bmjmedicine

Non-linear predictor outcome associations

Frederick K Ho ,[1] Tim J Cole[2]

[1]School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[2]UCL Great Ormond Street Institute of Child Health, London, UK
Correspondence to: Dr Frederick K Ho, University of Glasgow, Glasgow G12 8RZ, UK; Frederick.Ho@glasgow.ac.uk

Cite this as: *BMJMED* 2023;2:e000396. doi:10.1136/bmjmed-2022-000396

Received: 4 October 2022
Accepted: 6 January 2023

Ignoring non-linear association can lead to erroneous inference. Frederick Ho and Tim Cole consider methods to overcome this problem in practice.

Introduction

Statistical models, such as linear regression, help to uncover the association between a predictor (eg, a prognostic factor or a hypothesised cause) and a health outcome (eg, health related quality of life). To make the models estimable, various assumptions have to be made. An implicit assumption of linear regression when the predictor is continuous (eg, body mass index (BMI)) is that the outcome is linearly associated with the predictor. In other words, a unit change in the predictor is associated with a constant change in the outcome whatever the value of the predictor. The purple solid line in figure 1 shows such a fitted straight line portraying the association between BMI and mental wellbeing in 12 435 participants from the Understanding Society study.[1] The model indicates that each extra unit of BMI was associated with 0.03 units worse mental wellbeing, across the range of BMI. The difference between individuals with BMIs of 19 and 20 would be the same as individuals with BMIs of 39 and 40.

However, the association between predictor and outcome is often not linear. For example, average BMI across all children follows a J-shaped pattern with age,[2] and dietary protein intake is associated with lower mortality linearly only when the intake is low.[3] Therefore, an assumption of linearity, without first testing for it, is inadvisable.

Non-linear associations can be modelled by categorising the predictor. The simplest case is dichotomisation, such as splitting BMI into obese (≥30) and non-obese (<30) groups. Apart from the loss of statistical power, categorisation is poor at portraying the true association.[4] All values in a category are assumed to be associated with the outcome in the same way. In the obesity example, the difference between people who were underweight and overweight was lost when they were grouped together (yellow dashed line compared with the pink dashed line in figure 1). Categorisation with more groups captures the non-linearity better, but the result is still imprecise and the cut-points are arbitrary (pink dashed line in figure 1).

KEY MESSAGES

⇒ Non-linear predictor outcome association occurs when the association between the predictor and outcome varies depending on different values of the predictor

⇒ Ignoring non-linear association could lead to erroneous inference; categorising the predictor could be a crude approximation for the non-linearity but is subject to important limitations

⇒ Non-linear associations are common and the models used to capture them properly should be considered more widely

Modelling with explicit polynomials

Polynomial regression models specify the non-linear association using a polynomial of the predictor. The purple solid line in figure 2 shows the association between mental wellbeing and BMI as a cubic curve. Polynomials with a degree greater than 3 are not generally used as they are too sensitive to outliers.[5] Correlation between polynomial terms can also make the estimation less robust.[6] Because only two additional terms (quadratic and cubic) are used, cubic polynomial regression is not that accurate for capturing complex non-linear associations.[5] Fractional polynomials,[7] that is, where the power can be a fraction (eg, $x^{1/2}$), can also used, where the power of each term in the regression formula, and the number of terms, are selected by a data-driven algorithm. Fractional polynomials can capture complex associations better than cubic polynomials. Nonetheless, all polynomial regression (like linear regression) estimates its coefficients using all the data, which means that one outlier at an extreme value can markedly affect the curve.

Modelling with regression splines

A regression spline consists of a set of piecewise polynomials, each fitted to a different section of the predictor range.[8] The locations where the polynomials meet are called knots. Regression analysis with a categorical predictor can be thought as an extreme form of spline; each piecewise polynomial is a horizontal straight line and the knots are the category cut-points. Regression splines most commonly use cubic polynomials, hence cubic splines, which have to meet certain conditions, such as continuity and smoothness at the knots.

A strength of cubic spline regression is that each data point influences only one of the fitted cubic curves. A limitation of splines is that the boundary polynomials (ie, those at the extremes of the predictor) are unstable and particularly sensitive to outliers. To stabilise them, the boundary curves can be modelled as straight lines; such splines are called natural cubic splines or restricted cubic splines.

