Centile curves	192,727; 0-25 y; M&F	Brazil, UK, HK, the Netherlands, Singapore and USA	age 18 years passed through the widely used cut off points of 25 and 30 kg/m ² for adult overweight and obesity

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Costa-Font et al (2009)	UK	4.2	Quantile regression decomposition	415,147 Spaniards and 14,522 Italians; 18-65 y; M&F	Italy and Spain	BMI values in Spanish females are higher and more dispersed towards the upper tail (Italian counterfactuals)
Dinsdale et al (2011)	UK	1		Children and young people		
Dunson et al (2007)	USA	4.5	Bayesian density regression	522; 35-50 y; F	USA	Morbidly obese women tend to have low LH levels
Elliott and Stettin (2007)	USA	5.1	Multiple imputation model	2,474; 2-11 y; NA	USA	Transcription errors (minimal effects)
Fenske et al (2011)	Germany	4.2	Additive quantile regression, semiparametric extension	NA; NA ; M&F	India	Age (non-linear pattern), the education and the partner of the mother, the position of the child in the birth order and the general economic situation of the family

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Yang et al. (2007)	UK	2.2	linear regression	38,759; 7+ y; M&F	Europe	FTO (fat mass and obesity associated) gene that predisposes to diabetes through an effect on BMI
Gregory (2011)	USA	6.3	Generalized additive model	12,686; 20+ y; M	USA	BMI does affect wages for men, the effect of BMI isnt really changing as men age
Heagerty and Pfeiffer (1999)	USA	4.4	Semiparametric regression	3,425 ; 0-3 y; F	USA	Age-specific BMI percentile appears to be a reasonable measure of height- and age- adjusted weight in females 1-36 months of age
Hesterberg et al (2008)	alUSA	6.4	LASSO, Ridge Regression	442; NA; NA	USA	BMI has the most significant effect among several predictors for disease
Hu et al. (2000)	USA	2.1	Canonical analysis	1,512; 45-74 y ; M&F	American Indian	Women : more obese; lower apoB; higher apoAI levels

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Newby et al. (2008)	USA	2.2	Linear mixed effect model	4,280; 55-69 y; M&F	the Netherlands	Women with the highest intake of total flavonols/flavones and total catechins experienced a significantly lower increase in BMI, no significant differences in BMI change were observed in men
Lai and Small (2007)	USA	6.2	Generalized method of moments	370; 1-14 y; M&F	the Philippines	BMI is not a strong predictor of morbidity
Li et al. (2010)	USA	4.3	Nonparametric regression	20,878; 2-20 y; M&F	USA	Better estimation for BMI-for-age curves and the prevalence of high-BMI children in the U.S
Little (2011)	USA	5.2	Subsample Ignorable likelihood method	9,041; NA; NA	USA	The effect of household income and education on blood pressure are similar estimated via different analysis

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Nandram (2010)	USA	5.6	Nonignorable nonresponse Bayesian model	500,000+ ; 2-19 y; M&F	USA	Differences in the 85th or 95th percentile for overweight by county, race, and particularly age, and a small difference in sex
Newby et al. (2006)	USA	2.2	Multiple linear regression model	33,840; NA ; F	Sweden	Changes in eating patterns were significantly related to changes in BMI over 9 y and the effect was modified by baseline BMI
Ogden et al (2002)	USA	4.4	LMS	2,396; 2-20 y; NA	USA	The 2000 CDC growth charts are recommended
Onis et al (2007)	Switzerland	4.4	LMS	22,917 ; 1-24 y ; M&F	NCHS/WHO	The new curves are closely aligned with the WHO Child Growth Standards at 5 years, and the recommended adult cutoffs for overweight and obesity at 19 years

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Popkin (2010)	USA	4.2	Parametric regression	NA ; 6+ y; M&F	Australia, China, UK, and the US	the best estimates for upper extremes of the BMI distribution, large increases among women in the UK and Australia and in children of younger ages (6C9 y) in China
Razak et al. (2013)	Canada	3.1	Multilevel regression models	732,784; 20-49 y; F	37 countries	Increases in mean BMI is associated with an increase in the degree of spread (or dispersion) of BMI values, higher BMI segments gaining weight at an increased rate relative to lower BMI segments
Remmers et al. (2013)	the Netherlands	6.1	Generalized equations	470; 4-9 y; M&F	the Netherlands	Increments of MVPA were associated with decreases in BMI z-score in heavier children, both boys and girls

