Open Journal of Statistics, 2016, 6, 76-84
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.61010

Local Curvature and Centering Effects in Nonlinear Regression Models

Michael Brimacombe

Department of Biostatistics, KUMC, Kansas City, KS, USA
Email: mbrimacombe@kumc.edu

Received 15 December 2015; accepted 20 February 2016; published 23 February 2016

Abstract

The effects of centering response and explanatory variables as a way of simplifying fitted linear models in the presence of correlation are reviewed and extended to include nonlinear models, common in many biological and economic applications. In a nonlinear model, the use of a local approximation can modify the effect of centering. Even in the presence of uncorrelated explanatory variables, centering may affect linear approximations and related test statistics. An approach to assessing this effect in relation to intrinsic curvature is developed and applied. Mis-specification bias of linear versus nonlinear models also reflects this centering effect.

Keywords

Nonlinear Regression, Centering Data, Model Mis-Specification, Bias, Curvature

1. Introduction

Applied probability models are mathematical constructs that have roots in both theory and observed data. They often reflect specific theoretical properties, but may simply be the application of an all-purpose linear model. The fitting of a probability model to the observed data requires careful consideration of potential difficulties and model sensitivities. These may include aspects of the model itself or anomalies in the structure of the database. As large scale observational databases have become more common, the possibility of unplanned and non-standard data patterns have become more common.

The stability of linear models can be affected by various properties of the model-data combination. Model sensitivity to rescaling and transformations of the response [1], the presence and effect of heterogeneity [2], the need to employ ridge regression when collinearity is present [3], all have the goal of improving the application and stability of the model-data combination and resulting fitted model. In the application of linear models, these issues extend to consideration of residual error behavior and diagnostic measures to detect the effects of outliers,

How to cite this paper: Brimacombe, M. (2016) Local Curvature and Centering Effects in Nonlinear Regression Models. *Open Journal of Statistics*, **6**, 76-84. http://dx.doi.org/10.4236/ojs.2016.61010

collinearities and serial correlation. Discussion of these can be found in [4].

The simple centering of data in linear models is often applied as a component of standardizing the variables in a regression, re-centering the means of the variables at zero. It can also be seen as a way to lower correlation among explanatory variables in some cases, but will have limited if any effect on ANOVA related test statistics and measures of goodness of fit in models when interaction terms are present in the model. This is due to the geometry of the test statistics involved which typically reflect standardized lengths of orthogonal projections which are invariant to centering. See for example [5]. In high dimensional linear models, centering allows for easier geometric interpretation of correlations among a set of centered vectors and is often an initial step in the analysis. Note that in data with nonlinear patterns, correlation based adjustments often does not make sense as they implicitly assume an underlying linear framework. A serious concern in this regard is model mis-specification, here the assumption of a linear model when underlying nonlinearity is present. Centering the data may induce bias and inaccurate estimation and testing.

Nonlinear regression models are also available to model data based patterns. The use of centering in such models can be challenging to interpret. Such models are common in many biological, ecological and economic applications and there is often less flexibility in the set of potential modifications available as theory often informs and restricts model choice. Examples can be found in [6]. In terms of inference, the Wald statistic tends to be more interpretable, even though the log-likelihood ratio and score function are more theoretically justified. The local curvature of the regression surface may require consideration if approximations based on local linear models are used to develop pivotal quantities for inference, especially in small samples with normal error.

In this paper, centering effects are examined in relation to the use of linear approximation in nonlinear regression models. To begin, the effects of centering in linear models with interaction effects are reviewed. Centering effects in nonlinear models where linear approximation is employed to obtain tests of significance are then discussed. Even in the presence of uncorrelated explanatory variables and simple main effects, centering may significantly affect locally defined linear approximations and related test statistics. Local measures of nonlinearity are defined and used to assess these effects. We then investigate the mis-specification of linear versus nonlinear models and show that centering effects arise as a measure of bias. This is particulary relevant in high dimensional data modeling where centeriing is common as a first step in data analysis.

