Predicting Body Fat Using Data on the BMI

Terence C. Mills
Loughborough University

Journal of Statistics Education Volume 13, Number 2 (2005),
ww2.amstat.org/publications/jse/v13n2/datasets.mills.html

Copyright © 2005 by Terence C. Mills, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Body Mass Index; Functional form; Prediction; Regression.

Abstract

A data set contained in the Journal of Statistical Education’s data archive provides a way of exploring regression analysis at a variety of teaching levels. An appropriate functional form for the relationship between percentage body fat and the BMI is shown to be the semi-logarithmic, with variation in the BMI accounting for a little over half of the variation in body fat. The fairly modest strength of the relationship implies that confidence intervals for body fat, and tolerance intervals for BMI, can be quite wide, so that strict reliance on the BMI as a measure of body fat, and hence obesity, is unwarranted. Nevertheless, when fitting percentage body fat as a function of the class of “power weight for height indices”, i.e., indices of the form weight/height^p, the BMI, with a height exponent of p = 2, is an appropriate choice to make.

1. Introduction

In recent years both the medical profession and the media have become increasingly concerned about the levels of obesity being reached throughout the world. While percentage body fat is arguably the most accurate measure of obesity, it is quite difficult to calculate: one of the more accurate methods requires underwater weighing. Attention has thus focused on the easier to calculate body mass index (BMI), defined as the ratio of weight (in kilograms) to the square of height (in metres); see, for example, Kuczmarski and Flegal (2000) for a survey of the historical development of the BMI. An interesting statistical problem is then to assess the relationship between body fat and the BMI, which can include gauging the strength of the correlation, testing hypotheses concerning, for example, the functional form of the relationship, and using the fitted model for prediction. This paper utilises a relatively small, easily accessed data set containing body fat and BMI measurements to illustrate these facets of regression analysis. Thus it should be both accessible and of interest to students at various levels who wish to employ regression techniques in their data analysis.
2. The Data Set

Johnson (1996) provides both a data set containing various physical measurements on 252 men, including two estimates of percent body fat, and an explanation of the way these estimates are calculated. This data set is available at the Journal of Statistical Education data archive, ww2.amstat.org/publications/jse/jse_data_archive.html. We use here the body fat percentages calculated using the Brozek equation. Johnson (1996) indicates some probable errors in this data set but, as we wish to model body fat as a function of just BMI, and later of the BMI’s constituents, weight and height, we only concern ourselves here with those cases which had errors or aberrant values in their (Brozek) percent body fat, weight and/or height. For case 182, a particularly lean individual for which the body fat percentage predicted by Brozek’s equation is negative, this percentage has been set at 0.1 prior to analysis. For case 42, we follow Johnson (1996, Section 3) and replace the reported height of 29.5 inches with the much more reasonable 69.5 inches.

A scatterplot of all 252 pairs of observations is shown in Figure 1. A traditional starting point for students analysing the relationship between percent body fat and the BMI is to fit a linear regression. Two such regressions are shown superimposed on Figure 1: a fit to all 252 cases and a fit with case 39 omitted from the calculations. This case has a BMI of 48.9 associated with a body fat percentage of 33.8 and is seen to be an outlier, both pulling the fitted line towards it and giving an impression of a distinct curvilinear relationship between the two variables. It will therefore be omitted from further modelling, but students could be encouraged to discuss whether this is the best course of action for dealing with an aberrant observation and whether alternative solutions could be considered.

![Figure 1. Scatterplot of body fat percentage against BMI with linear regressions fitted to all observations (bodyfat% = -20.4 + 1.55BMI)](image-url)
and with the outlier removed (bodyfat% = –24.9 + 1.73BMI) superimposed.