Natural cubic splines can capture complex non-linear associations and are relatively robust to outliers but are sensitive to the number and location of the knots. Two examples of natural cubic splines are shown in figure 2. The yellow dashed line has three knots and the pink dashed line has 14 knots, placed at the quantiles of the distribution. Because regression splines are sensitive to knot choice, analysts

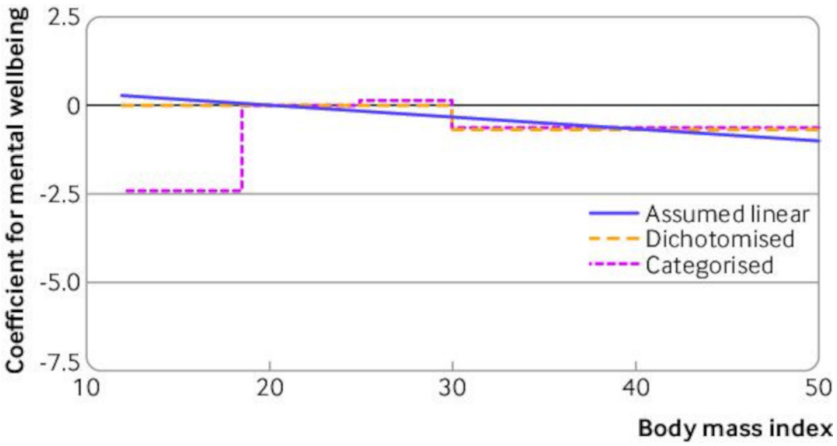

Figure 1 | Association between body mass index and mental wellbeing using conventional methods. Mental wellbeing is norm referenced with a population mean of 50 and a standard deviation of 10, adjusted for age, sex, ethnicity, education, and baseline physical and mental wellbeing. Age and wellbeing variables were adjusted as P-splines. Primary sampling units and strata were adjusted using random intercepts

need to optimise their number and placement. Too many knots lead to overfitting, whereby the curve is too rough (pink line in figure 2). Conversely, too few knots lead to underfitting, in which the spline is too smooth and does not capture the underlying curvature. Fitting a straight line to a non-linear association is an obvious example of underfitting, also known as oversmoothing. In many applications, including three to five knots (depending on sample size and curve complexity) at the corresponding quantiles of the predictor should result in a reasonable estimate.[9]

A data driven approach for knots

Penalised splines are an extension of regression splines that minimise the influence of knot choice. They do this by use of a data driven algorithm to penalise a rough, overfitted curve, and can achieve appropriate smoothness with minimal input from the analyst.

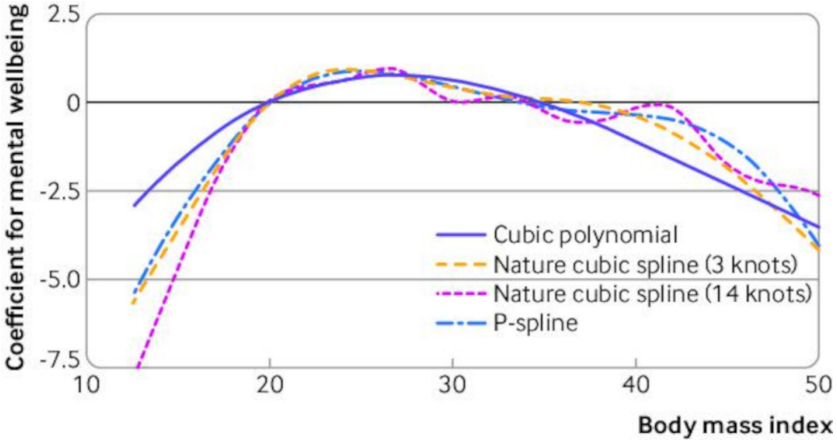

Figure 2 | Association between body mass index and mental wellbeing using non-linear methods. Mental wellbeing is norm referenced with a population mean of 50 and standard deviation of 10. Adjusted for age, sex, ethnicity, education, and baseline physical and mental wellbeing. Age and wellbeing variables were adjusted as P-splines. Primary sampling units and strata were adjusted using random intercepts.

P-splines[10] are a simple yet powerful example of penalised splines.[11] P-splines include a relatively large number of knots (often 20 or more), equally spaced on the predictor range, and the large number of knots avoids underfitting. Meanwhile, a computationally light smoothing penalty avoids overfitting. The blue dashed line in figure 2 shows a P-spline fitted with default settings.