2. Centering in Linear Models

We can write a standard linear model in the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

typically assuming the ε_i random errors are *i.i.d.* $N\left(0,\sigma^2\right)$. The y_i are the responses of interest, the β_j unknown parameters and the x_{ij} are explanatory variables taken here as known. The y_i and x_{ij} can be collected into vectors and matrices and re-expressed as $y = X\beta + \varepsilon$. The model is quite flexible and can be transformed in many ways.

The use of centering in linear regression settings is typically suggested to lower correlation among the explanatory variables. For example, if x_i^2 is entered in the model already containing x_i, centering will often lower the correlation between them. This will provide more stability in the interpretation of the fitted model. Centering is often thought to be useful when interaction terms are entered into the model, giving more stability in least squares based estimation. The cross-product term in regression models with interaction may be collinear with the main effects, making it difficult to detect identify both main and interaction effects. However in such models, as shown in [5], mean-centering does not change the computational precision of parameters, the sampling accuracy of main effects, interaction effects, nor the R^2. The pivotal quantities and related test statistics for the main effects may require adjustment for this to be clear as the respective parameters may alter meaning.

To see this, consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Centering by definition will not affect the shape of the initial (x, y) data cloud, it simply re-centers it to $(0,0)$. The best fitting line will therefore not alter in terms of its slope and neither will the residuals of the fitted line. As the *SSE* is the squared length of the residuals, the *MSE* the average squared length and the goodness of

fit measure R^2 $R^2 = 1 - SSE/SST$, where $SST = \left(y_i - \bar{y}\right)^2$, these also do not alter with centering. The *OLS* estimate for the slope, $\hat{\beta}_1$, is based on sums of differences from the *x* and *y* means and is invariant to centering, as is the correlation between *x* and *y*. The error distribution assumed does not affect these results. It is based on the initial assumption of normally distributed (theoretical) erros and the geometric properties of the least squares estimators. Note that the estimate for the intercept β_0 will alter upon centering the data.

For the multivariate linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

the same basic argument related to residuals holds and the results are similar. The centering of all variables has no effect on the measures of association between the *x* and *y* variables, including the least squares estimators $\hat{\beta}_j$, $j = 1, \cdots, p$. Note again that if terms of the form $\beta_j x_j^2$ are added to the model, then centering may lower the correlation between the x_j and x_j^2 terms.

The addition of interaction terms $x_i x_j$ to the linear model are a way of examining whether the relationship between *y* and x_i can be interpreted directly without accounting for the levels of another variable x_j. If the coefficient for the respective interaction term is found to be significant, the main effect relating *y* and x_i cannot be directly assessed and stratification of the model may be necessary. Typically the multiple $x_i \cdot x_j$ is taken to represent interaction effects as the partial derivative of the response with regard to either of the *x* will have the form

$$\frac{\partial y}{\partial x_i} = \beta_i + \beta_j x_j$$

This implies that the main effect of x_i is dependent on the level of x_j. Note that the transformation $y \to \log(y)$ may remove a significant interaction.

The centering of the data to limit potentially high levels of correlation between the interaction term $x_i x_j$, and both x_i and x_j is sometimes suggested. As noted above this will not alter most measures of fit in the linear model (even a linear model where one of the variables is the interaction term). In particular, as shown in [5], if we have as our model

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta x_{1i} x_{2i} + \varepsilon_{ij}$$

then the least squares estimate of the interaction term will not alter if x_1 and x_2 are centered, neither will the R^2 value for the model. Note that the significance for the main effects in this model will appear to alter, but only due to the parameters having a different meaning in the centered model and thus related t-tests are testing slightly different hypotheses.