3. Nonlinear Modelling

There appears to be little evidence from Figure 1 of a nonlinear relationship between percent body fat and the BMI, which may be confirmed by asking students to plot the regression residuals against the BMI. Earlier studies of this relationship (Webster, Hesp and Garrow, 1984; Gray and Fujioka, 1991), however, have suggested that a nonlinear relationship of the “inverse’ form

$$\text{bodyfat}\% = a \left(\frac{\text{BMI} - b}{\text{BMI}}\right) = a - ab \frac{1}{\text{BMI}}$$

may be more appropriate on theoretical grounds. Here $b$ represents the theoretical BMI associated with 0% body fat, and $a$ represents the percentage of excess body weight which is fat (see Gray and Fujioka, 1991, pages 548–9, for more detailed interpretation of this relationship). With $a$ and $b$ both positive, the first and second derivatives of this function with respect to BMI are positive and negative respectively. Thus the model implies that percent body fat increases, but at a decreasing rate, with increasing BMI and is bounded by the value $a$. An alternative model which would also capture this type of nonlinearity is the semi-logarithmic relationship bodyfat% = $c + d\ln(\text{BMI})$ with $d$ positive. Students can fit these two nonlinear functions to the data and assess their fits relative to that of the linear model. Figure 2 presents the fits of the three functions graphically. Students can note that the estimates of the inverse function parameters are $a = 64.3 \pm 2.6\%$ and $b = 17.6 \pm 0.3\%$, where we use the notation $\hat{\beta} = \hat{\beta} \pm \text{s.e.}(\hat{\beta})$, where $\hat{\beta}$ and $\text{s.e.}(\hat{\beta})$ are the ordinary least squares (OLS) estimate and associated standard error of a parameter $\beta$, i.e., that a BMI of 17.6 is associated with 0% body fat and that 64.3% of excess body weight is fat. It should be noted that, while estimates of the parameters of this model can be obtained by regressing percent body fat on the inverse of the BMI and algebraically calculating the estimates from the regression coefficients, parameter standard errors can only be obtained using a dedicated nonlinear regression routine. The $R^2$ statistics from the three models are: linear 0.560, inverse 0.557, and semi-log 0.563, with residual error variances $5.123^2$, $5.141^2$ and $5.105^2$, respectively. Thus, on goodness of fit grounds, the semi-logarithmic model is to be preferred. The fits, however, are very similar over the central range of BMI values, only differing substantially for the very highest BMI values, so that a more formal method of model selection could be considered.
One approach is to note that all three functional forms may be “nested” within a general functional form by using the Box and Cox (1964) family of power transformations, which are defined for the generic variable $Z$ as

$$Z^{(e)} = \begin{cases} \left( \frac{Z^\theta - 1}{\theta} \right) & \theta \neq 0, \quad Z > 0 \\ \ln Z, & \theta = 0 \end{cases}$$

If we denote the $i$th observations, $i = 1, 2, ..., n$, on percent body fat and the BMI as $y_i$ and $x_i$, respectively, then we may consider the nonlinear regression model

Figure 2. Fitted linear, semi-logarithmic (bodyfat% = $-126.2 + 45.0 \ln(BMI)$), and inverse (bodyfat% = $64.3(BMI - 17.6/BMI)$) functions.
where $u_i$ is an error term, assumed to be normally and independently distributed with zero mean and constant variance $\sigma^2$. The linear model is obtained when $\lambda = \phi$, setting $\lambda = 1$ and $\phi = 0$ defines the semi-logarithmic model, while setting $\lambda = 1$ and $\phi = -1$ defines the inverse model (in each case with a redefined intercept of $1 + \beta_0 - \phi \beta_1$). However, arbitrary values of the power transformation parameters need not be imposed upon (1); rather, they may be estimated along with the other parameters of the model, and tests of the hypotheses implied by the alternative functional forms may then be performed in order to discriminate between them.

The procedure for doing this is to recognise that, for any given values of $\lambda$ and $\phi$, estimates of $\beta_0$ and $\beta_1$ conditional upon these values may be obtained from the regression of equation (1). Maximum likelihood (ML) estimates of the power transformation parameters, denoted $\hat{\lambda}$ and $\hat{\phi}$, and hence of the $\beta$’s, are found by maximizing the concentrated log-likelihood, defined as

$$L(\lambda, \phi) = \frac{n}{2} \ln \hat{\sigma}^2(\lambda, \phi) + (\lambda - 1) \sum \ln y_i$$