Other examples are available of penalised splines based on different outlooks, most notably the smoothing spline[12] and the thin plate spline.[13] The smoothing spline assigns a knot at each of the data points and its smoothing penalty is more complex than that of the P-spline. Theoretically, this method should result in a better fit at the expense of a longer processing time. The thin plate spline can be considered as a generalisation of the smoothing spline with the ability to fit multiple predictors simultaneously.[13] Thereby, this spline can estimate the non-linear association of outcome with two (or more) predictors as a smooth surface.

Modelling non-linearity in practice

Tools to model non-linear associations are readily available. All statistical software that can fit regression models should also be able to incorporate polynomial terms. Regression splines are available in R (splines package), Stata (mkspline), and SPSS (regression procedure). Penalised splines can be implemented by several packages in R (gam, mgcv, gamlss). The generic term for them is generalised additive models, an extension of generalised linear models where the additive indicates non-linear fitting. These techniques can also be applied to other forms of regression, for example, logistic regression for binary data, and proportional hazards regression (Cox model) for time-to-event data. However, the non-linear link function for these regressions can affect the interpretation of linearity.

Statistical inference on splines can be done by conducting overall significance tests. An F-test can compare model fit for a spline with a straight line and test for non-linearity. Similarly, the Bayesian Information Criterion can compare splines (particularly regression splines) with different knot choices.[14] Additionally, visual evidence can show under-fitting or over-fitting. The aim is to adjust the number of knots to show curve features that are consistent with subject knowledge.

Although all these methods involve different mathematics, they produce similar estimates in common applications.[15] Polynomial regression fit is often slightly worse and more influenced by outliers, so this method is usually used when a simple explicit mathematical form is needed, or when splines are not available. The fitted curves from P-splines, smoothing splines, and thin plate splines are often similar.[15]

Conclusions

Non-linear associations are common and models to properly capture them are readily available. Their use should be considered more widely.

Twitter Frederick K Ho @fredho42

Acknowledgements This study is completed under UK Data Service project 227777. We are grateful to the Understanding Society participants.

Contributors FH conceptualised the study, analysed the data and drafted the manuscript. TJC commented on the study conceptualisation, evaluated the data, and critically revised the manuscript.

Funding We do not declare a specific grant for this research from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests We have read and understood the BMJ policy on declaration of interests and declare the following interests: none.

Ethics approval This study involves human participants and was approved by Understanding Society data was approved by a blanket ethical approval. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. The data can be requested from the UK Data Service (SN 8715). Codes can be downloaded at: https://github.com/fredkho42/nonlinearity.

ORCID iD

Frederick K Ho http://orcid.org/0000-0001-7190-9025

REFERENCES

1 University of Essex Institute for Social and Economic Research. *Understanding Society: longitudinal teaching dataset, waves 1-9, 2009-2018. UK data service. Sn: 8715*, 2021.
2 Cole TJ, Lobstein T. Extended International (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatr Obes* 2012;7:284–94. doi:10.1111/j.2047-6310.2012.00064.x
3 Ho FK, Gray SR, Welsh P, *et al*. Associations of fat and carbohydrate intake with cardiovascular disease and mortality: prospective cohort study of UK Biobank participants. *BMJ* 2020;368:m688. doi:10.1136/bmj.m688
4 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.1. doi:10.1136/bmj.332.7549.1080
5 Stasinopoulos MD, Rigby RA, Heller GZ, *et al*. *Flexible regression and smoothing: using GAMLSS in R*. Boca Raton, Florida: CRC Press, 2017.
6 Bradley RA, Srivastava SS. Correlation in polynomial regression. *Am Stat* 1979;33:11–4.
7 Royston P, Altman DG. Approximating statistical functions by using fractional polynomial regression. *J R Stat Soc: Ser D* 1997;46:411–22.
8 Marsh LC, Cormier DR. *Spline regression models*. Sage, 2001.
9 Harrell FE. *Regression modeling strategies*. NY: Springer, 2015.
10 Eilers PH, Marx BD. *Practical smoothing: the joys of P-splines*. Cambridge University Press, 2021.
11 Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996;11:89–121. doi:10.1214/ss/1038425655
12 Wang Y. *Smoothing splines: methods and applications*. CRC press, 2011.
13 Wood SN. Thin plate regression splines. *J R Stat Soc Ser B* 2003;65:95–114. doi:10.1111/1467-9868.00374
14 Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–4. doi:10.1214/aos/1176344136
15 Perperoglou A, Sauerbrei W, Abrahamowicz M, *et al*. A review of spline function procedures in R. *BMC Med Res Methodol* 2019;19:1–16. doi:10.1186/s12874-019-0666-3