3. Example 1

Consider the Penrose bodyfat ([7]) dataset of physiologic measurements where some measures are highly correlated. We look to predict bodyfat density as a function of several body measurements; Abdomen, Wrist, Weight, Hip, Knee, Ankle, Forearm, Biceps, Thigh, Chest. Three principal components account for 84% of the total variation in the data. Stepwise regression gives three variables (Abdomen, Weight, Wrist) accounting for an R^2 value of 73%. These variables have high correlations (0.88, 0.73, 0.62) which do not alter if we center the data. If we proceed to include interactions, dropping the Abdomen-Weight interaction due to extreme collinearity, we obtain a similar R^2 value (73.1%). The correlations among the interactions themselves can be examined pre-centering (0.95, 0.96, 0.94) and post-centering (0.38, 0.90, 0.30) showing the effect of centering. We also obtain an overall F-test value of 133.95 (significant at 0.0001) which does not alter and $SSE = 0.02$, also invariant to centering. Further results are given in **Table 1**. Note that the *OLS* estimates for the interactions terms and their standard errors do not alter.

4. Nonlinear Regression Models: Local Curvature Assessment

Nonlinear regression models typically are developed and applied in areas such as toxicology, economics and ecology. See [8]. Consider the nonlinear regression model

$$y_i = \eta\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) + \varepsilon_i \quad (1)$$

Table 1. Centering in linear models. (a) Original Data (S = 0.0099, R-sq = 73.1); (b) Centered Data (S = 0.0099 R-sq = 73.1).

(a)

Variables	Wt	Ab	Wrist	Ab * Wt	Ab * Wrist
Coeff (Std Error)	−0.0005 (0.00041)	−0.0026 (0.00041)	−0.0017 (0.0056)	0.000002 (0.000002)	0.00003 (0.00003)
p-value	0	0.001	0.77	0.29	0.35

(b)

Variables	Wt	Ab	Wrist	Ab * Wt	Ab * Wrist
Coeff (Std Error)	0.0002 (0.00005)	−0.0022 (0.00013)	0.0037 (0.001)	0.000002 (0.000002)	0.00003 (0.00003)
p-value	0	0.001	0.001	0.29	0.35

$i=1,\cdots,n$ where $\boldsymbol{x}_i$ are fixed values of the explanatory variable $\boldsymbol{x}$, the model function η is known and depends on the parameter vector $\boldsymbol{\beta}\in\mathbf{R}^p$ and $\boldsymbol{x}_i$. The ε_i are independent error terms, each normally distributed with mean zero and variance element σ^2. The set of possible mean values defines a surface, $\boldsymbol{\eta}(\boldsymbol{\beta}):\boldsymbol{\beta}\in\Omega\subseteq\mathbf{R}^p$, where Ω is the parameter space and $\boldsymbol{\eta}(\boldsymbol{\beta})$ is the $n\times 1$ column vector with i^{th} component given by $\eta(\boldsymbol{x}_i,\boldsymbol{\beta})$. Some standard examples of nonlinear models include the Michaelis-Menten model $y_i=\beta_1 x_i/(\beta_2+x_i)+\varepsilon_i$ and the Logistic model; $y_i=\beta_1/(1+\beta_2 \mathrm{e}^{\beta_3 x_i})+\varepsilon_i$.

Nonlinear regression models are subject to the effects of centering when using local linear approximation. The relative position of the response y vis-a-vis the solution locus $\eta(\boldsymbol{x}_i,\boldsymbol{\beta})$ and the point on the surface at which the linear or tangent plane approximation is developed will affect the degree to which centering affects least squares based analysis of the model. In relation to the residual vector, an important aspect of the linear argument above, when there is intrinsic curvature present, the usual geometric properties of the residual vector are affected as they are the projection of an idempotent matrix only locally. Below we show that simply centering the data affects the observed residuals, affects the level of a locally defined measure of intrinsic curvature and thus the linear approximation based analysis, and in the setting of misclassification, imputes bias into the analysis even to the first order.