where

$$\hat{\sigma}^2(\lambda, \phi) = \frac{1}{n} \sum \hat{u}_i$$

is an estimate of the “conditional” error variance, the $\hat{u}_i$ being the residuals from the conditional regression (1). The term “concentrated” is used because the maximization is, in fact, a step-wise procedure in that $\hat{\sigma}^2(\lambda, \phi)$ is first obtained from a linear regression with fixed values of $\lambda$ and $\phi$, with the second step being to maximize $L(\lambda, \phi)$ over all values of $\hat{\sigma}^2(\lambda, \phi)$. For an extended discussion of concentrated likelihood methods, see Seber and Wild (2003, pages 37-42), and for a review of the Box-Cox transformation, see Sakia (1992).

This maximization may conveniently be computed by searching over a grid of $\lambda$, $\phi$ values: advanced students can be encouraged to develop routines for carrying out this two-dimensional grid search, which involves calculating the power transformations, saving regression output to compute the log-likelihood, and writing looping procedures. Students may also experiment with plotting the contours of the likelihood function so constructed, which will provide a graphical perspective on the likely precision with which the transformation parameters, $\lambda$ and $\phi$, are estimated. This precision may also be examined by calculating the confidence region obtained by using the result that

$$S(\lambda, \phi) = 2 \left[ L(\hat{\lambda}, \hat{\phi}) - L(\lambda, \phi) \right]$$

is approximately distributed as chi-squared with two degrees of freedom (Box and Cox, 1964). Thus, for
example, 95% and 75% confidence regions are defined by \( S(\lambda, \phi) < 5.99 \) and \( S(\lambda, \phi) \) respectively.

The ML estimates are obtained as \( \hat{\lambda} = 0.92 \) and \( \hat{\phi} = 0.01 \), with \( L(0.92, 0.01) = -407.83 \). Since \( L(1, 0) = -407.17 \), it is clear that the semi-logarithmic model is contained within any conventional confidence region. The other functional forms are quite "close" in terms of fit, however. The linear model \( (\lambda = \phi = 1 \text{ in equation (1)}) \) has \( L(1, 1) = -409.04 \), and so is contained within the 75% confidence region, while the inverse model \( (\lambda = 1, \phi = -1) \) has \( L(1, -1) = -409.95 \) and is thus contained within a 95% confidence region. The double-logarithmic model \( (\lambda = \phi = 0 \text{ in equation (1)} \) and a functional form that is often used in regression analysis), however, has \( L(0, 0) = -501.59 \) and so is excluded from all conventional confidence regions.

The fitted regressions for the linear, semi-logarithmic and inverse functional forms are

\[
\hat{y}_i = -24.9 + 1.73 x_i \quad (2.5) \quad (0.10)
\]

\[
\hat{y}_i = -126.2 + 45.0 \ln x_i \quad (8.1) \quad (2.5)
\]

and

\[
\hat{y}_i = 64.3 - \frac{1133}{64} x_i \quad (2.6) \quad (64)
\]

where parameter standard errors are shown in parentheses. Diagnostic checks on the residuals of each of these regressions found no evidence of heteroskedasticity or non-normality in any of the regressions, as was also true for the residuals from the regression using the ML estimates.

At this point students could be asked to consider alternative non-linear specifications. A plausible competitor would be a polynomial in \( x_i \), so that students may fit the quadratic regression:

\[
\hat{y}_i = -42.5 + 3.08 x_i - 0.025 x_i^2 \quad (13.1) \quad (1.00) \quad (0.019)
\]

Figure 3 shows the implied semi-logarithmic, quadratic and linear functions for \( 10 \leq \text{BMI} \leq 50 \), bearing in mind that the range of BMI values in the sample used for estimation is 18.1 to 39.1. The three functions are almost identical over the central region of the BMI range (20 to 30), but the linear model is a poor approximation to the semi-logarithmic outside of this interval. The quadratic, on the other hand, provides a good approximation to the semi-logarithmic over the entire observed range of BMI values, even though, as students can check, the quadratic term in (7) is insignificant (its t-ratio is just –1.36). Since the semi-logarithmic function also produces a superior fit to the quadratic (the error variance of the latter is 5.114^2) and contains one less parameter, we prefer the former as a better representation of the relationship between percent body fat and BMI.
4. Using the models for prediction