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample origin	Association
Renehan et al. (2008)	UK	1	Random-effects analyses and meta-regressions	meta-282,137; NA ; M&F	North America, Europe and Australia, AsiaCPacific	Increased BMI is associated with increased risk of common and less common malignancies
Stifel and Averette (2009)	USA	4.2	Quantile regression	12,000 ; 3-15 y; M&F	USA	Distinguishing correlates of child BMI at the upper and lower tails of the weight distribution are obtained
Sturm and Data (2005)	USA	4.2	Quantile regression	13,282; NA; NA	USA	prices for vegetables and fruits (+), meat prices (-), dairy or fast-food prices (x) for outlet density (x)
Tangsgorn et al. (2001)	Norway	2.1	Canonical correlation analysis	100; 21-70 y; M	NA	The complexity of simultaneous relationship of the relevant variables are successfully clarified

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample ori- gin	Association
Torloni et al. (2009)	Brazil	1	Random-effects regression	meta- 671,945; NA; F	NA	The risk of gestational diabetes mellitus is positively associated with pre-pregnancy BMI
Warrington et al. (2013)	Australia	2.2	Linear mixed model, linear mixed ef- fects model with skew-t random errors, semi- parametric linear mixed models and a non-linear mixed effects model	1,506; 1-17 y; M&F	Australia	The SPLMM was slightly better than the other methods at ac- counting for outliers and had the best model fit
Yang et al. (2011)	USA	4.1	Quantile regression	507; 40-70 y; NA	USA	The relationships between BMI and covariates were not constant across the BMI distribution and between women and men

Continued on next page

Table 2 – Continued from previous page

Authors, year	Country	Chapter	Method	Sample size, age range, gender	Sample ori- gin	Association
Yang et al. (2014)	USA	2.1	Quantile regression	10,007; NA; M&F	USA	Having sufficient and quality sleep and reducing stress may be two intervention strategies that can be promoted among adults in Philadelphia
Yu and Jones (1998)	UK	4.3	Local-linear kernels quan- tile regression	300; 6m-6y; NA	NA	All quantiles including the median display a marked peak at the larger ages
Yuan (2007)	USA	5.5	Model-based estimator	12,655; 0-6 y; NA		

Reference

- Ells LJ , Cavill N. Preventing childhood obesity through lifestyle change interventions. A briefing paper for commissioners. *National Obesity Observatory, Oxford* 2009.
- World Health Organization. *Obesity: preventing and managing the global epidemic*. World Health Organization 2000.
- Vandenbroeck IP, Goossens I, Clemens M. *Foresight Tackling Obesities: Future Choices Building the Obesity System Map..* Government Office for Science, UK Government's Foresight Programme 2007.
- Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *The Lancet* 2008; **371**: 569–578.
- Larsson SC, Wolk A. Excess body fatness: an important cause of most cancers. *The Lancet* 2008; **371**: 536–537.
- Torloni MR, Betrán AP, Horta BL et al. Prepregnancy BMI and the risk of gestational diabetes: a systematic review of the literature with meta-analysis. *Obesity Reviews* 2009; **10**: 194–203.
- Dinsdale H, Ridler C, Ells LJ. A simple guide to classifying body mass index in children. *National Obesity Observatory, Oxford* 2011.
- Tangugorn V, Krogstad O, Espeland L, Lyberg T. Obstructive sleep apnea: a canonical correlation of cephalometric and selected demographic variables in obese and nonobese patients. *The Angle Orthodontist* 2001; **71**: 23–35.
- Hu D, Hannah J, Gray RS et al. 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.
- Yang W, Müller H, Stadtmüller U. Functional singular component analysis. *J. R. Statist. Soc. B* 2011; **73**: 303–324.
- Frayling TM, Timpson NJ, Weedon MN et al. 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.

- Abarin T, Wu Y, Warrington N, Lye S, Pennell C, Briollais L. 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.
- Newby PK, Weismayer C, Åkesson A, Tucker KL, Wolk A. 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.
- Hughes LAE, Arts ICW, Ambergen T et al. Higher dietary flavone, flavonol, and catechin intakes are associated with less of an increase in BMI over time in women: a longitudinal analysis from the Netherlands Cohort Study. *The American journal of clinical nutrition* 2008; **88**: 1341–1352.
- Warrington NM, Wu Y, Pennell CE et al. Modelling BMI Trajectories in Children for Genetic Association Studies. *PloS one* 2013; **8**: e53897.
- Boffetta P, McLerran D, Chen Y et al. 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.
- Razak F, Corsi DJ, Subramanian SV. Change in the body mass index distribution for women: analysis of surveys from 37 low-and middle-income countries. *PLoS medicine* 2013; **10**: e1001367.
- Beyerlein A, Fahrmeir L, Mansmann U, Toschke, AM. Alternative regression models to assess increase in childhood BMI. *BMC medical research methodology* 2008; **8**: 59.
- Beyerlein A, Toschke, AM, von Kries R. Breastfeeding and Childhood Obesity: Shift of the Entire BMI Distribution or Only the Upper Parts&quest. *Obesity* 2008; **16**: 2730–2733.
- Thas O, Neve JD, Clement L, Ottoy JP. Probabilistic index models. *J. R. Statist. Soc. B* 2012; **74**: 623–671.
- Bottai M, Frongillo EA, Sui X et al. Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. *Obesity* 2014; **22**: E149CE156.
- Koenker R, Hallock K. Quantile regression: An introduction. *Journal of Economic Perspectives* 2001; **15**: 43–56.
- Yu K, Lu Z, Stander J. Quantile regression: applications and current research areas. *J. R. Statist. Soc. D* 2003; **15**: 331–350.