Local Geometry

Some geometry is briefly reviewed. Let F_0 be the $n\times p$ matrix with column elements given by $\boldsymbol{f}_i=\partial\boldsymbol{\eta}(\boldsymbol{\beta})/\partial\beta_i\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o}$ for $i=1,\cdots,p$. If $\boldsymbol{L}(F_0)$ is the tangent plane to the surface $\boldsymbol{N}$ defined at $\boldsymbol{\eta}(\boldsymbol{\beta}_0)$, then $P_0=F_0(F_0'F_0)^{-1}F_0'$ is the orthogonal projection matrix for $\boldsymbol{L}(F_0)$ evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. Further $P_0(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\beta}_0))$ is the projection of $(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\beta}_0))$ onto the tangent plane at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. Let $\boldsymbol{u}=P_0(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\beta}_0))/\|P_0(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\beta}_0))\|$, where $\|\cdot\|$ denotes length, be a unit vector centered at $\boldsymbol{\eta}(\boldsymbol{\beta}_0)$ on the tangent plane. The quadratic approximation to $\boldsymbol{\eta}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ is given by

$\boldsymbol{\eta}(\boldsymbol{\beta})-\boldsymbol{\eta}(\boldsymbol{\beta}_0)=F_0\theta+\left(\frac{1}{2}\right)\theta' H_0\theta$ where H_0 is the Hessian $p\times p$ matrix with vector elements

$\boldsymbol{h}^{ij}=\dfrac{\partial^2\eta(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j}$ evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}=(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$. The intrinsic acceleration vector in the direction $\boldsymbol{u}$ can be expressed as $(I-\boldsymbol{u}\boldsymbol{u}')\boldsymbol{\eta}''(\boldsymbol{\beta})$ or $-\boldsymbol{v}/\rho$, where $\boldsymbol{v}$ is the unit vector perpendicular to the acceleration vector in the direction $\boldsymbol{u}$ and $\rho=\rho(\boldsymbol{\beta}_0)$ is the corresponding radius of curvature at $\boldsymbol{\eta}(\boldsymbol{\beta}_0)$. We then have

$$-\boldsymbol{v}/\rho=\frac{\boldsymbol{\theta}'G_0\theta}{\|F_0\theta\|^2}$$

where $G_0=\left[(I-P_0)h_{ij}\right]$. Taking the norm gives the intrinsic local curvature

$$\kappa=1/\rho=\frac{\|\theta'G_0\theta\|}{\|F_0\theta\|^2} \tag{2}$$

where again $\boldsymbol{\theta} = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, all matrices are evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

An intrinsic curvature based adjustment to standard ANOVA can be developed. See [9]. The usual orthogonal decomposition of regression and error can be replaced with the orthogonal decomposition $\boldsymbol{y} = z_1\boldsymbol{u} + z_2\boldsymbol{v} + \boldsymbol{V}\boldsymbol{z}_3$ with the residual space spanned by the intrinsic curvature vector $\boldsymbol{v}$ and the column vectors of $\boldsymbol{V}$, which are orthonormal vectors spanning the remaining residual space dimensions, orthogonal to both tangent plane and $\boldsymbol{v}$, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The relevance of the curvature in the direction $\boldsymbol{u}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ can be assessed by comparing the orthogonal projection(s) of $(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0))$ onto $\boldsymbol{u}$ and $\boldsymbol{v}$ respectively.

To investigate this curvature effect in relation to the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ an approximate linear model based approach can be used. A sum of squares regression component can generate a global *F*-test with *p* and $(n-p)$ degrees of freedom. Assuming $\boldsymbol{y} \sim N\left((\boldsymbol{\eta}(\boldsymbol{\beta}_0)), \sigma^2 I\right)$ where σ^2 is unknown, we have under the null;

$$\frac{\left\| P_0\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 / p}{\left\|(I - P_0)\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 / (n-p)} \sim F_{p,n-p}. \tag{3}$$

with large values of the test statistic leading to rejection of $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

A further orthogonal decomposition gives a test of significance for curvature in the direction $\boldsymbol{u}$ using orthogonal projection onto the vector $\boldsymbol{v}$;

$$\frac{\left\| P_0\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 / p}{\sqrt{\left(\left\|(I - P_0)\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 - \left\|P_v\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2\right) / (n - (p+1))}} \sim F_{1,n-(p+1)} \tag{4}$$

where $P_v\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right) = vv'$. A large value here reflects a significant projection length onto the curvature vector $\boldsymbol{v}$ in the direction $\boldsymbol{u}$. The orthogonal projection onto the vector $\boldsymbol{v}$ also provides a correction factor for the global test

$$\frac{\left(\left\|P_0\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 + \left\|P_v\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2\right) / (p+1)}{\left(\left\|(I - P_0)\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2 - \left\|P_v\left(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta}_0)\right)\right\|^2\right) / (n - (p+1))} \sim F_{p+1,n-(p+1)}.$$

See [10] for further details and application in regard to the testing of global null hypotheses. As the effect of intrinsic curvature depends where on the actual regression surface the linear approximation is developed in relation to the position of the response vector *y*, all of these test statistics may reflect centering effects.