The fitted equations may then be used for prediction. We concentrate on the two best-fitting equations, the linear (4) and semi-logarithmic (5). Focusing our exposition on the linear model, students can construct predictions and, say, one-standard error bounds, by using the standard formulae to predict a value of \(y\), \(\hat{y}_f\) say, given the BMI value \(x\) (see, for example, Maddala, 1977, chapter 7, or Brown, 1993, chapter 2):

\[
\hat{y}_f = \hat{\beta}_0 + \hat{\beta}_1 x_f \\
\sigma(\hat{y}_f) = \sigma \left( 1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum_x} \right)^{\frac{1}{2}} 
\]

where
The one-standard error bounds for $y$ are then calculated as $\hat{y}_f \pm \sigma(\hat{y}_f)$. Students may be encouraged to provide interpretations of these bounds. For example, the current U.S. Dietary guidelines define the range $18.5 < \text{BMI} < 25$ to be “healthy,” $25 < \text{BMI} < 30$ to be “overweight,” and higher values of BMI to be “obese” (see Kuczmarski and Flegal, 2000, table 2). Using the linear model, these cut-offs predict body fat percentages of $\hat{y} = 7.1$, 18.3 and 27.0 for $x = 18.5$, 25 and 30, respectively. For the data set here, we have $\bar{x} = 25.3$ and $\sum x = 2787.8$. The one-standard error bounds (68% prediction intervals) are then calculated as (1.9, 12.3), (13.2, 23.4) and (21.8, 32.2). These bounds, along with the predicted values, are plotted in Figure 4 and show that, at this level of confidence, a BMI value can only predict body fat within a range of approximately 10 percentage points with 68% accuracy (i.e., if a sequence of forecasts were made given $x$, the true value $y$ would only be contained in 68% of the calculated prediction intervals). This reflects the modest strength of the fitted relationship given by the $R^2$ value of just 0.56. Students can then be asked to repeat the calculations with the semi-logarithmic model and comment on any differences. Here the 68% prediction intervals are (−0.1, 10.3), (13.5, 23.7) and (21.8, 32.0), respectively, so that the prediction intervals are shifted downwards, relative to the linear model, for BMI values further from the sample mean, reflecting the curvature of the semi-logarithmic function.
A related question to ask students is: if we are given a value $y_f$, what is the value $x$ that could have given rise to $y_f$? An answer to this may be obtained by inverting (8) to give

$$
\hat{x}_f = \frac{y_f - \hat{\beta}_0}{\hat{\beta}_1}
$$

Although this is the ML estimate of $x_f$, it is biased because, in general,

$$
E(\hat{x}_f) = \frac{E(y_f - \hat{\beta}_0)}{\hat{\beta}_1} - x_f
$$

(see, for example, Seber and Lee, 2003, section 6.1.5). An alternative estimate is that obtained from estimating the reverse, or inverse, regression of $x$ on $y$:

$$
\hat{x}_f = \hat{\beta}_0 + \hat{\beta}_1 y_f
$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLS estimates of the inverse regression. Predicting from $x$ is often referred to as calibration. The inverse regression estimator $\hat{x}_f$ is typically thought to be appropriate when the $x$ observations are regarded as a random sample from a population: this is referred to as the “natural” or “random” calibration problem. The estimator $\hat{x}_f$ is appropriate when the $x$ values may be taken as fixed by the design of the experiment: this is “controlled” calibration. Osborne (1991) presents a historical review of statistical calibration, which has given rise to confusion and argument for many years, while Brown (1993, chapter 2.3) is a useful reference. Students could be asked to discuss which of the two calibration problems the data set most appropriately falls under.