- Chen CM, Chang CK, Yeh CY. 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.
- Yang TC, Matthews SA, Chen VY. 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.
- Costa-Font J, Fabbri D, Gil J. Decomposing body mass index gaps between Mediterranean countries: A counterfactual quantile regression analysis. *Economics & Human Biology* 2009; **7**: 351–365.
- Fenske N, Kneib T, Hothorn T. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Statist. Assoc.* 2011; **106**: 494–510.
- Sturm R, Datar A. Body mass index in elementary school children, metropolitan area food prices and food outlet density. *Public health* 2005; **119**: 1059–1068.
- Stifel DC, Averett SL. Childhood overweight in the United States: A quantile regression approach. *Economics & Human Biology* 2009; **119**: 387–397.
- Popkin BM. Recent dynamics suggest selected countries catching up to US obesity. *The American journal of clinical nutrition* 2010; **119**: 284S–288S.
- Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika* 1994; **81**: 673–680.
- Li Y, Graubard BI, Korn EL. Application of nonparametric quantile regression to body mass index percentile curves from survey data. *Statistics in medicine* 2010; **29**: 558–572.
- Yu K, Jones MC. Local linear quantile regression. *J. Am. Statist. Assoc.* 1998; **93**: 228–237.
- Heagerty PJ, Pepe MS. 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.
- Cole TJ. Fitting smoothed centile curves to reference data. *J. R. Statist. Soc. A* 1988; **151**: 385–418.
- Cole TJ. The LMS method for constructing normalized growth standards. *European journal of clinical nutrition* 1990; **44**: 45–60.

- Cole TJ, Green PJ. Statistics in medicine. *European journal of clinical nutrition* 1990; **11**: 1305–1319.
- Cole TJ, Freeman JV, Preece MA. Body mass index reference curves for the UK, 1990.. *Archives of disease in childhood* 1990; **73**: 25–29.
- Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *Bmj* 2000; **320**: 1240.
- Cole TJ, Flegal KM, Nicholls D, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: international survey. *Bmj* 2007; **335**: 194.
- Ogden CL, Kuczmarski RJ, Flegal KM et al. 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.
- Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization* 2007; **85**: 660–667.
- Dunson DB, Pillai D, Park JH. Bayesian density regression. *J. R. Statist. Soc. B* 2007; **69**: 163–183.
- Elliott MR, 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.
- Little RJ, Zhang N. Subsample ignorable likelihood for regression analysis with missing data. *J. R. Statist. Soc. C* 2011; **60**: 591–605.
- Little RJ, Rubin DB. *Statistical analysis with missing data*. Wiley 2002.
- Bound J, Brown C, Mathiowetz N. Measurement error in survey data. *Handbook of econometrics* 2001; **5**: 3705–3843.
- Cawley J. The impact of obesity on wages. *Journal of Human Resources* 2004; **39**: 451–474.
- Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–592.
- Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; **81**: 471–483.
- Yuan Y, Little RJA. Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *J. R. Statist. Soc. C* 2007; **56**: 79–97.

- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* 1952; **47**: 663–685.
- Nandram B, Choi JW. A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* 2010; **105**: 120–135.
- Remmers T, Sleddens E, Gubbels J. Relationship Between Physical Activity and the Development of BMI in Children. *Medicine and science in sports and exercise* 2013; **46**: 177–184.
- Branum AM, Parker JD, Keim SA, Schempf AH. 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.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.
- Lai TL, 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.
- Hansen LP. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 1982; **50**: 1029–1054.
- Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. CRC Press 1990.
- Gregory C. Wages, BMI, and Age: A Generalized Additive Model Using the Oracle Estimator. *Available at SSRN 1975044* 2011.
- Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–135.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 1996; **58**: 267–288.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* 2005; **67**: 301–320.
- Zou H, Hastie T. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* 2005; **96**: 1348–1360.
- Hesterberg T, Choi NH, Meier L, Fraley C. Least angle and L_1 penalized regression: A review. *Statistics Surveys* 2008; **2**: 61–93.

Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Statist. Soc. B* 2008; **73**: 325–349.

Wu Y, Liu Y. Variable selection in quantile regression. *Statistica Sinica* 2009; **19**: 801.