5. Centering in Nonlinear Models

As in linear models, the use of centering on both response and some if not all of the explanatory variables initially would seem to have little or no effect on the underlying geometry of the model-data combination. A graph of the (x, y) point cloud initially centered at $(\bar{x}, \bar{y})$ will simply re-center at $(0, 0)$ even if the overall pattern is nonlinear. However there may be effects on the subsequent analysis due to the nature of the nonlinear model and the locally linear frame of reference used for inference. The relative centering based shift in the $\boldsymbol{\eta}(\boldsymbol{\beta})$ surface versus the shift in the response *y* may alter the geometric relationship between *y* and $\boldsymbol{\eta}(\boldsymbol{\beta})$ and the tangent plane relevant to the local approximation, related test statistics and orthogonal projections. These effects do not exist in the standard linear model setting as projections are taken onto the same flat surface with zero curvature at all points. Here the more curved the regression surface, the more the local frame of reference can be affected by small changes in the relative positioning of the response vector.

In regard to standard *m.l.e.* based analysis, the effects of centering will depend on the actual model itself. For example consider the asymptotic growth model

$$y_i = \beta_1\left(1 - \exp\left(-\beta_2 x_i\right)\right) + \varepsilon_i.$$

where centering the data yields

$$(y_i - \bar{y}) = \beta_1\left(1 - \exp\left(-\beta_2\left(x_i - \bar{x}\right)\right)\right) + \varepsilon_i$$

If the differences $(y_i - \bar{y})$ are relatively greater than $(x_i - \bar{x})$ then in terms of the response vector and regression surface the portion of the regression surface relevant to supporting the local linear approximation and analysis will alter. Note also that the parameters and their estimators in a nonlinear model are not easily interpreted as simple intercept and slope. They are often defined and justified in terms of underlying differential equations or asymptotic properties.

The fundamental nature of a nonlinear regression model may be reflected in its possible forms under reparameterisation, especially in regard to re-expression as a linear model. If this is possible, then intrinsic curvature corrections tend to be of little value and centering can be seen to have the same non-effect as in standard linear models with regard to the rescaled parameters. For example, the Michaelis-Menten model is given by;

$$y_i = \frac{\theta_1 x_i}{(\theta_2 + x_i)} + \varepsilon_i$$

where ε_i are *i.i.d.* $N(0, \sigma^2)$. This can be re-expressed and re-parameterized as

$$y_i = \frac{\theta_1 x_i}{(\theta_2 + x_i)} = \frac{x_i}{((\theta_2/\theta_1) + (x_i/\theta_1))}$$

$$1/y_i = \theta_1 + (\theta_2/\theta_1)(1/x_i)$$

Letting $y_i^* = 1/y_i$ and $x_i^* = 1/x_i$ the model has a linear form if this reformatting of the variables is acceptable. In some settings however this re-writing of the model may not be possible.

For models which may not be re-expressed as linear models, we can assess the change in curvature effect at a given $\boldsymbol{\eta}(\boldsymbol{\beta}_0; x)$ when centering the data $(y, x) \to (y^*, x^*)$ using

$$\left\| P_v\left(\boldsymbol{y}^* - \boldsymbol{\eta}\left(\boldsymbol{x}_i^*, \boldsymbol{\beta}_0\right)\right)\right\|^2 \Big/ \left\| P_v\left(\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\right)\right\|^2.$$

The *SSE* values may also differ and together these alter the relevant *F*-statistics for the local ANOVA analysis discussed above. Note that while the raw data plot is simply re-centered, the local approximation and analysis reflecting the model-data combination is more strongly affected by centering.