Error bounds (prediction intervals) for $x$ using the estimate $\hat{x}_f$ are straightforwardly calculated by using the formulae in (8) with $x$ and $y$ interchanged. A more challenging question is the construction of a prediction interval for $x$ using $\hat{x}_f$. The approach developed in Seber and Lee (2003, chapter 6.1.5), for example, considers the error made in predicting $y_f$, $y_f - \hat{y}_f$, which can be written as

$$
y_f - \hat{y}_f - y_f - \hat{\beta}_0 - \hat{\beta}_1 \hat{x}_f - y_f - \bar{y} - \hat{\beta}_1 (x_f - \bar{x})
$$

the last equality being obtained using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. For large $n$, and defining $\vartheta_f = x_f - \bar{x}$, the ratio
\[ \frac{\left(y_f - \bar{y} - \hat{\beta}_1 \hat{\sigma}_f\right)}{\left(1 + \frac{\hat{\sigma}_f^2}{\sum_x}\right)^{\frac{1}{2}}} \]

will be distributed as standard normal. Thus

\[ \Gamma\left[\left(\frac{y_f - \bar{y} - \hat{\beta}_1 \hat{\sigma}_f}{\hat{\sigma}_f \left(1 + \frac{\hat{\sigma}_f^2}{\sum_x}\right)^{\frac{1}{2}}}\right)^2 \leq \chi^2_{\alpha} \right] = 1 - \alpha \]

where \( \chi^2_{\alpha} \) is the \( \alpha \) percentage point of the chi-square distribution with one degree of freedom. The set of all values of \( \hat{\sigma}_f \) satisfying the inequality

\[ \left(y_f - \bar{y} - \hat{\beta}_1 \hat{\sigma}_f\right)^2 \leq \chi^2_{\alpha} \hat{\sigma}_f^2 \left(1 + \frac{\hat{\sigma}_f^2}{\sum_x}\right) \]

will then provide a \( (1 - \alpha) \) confidence interval for the unknown \( x_f \), with lower and upper bounds defined as \( \bar{x} + \hat{\sigma}_1 \) and \( \bar{x} + \hat{\sigma}_2 \). \( \hat{\sigma}_1 \) and \( \hat{\sigma}_2 \) are the solutions (roots) of the quadratic equation

\[ \left(y_f - \bar{y} - \hat{\beta}_1 \hat{\sigma}_f\right)^2 - \chi^2_{\alpha} \hat{\sigma}_f^2 \left(1 + \frac{\hat{\sigma}_f^2}{\sum_x}\right) = 0 \]

i.e.,

\[ \left(\hat{\beta}_1^2 - \frac{\chi^2_{\alpha} \hat{\sigma}_f^2}{\sum_x}\right) \hat{\sigma}_f^2 - 2 \hat{\beta}_1 (y_f - \bar{y}) \hat{\sigma}_f + \left(y_f - \bar{y}\right)^2 - \chi^2_{\alpha} \hat{\sigma}_f^2 \left(1 + \frac{1}{n}\right) = 0 \]  \( (9) \)

It is possible for these roots to be complex if \( \hat{\beta} \) is not significantly different from zero, in which case the regression line is close to being horizontal and any value of the regressor is acceptable. As can be seen from the fitted regression (4), however, \( \hat{\beta}_1 \) is highly significant as the 95% confidence interval for \( \hat{\beta}_1 \) is 1.73 \( \pm \) 0.20. The resultant interval is often called a discrimination interval rather than a prediction interval (Seber and Wild, 2003, page 146).

Johnson (1996) reports a suggestion that 15% body fat is a maximum for good health for men, so it is interesting to calculate BMI prediction and discrimination intervals for this value of \( y_f \). The estimated coefficients of the inverse regression are \( \hat{\beta}_0 = 19.22 \) and \( \hat{\beta}_1 = 0.32 \), so that
while the value of $x$ is

$$x_f = (15 + 24.9)/1.73 = 23.1$$

For the calculation of the intervals we require $\bar{y}$ and $\sum_y$, so that $\sigma(\bar{x}_f) = 2.225$. A 95% prediction interval using $\bar{x}_f = 24.1$ is thus (19.7, 28.5). Using $\bar{x}_f = 23.1$, for a 95% discrimination interval, with $\chi^2_{0.05} = 3.84$, the quadratic (9) simplifies to

$$2.96\beta^2 + 13.15\beta - 86.72 = 0$$

The two roots of this equation are $\beta_1 = -8.1$ and $\beta_2 = 3.6$. The 95% discrimination limits for $x$ are thus 17.2 and 28.9, which show that the interval is asymmetric about $\bar{x}_f$. Thus, if $\bar{x}_f$ is used, then at the 95% level of confidence, a 15% body fat is consistent with a BMI ranging from 17.2 to 28.9, i.e., from a BMI below the current lower healthy BMI cut-off to a value close to the upper-end of the overweight range. If $\bar{x}_f$ is used, this range is a little narrower, running from 19.7 to 28.5. A similar calculation with the semi-logarithmic function obtains 95% limits of 20.2 and 28.3 (using $\bar{x}_f$) and 18.3 and 28.9 (using $\bar{x}_f$). Students may be encouraged to discuss the implications of the width of these intervals for the efficiency of the BMI as an indicator of percent body fat.

5. Disentangling the influence of weight and height on body fat

Given the definition of the BMI as the ratio of weight (in kilograms) to the square of height (in metres), the regressor in the best-fitting semi-logarithmic model (5) can be decomposed as

$$\ln x_i = \ln w_i - 2 \ln h_i$$

where $w$ and $h$ are the weight and height of the $i$th case. Thus, if the multiple regression

$$y_i = \beta_0 + \beta_1 \ln w_i + \beta_2 \ln h_i + \epsilon_i$$

is considered, equation (5) is obtained if the restriction $2\hat{\beta}_1 + \hat{\beta}_2 = 0$ is imposed. Students may then be asked to consider how they might test such a linear restriction. One approach would be construct a confidence interval for $2\hat{\beta}_1 + \hat{\beta}_2$: a 100(1 - $\alpha$) interval is given by

$$2\hat{\beta}_1 + \hat{\beta}_2 \pm z_{\alpha/2} \left(\text{var}(2\hat{\beta}_1 + \hat{\beta}_2)\right)^{1/2}$$
where $z_{\alpha/2}$ is the $\alpha/2$ percentage point of the standard normal distribution and

$$\text{var}(2\hat{\delta}_1 + \hat{\delta}_2) = 4 \text{var}(\hat{\delta}_1) + \text{var}(\hat{\delta}_2) + 4 \text{cov}(\hat{\delta}_1, \hat{\delta}_2)$$

(see, for example, Maddala, 1977, chapter 10.3). The Johnson (1996) data set contains data on weight and height, although measured in pounds and inches. After rescaling to kilograms and metres (multiplying by the factors 0.4536 and 0.0254 respectively), the following multiple regression was obtained

$$y_i = -118.7 + 44.3 \ln w_i - 101.3 \ln h_i$$

$$\begin{pmatrix} 9.4 \\ 2.5 \\ 10.2 \end{pmatrix}$$

The variances and covariances of the estimated coefficients are estimated as $\text{var}(\hat{\delta}_1) = 6.24$, $\text{var}(\hat{\delta}_2) = 103.63$ and $\text{cov}(\hat{\delta}_1, \hat{\delta}_2) = -13.23$, respectively, so that $\text{var}(2\hat{\delta}_1 + \hat{\delta}_2) = 75.67$. Thus an approximate 95% confidence interval (using $z_{0.025} = 1.96$) for $2\hat{\delta}_1 + \hat{\delta}_2$ is calculated to be $-11.7 \pm 17.0$, which includes 0 and thus provides evidence in favour of using the BMI as a “composite” weight-height index. Note, however, that an approximate 68% interval (using $z_{0.159} = 1$) is $-11.7 \pm 8.7$, which does not include 0.

Earlier research on “weight for height” indices by Benn (1971) considered power indices of the form $\text{weight}/\text{height}^p$ (see also Flegal, 1990). The testing approach outlined above may be utilised to construct a range of values for the exponent $p$ that are consistent with the data. The power index implies the linear combination $p\hat{\delta}_1 + \hat{\delta}_2$ between the slope coefficients of equation (10), which can then be written as

$$y_i = \delta_0 + \delta_1 (\ln w_i - p \ln h_i) + u_i$$

The exponent can be estimated as $\hat{p} = -\frac{\hat{\delta}_2}{\hat{\delta}_1}$. The variance of $\hat{p}$ is given by

$$\text{var} \hat{p} \approx \left[ \frac{\partial \delta_2}{\partial \delta_1} \right] \left[ \begin{array}{ccc} \text{var}(\hat{\delta}_1) & \text{cov}(\hat{\delta}_1, \hat{\delta}_2) \\ \text{cov}(\hat{\delta}_1, \hat{\delta}_2) & \text{var}(\hat{\delta}_2) \end{array} \right] \left[ \begin{array}{c} \frac{\partial \delta_1}{\partial \delta_1} \\ \frac{\partial \delta_2}{\partial \delta_2} \end{array} \right]$$