6. Example 2

We examine these concepts further in the context of the asymptotic growth model applied to the BOD dataset found in Bates and Watts (1988). This is given by

$$y_i = \beta_1\left(1 - \exp\left(-\beta_2 x_i\right)\right) + \varepsilon_i.$$

The original and centered dataset is given in **Table 2** and results from fitting the model based on the *m.l.e.* are given in **Table 3**.

The non-standard behavior of this model yields log-likelihood based confidence regions that are open at confidence levels above 95% in the β_2 direction and a linear approximation based analysis can be applied. The first order derivative matrix is *n* by 2 and can be written, for $i = 1, \cdots, n$

$$F = \left[\left(1 - \mathrm{e}^{-\beta_2 x_i}\right), \beta_1 x_i \mathrm{e}^{-\beta_2 x_i}\right]$$

with related 2 by 2 by *n* second order Hessian matrix

$$H = \begin{bmatrix} 0 & x_i \mathrm{e}^{-\beta_2 x_i} \\ x_i \mathrm{e}^{-\beta_2 x_i} & -\beta_1 x_i^2 \mathrm{e}^{-\beta_2 x_i} \end{bmatrix}$$

where each h_{ij} is an *n*-dimensional vector. The 0 value denotes a linear aspect to the model in certain directions, sometimes called partially linear.

Note that the *m.l.e.* here is not available in closed form, rather it is defined by differentiating the log-likelihood with regard to each parameter and setting the resulting equations equal to zero. Here the log-likelihood is given by

Table 2. BOD Data (Centered).

Demand	8.3 (−6.53)	10.3 (−4.53)	19 (4.17)	16 (1.17)	15.6 (0.77)	19.8 (4.97)
Time	1 (−2.667)	2 (−1.667)	3 (−0.667)	4 (0.333)	5 (1.333)	7 (3.333)

Table 3. BOD Model Standard Output ($H_0 : \beta_1 = 0, \beta_2 = 0$).

	MLE ($\hat{\beta}_1, \beta_2$)	Std Error ($\hat{\beta}_1, \hat{\beta}_2$)	t-statistic ($\hat{\beta}_1, \hat{\beta}_2$)	p-value ($\hat{\beta}_1, \hat{\beta}_2$)	SSE
Original Data	19.14, 0.53	2.50, 0.20	7.67, 2.62	0.0015, 0.06	2.549
Centered Data	8.10, 0.21	11.94, 0.27	0.68, 0.78	0.53, 0.48	2.878

$$\sum\left[\left(-1/2\sigma^2\right)\left(y_i-\beta_1\left(1-\exp\left(-\beta_2 x_i\right)\right)\right)^2\right]$$

Note that the effects of centering on the *m.l.e.* occur in this set of equations. Standard errors can be determined from the inverse of the Fisher Information matrix.

For the original data, the resulting maximum likelihood or least squares value for (β_1, β_2) is given by $\hat{\beta}_1 = 19.143(\pm 2.5)$, $\hat{\beta}_2 = 0.5311(\pm 0.2)$ with residual standard error $s^2 = 2.549$ on 4 degrees of freedom. The residual vector is given by (0.41, −2.22, 3.75, −0.85, −2.20, 1.12). T-tests for for a difference from zero give p-values of 0.0015 and 0.059 respectively. For the centered data, the maximum likelihood values for (β_1, β_2) are $\hat{\beta}_1 = 8.1055(\pm 11.94)$ and $\hat{\beta}_2 = 0.2147(\pm 0.27)$ with residual standard error $s^2 = 2.878$ on 4 degrees of freedom. The residual vector is given by (−0.29, 1.05, 5.41, 0.61, −1.24, 0.83). Comparing the maximum likelihood values is difficult as the meaning of the parameters alters. More importantly we can see that the residual vector and related *SSE* have altered due to centering.

The curvature adjusted approach using ANOVA is given in **Table 4** for a null value of $\boldsymbol{\beta}_0 = (18.0, 0.4)$.