Since
Using the estimates $\hat{\beta}_1 = 44.8$ and $\hat{\beta}_2 = -101.3$, we can calculate $\hat{\beta} = 2.26$ and $\text{var}(\hat{\beta}) \approx 0.038$. Thus an approximate 95% confidence interval for $\beta$ is $2.26 \pm 1.96(0.038)^{1/2} = 2.26 \pm 0.38$. This interval contains $\beta = 2$, thus again confirming the choice of the BMI as an appropriate “Benn power-index”, but excludes $\beta = 1.5$, a value of the exponent suggested by early studies of the weight-height relationship but not currently recommended (see Kuczmarski and Flegal, 2000).

6. Conclusions

In this paper we have shown that, for the Johnson (1996) data set, an appropriate functional form of the relationship between percentage body fat and the BMI is semi-logarithmic, with variation in the BMI accounting for a little over half of the variation in body fat. The fairly modest strength of the relationship implies that confidence intervals for body fat, and tolerance intervals for BMI, can be quite wide, so that strict reliance on the BMI as a measure of body fat, and hence obesity, is unwarranted. Nevertheless, within the class of “power weight for height indices”, the BMI, with a height exponent of 2, is an appropriate index to use.

Acknowledgement

I would like to thank three referees and the editor, Roger Johnson, for very helpful comments on earlier drafts that have significantly improved the paper.

References

Benn, R.J. (1971), “Some mathematical properties of weight for height indices as measures of adiposity,” British Journal of Preventive and Social Medicine, 25, 42-50.

Box, G.E.P. and Cox, D.R. (1964), “An analysis of transformations,” Journal of the Royal Statistical Society, Series B, 26, 211-243.

Brown, P.J. (1993), Measurement, Regression, and Calibration, Oxford, U.K.: Oxford University Press.

Flegal, K.M., (1990), “Ratio of actual to predicted weight as an alternative to a power-type weight-height index (Benn index),” American Journal of Clinical Nutrition, 51, 540-547.

Gray, D.S. and Fujioka, K. (1991), “Use of relative weight and body mass index for the determination of adiposity,” Journal of Clinical Epidemiology, 44, 545-550.
Johnson, R.W. (1996), “Fitting percentage of body fat to simple body measurements,” *Journal of Statistics Education* [Online], 4(1). [www.amstat.org/publications/jse/v4n1/datasets.johnson.html](http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html)

Kuczmarski, R.J. and Flegal, K.M. (2000), “Criteria for definition of overweight in transition: background and recommendations for the United States,” *American Journal of Clinical Nutrition*, 72, 1074-1081.

Maddala, G.S. (1977), *Econometrics*, McGraw-Hill.

Osborne, C. (1991), “Statistical calibration: a review,” *International Statistical Review*, 59, 309-336.

Sakia, R.M. (1992), “The Box-Cox transformation technique,” *The Statistician*, 41, 169-178.

Seber, G.A.F. and Lee, A.J. (2003), *Linear Regression Analysis*, 2nd Edition, New York: John Wiley & Sons.

Seber, G.A.F. and Wild, C.J. (2003), *Nonlinear Regression*, New York: John Wiley & Sons.

Webster, J.D., Hesp, R. and Garrow, J.S. (1984), “The composition of excess weight in obese women estimated by body density, total body water and total body potassium,” *Human Nutrition: Clinical Nutrition*, 38C, 299-306.

---

Terence C. Mills  
Department of Economics  
Loughborough University  
Leicestershire  
United Kingdom  
*T.C.Mills@lboro.ac.uk*