The measure

$$\left\|P_v\left(\boldsymbol{y}^* - \boldsymbol{\eta}\left(\boldsymbol{x}_i^*, \boldsymbol{\beta}_0\right)\right)\right\|^2 \Big/ \left\|P_v\left(\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{x}_i, \boldsymbol{\beta}_0\right)\right)\right\|^2.$$

is examined here by comparing the *SSCurv* elements pre and post centering. This has a value pre-centering (0.40) that is approximately only 10% of its value post-centering (3.90). Whether this incurs statistically significant effects will depend on the local curvature of the surface, the manner in which the parameters enter into the model and the relative position of y in relation to $\boldsymbol{\eta}(x;\boldsymbol{\beta})$ and its linear approximation before and after centering. The results in **Table 4** show the centering of the data affecting the formal significance of the global test.

7. Mis-Specification and Centering Related Bias

The use of linear models when the underlying model-data combination is nonlinear can lead to mis-specification error. It is interesting to consider this in relation to centering effect which can yield bias even where second order intrinsic curvature is not significant. In many high dimensional data analytic techniques the centering of the data is a standard first step. See for example [10]. However it is rare in those settings that linearlity can be confidently assumed.

To examine mis-specification generally in this setting, we begin by expressing a linear model as function of two sets of variables

$$y = X\theta + \varepsilon = X_1\theta_1 + X_2\theta_2 + \varepsilon$$

Assume that the variables of interest form the X_1 $(n \times p_1)$ matrix with p_1 variables and the X_2 $(n \times p_2)$ matrix has p_2 additional variables and $p_1 + p_2 = p$. The error distribution is given by $\varepsilon \sim N\left(0, \sigma^2 I\right)$. The goal here is to identify significant variables in the X_1 matrix.

Assume now that a true nonlinear model underlies the set of X_1 variables. Re-expressing our initial model we have

$$y = W\left(X_1, \theta_1\right) + X_2\theta_2 + \varepsilon$$

Table 4. (a) ANOVA Table for BOD Model and Data $\left(H_0:\boldsymbol{\beta}_0=(18.0,0.4)\right)$; (b) ANOVA Table for Centered BOD Model and Data $\left(H_0:\boldsymbol{\beta}_0=(18.0,0.4)\right)$.

(a)

Source	df	SS	MS	F-statistic	p-value
Regression	2	31.93	15.97	2.44	0.21
Residual	4	26.11	6.53		
Curvature	1	0.4	0.4	0.047	0.85
Modified Residual	3	25.71	8.57		
Regression + Curvature	3	32.33	10.78	1.26	0.43
Total	6	58.03			

(b)

Source	df	SS	MS	F-statistic	p-value
Regression	2	1103.5	551.53	65.66	0.001
Residual	4	33.58	8.4		
Curvature	1	3.91	3.91	0.4	0.572
Modified Residual	3	29.67	9.89		
Regression + Curvature	3	1106.96	368.99	37.31	0.007
Total	6	1136.63			

where $W\left(X_1,\theta_1\right)$ is a nonlinear model for the X_1 subset of variables. Replacing $W\left(X_1,\theta_1\right)$ with its Taylor expansion about θ_{10} gives

$$y=\left[W\left(X_1,\theta_{10}\right)+W'\left(X_1,\theta_{10}\right)\left(\theta_1-\theta_{10}\right)\right]+X_2\theta_2+\varepsilon \tag{5}$$

where $W\left(X_1,\theta_{10}\right)$ is a constant function of X_1 and W' the relevant derivative. We can further write

$$y=X_1^*\theta_1+X_2\theta_2+\varepsilon^*$$

where $\varepsilon^*=\varepsilon-W'\left(X_1,\theta_{10}\right)\left(\theta_{10}\right)+W\left(X_1,\theta_{10}\right)$ and $X_1^*=W'\left(X_1,\theta_{10}\right)$.

If we fit the original linear model, mis-specification effects arise as we will use (i) X_1 instead of X_1^* and (ii) apply a biased error distribution as the more appropriate error distribution with nonlinearity present is; $\varepsilon^*\sim N\left(W'\left(X_1\theta_{10}\right)\left(\theta_1-\theta_{10}\right)+W\left(X_1,\theta_{10}\right),\sigma^2I\right)$. This reflects a type of centering effect that will be incorporated into the approximate least squares based analysis to follow. Typically we evaluate this at $\theta_1=\hat{\theta}_1$.

If the actual data are also centered, it follows that a data-based centering effect will further occur. Letting $\overline{x}_1$ be the centering element we have as the resulting error distribution

$$\varepsilon^*\sim N\left(W'\left(\left(X_1-\overline{x}_1\right),\theta_{10}\right)\left(\theta_{10}\right)\right)+\left(\left(W\left(X_1-\overline{x}_1\right),\theta_{10}\right),\sigma^2I\right).$$

The effect of centering the data here may be to worsen the mis-specification related biasing effect. This will depend on how the linear and nonlinear elements in the *W* vector and W' matrix interact with the centered data $\left(X_1-\overline{x}_1\right)$. Note that if the Taylor expansion is to the second order, then intrinsic curvature also affects the usefulness of residuals. See [11]. Here we have shown that in a nonlinear model with the possibility of linear versus nonlinear mis-specification, bias results from simple first order issues and the centering of data.

8. Discussion

Model sensitivity and stability are essential components of applied research using probability modes. These are functions of the model structure, data structure and the inferential or estimation method used to fit the model.

This is most pronounced when nonlinear models are to be employed and linear approximation is a component of the inferential process. Wald statistics are the most interpretable in this setting and in the case of nonlinear regression with normal error; the curvature of the regression surface is a key component affecting the accuracy of the inferential process. The underlying nature of the model is also relevant with linearity on same scale being reflected in the intrinsic curvature related calculations. These issues arise often in the analysis of high dimensional datasets where centering is a standard first step.

If we examine centering in the context of the original point cloud the effects of centering seem non-existent. But the information in the data is assessed in relation to the assumed linear or nonlinear model. The properties of the assumed model are thus relevant to the estimation and testing of parameters defined within the fitted local model. The positioning of the response vector y in n-space in relation to the p-dimensional nonlinear regression surface defines a local frame of reference for inference with the intrinsic curvature and even simple centering has effects in nonlinear models both generally and when linear approximation is employed. Nonlinear models often reflect theoretical results for carefully chosen parameter and data scaling. In conclusion, the centering of data in relation to nonlinear regression model should be applied and interpreted carefully.

Acknowledgements

We thank the Editor and the referee for their comments.

References

[1] Box, G.E.P. and Cox, D.R. (1964) An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211-252.

[2] Box, G.E.P. (1953) Non-Normality and Tests on Variances. *Biometrika*, **40**, 318-335. http://dx.doi.org/10.2307/2333350

[3] Draper, N.R. and Smith, H. (1981) Applied Regression Analysis. 2nd Edition, John Wiley & Sons, New York.

[4] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.

[5] Echambadi, R. and Hess, J.D. (2007) Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models. *Journal of Marketing Science*, **26**, 438-445. http://dx.doi.org/10.1287/mksc.1060.0263

[6] Seber, G.A.F. and Wild, C.J. (1989) Nonlinear Regression. John Wiley, New York. http://dx.doi.org/10.1002/0471725315

[7] Penrose, K., Nelson, A. and Fisher, A. (1985) Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, **17**, 189. http://dx.doi.org/10.1249/00005768-198504000-00037

[8] Bates, D.M. and Watts, D.G. (1988) Nonlinear Regression Analysis and Its Applications. John Wiley, New York. http://dx.doi.org/10.1002/9780470316757

[9] Brimacombe, M. (2016) A Note on Linear and Second Order Significance Testing in Nonlinear Models. *International Journal of Statistics and Probability*, **5**, 19-27.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407-451. http://dx.doi.org/10.1214/009053604000000067

[11] Cook and Tsai (1985) Residuals in Nonlinear Regression. *Biometrika*, **72**, 23-29. http://dx.doi.org/10.1093/biomet/72.1.23