Revista Colombiana de Estadística
July 2018, Volume 41, Issue 2, pp. 137 to 155
DOI: http://dx.doi.org/10.15446/rce.v41n2.68535

Using an Anchor to Improve Linear Predictions with Application to Predicting Disease Progression

Usando un anclaje para mejorar predicciones lineales con aplicación a la predicción de progresión de enfermedad

Alex G. Karanevich[a], Jianghua He[b], Byron J. Gajewski[c]

Department of Biostatistics, University of Kansas Medical Center, Kansas City, United States

Abstract

Linear models are some of the most straightforward and commonly used modelling approaches. Consider modelling approximately monotonic response data arising from a time-related process. If one has knowledge as to when the process began or ended, then one may be able to leverage additional assumed data to reduce prediction error. This assumed data, referred to as the "anchor", is treated as an additional data-point generated at either the beginning or end of the process. The response value of the anchor is equal to an intelligently selected value of the response (such as the upper bound, lower bound, or 99^{th} percentile of the response, as appropriate). The anchor reduces the variance of prediction at the cost of a possible increase in prediction bias, resulting in a potentially reduced overall mean-square prediction error. This can be extremely effective when few individual data-points are available, allowing one to make linear predictions using as little as a single observed data-point. We develop the mathematics showing the conditions under which an anchor can improve predictions, and also demonstrate using this approach to reduce prediction error when modelling the disease progression of patients with amyotrophic lateral sclerosis.

***Key words*:** Anchor; Amyotrophic lateral sclerosis; Biased regression; Linear models; Ordinary least squares.

Resumen

Modelos lineales son los modelos más fáciles de usar y comunes en modelamiento. Si se considera el modelamiento de una respuesta aprosimadamente monótona que surge de un proceso relacionado al tiempo y se sabe

[a] PhD. E-mail: akaranevich@kumc.edu
[b] PhD. E-mail: jhe@kumc.edu
[c] PhD. E-mail: bgajewski@kumc.edu

cuándo el proceso inició o terminó, es posible asumir datos adicionales como palanca para reducir el error de predicción. Estos datos adicionales son llamados de "anclaje" y son datos generados antes del inicion o después del final del proceso. El valor de respuesta del anclaje es igual a un valor de respuesta escogido de manera inteligente (como por ejemplo la cota superior, iferior o el percentil 99, según conveniencia). Este anclaje reduce la varianza de la predicción a costo de un posible sesgo en la misma, lo cual resulta en una reducción potencial del error medio de predicción. Lo anterior puede ser extremadamente efectivo cuando haypocos datos individuales, permitiendo hacer predicciones con muy pocos datos. En este trabajo presentamos en desarrollo matemático demostrando las condiciones bajo las cuales el anclaje puede mejorar predicciones y también demostramos una reducción del error de predicción aplicando el método a la modelación de progresión de enfermedad en pacientes con esclerosis lateral amiotrófica.

***Palabras clave*:** Anclaje; esclerosis lateral amiotrófica; modelos lineales; mínimos cuadrados ordinarios; regresión sesgada.

1. Introduction

Prediction has always been an important part of statistical modeling. With the advent of big data and the rise of machine learning, one may think that researchers have moved beyond prediction via simple linear models. This is not the case, however, especially in the field of medical research: a quick search of PubMed results in over 1000 publications which utilize linear (but not generalized linear) models from January 2016 – July 2017. This is because linear models are usually one of the first attempted approaches when analyzing new data, and they are sufficient surprisingly often. Linear models are simple to calculate, requiring tiny amounts of computing power compared to some of the more complex machine-learning algorithms (such as neural networks). Most importantly, linear models are very straightforward to interpret and explain, a direct contrast to the more sophisticated "black-box" methods that are dependent on large datasets. The ability to interpret and understand statistical models, or model intelligibility, is especially important in the field of healthcare (Caruana, Lou, Gehrke, Koch, Sturm & Elhadad 2015).

Yet linear models have their failings, especially when modelling a bounded response. Consider attempting to model the disease progression over time of a patient with amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease. This is measured by the instrument known as the ALS Functional Rating Scale – Revised, or ALSFRS-R (Cedarbaum, Stambler, Malta, Fuller, Hilt, Thurmond & Nakanishi 1999). The ALSFRS-R is always an integer between 0 and 48, with 48 representing no spread of the disease and 0 being the theoretical maximal spread of the disease. The progression of the ALSFRS-R tends to be very linear (Armon, Graves, Moses, Forté, Sepulveda, Darby & Smith 2000, Magnus, Beck, Giess, Puls, Naumann & Toyka 2002), but because of its bounded nature, simple linear models have the inherent structural defect of creating predictions that violate these lower and upper bounds. Many adjustments to this problem exist: examples include truncating the prediction to 48 if the prediction is too large (0 if too small) (Amemiya 1973) or performing a logistic transform on the data (Lesaffre, Rizopoulos & Tsonaka 2007).

If the goal is prediction, say of the patient's ALSFSR-R at one year, these adjustments may not perform well when small amounts of observed data exist. The small number of data-points can result in the variance of the prediction being very large, producing a large mean-squared-prediction-error (MSPE). Recall the MSPE is equivalent to the sum of the variance and squared bias of the prediction. In this paper we consider a simple method to reduce the variability of linear predictions at the cost of potentially increasing the predictive bias. Biased linear regression itself is not new (ridge regression (Hoerl & Kennard 2000) is one well-known example), but we do this in a unique way: by exploiting our knowledge of when the process we are modelling (e.g. the patient's disease progression) first began.

Tracking the date when a patient first began noticing symptoms of ALS (their disease onset time) is common practice in ALS clinics and trials. From a modelling perspective, one could use this information in a variety of ways: the most obvious way is using it as a covariate in the model. Let us try a different approach: if we were to assume their ALSFRS-R score at roughly the time of their disease onset, what might their ALSFRS-R be? One could argue that the patient has had minimal, if any, disease progression at time of disease onset. It seems reasonable then that one could assume their ALSFRS-R to be 48 (meaning the minimum possible disease progression) at this time. We could then create a new observation with ALSFRS-R score of 48 at the time of disease onset, and include that as one of the observations (data-points) used to build our linear model.

In this paper we consider utilizing knowledge of when a process starts to create an assumed data-point, which then can be used to reduce variability of linear model predictions. We found no previous literature on this technique in our literature search. We first show how the inclusion of this point mathematically reduces variance component of the MSPE under the assumptions of ordinary least-squares (OLS) linear regression; then we calculate the bias component it brings to the MSPE; we deduce the condition under which this approach can reduce the MSPE in predication combined variance and bias. Afterwards we give an example of utilizing this approach in the context of modeling ALS disease progression, showing how it improves the MSPE when compared to a linear model lacking the extra data-point. We show how it is also superior to a logit transform approach. We stress that this method is a simple to understand, easy to perform, and inexpensive to implement approach. It is our hope that this idea may be utilized by pragmatic researchers to improve their linear predictions and estimations at very little additional cost.

2. The Effect of Using an Anchor on the Mean Square Prediction Error in Simple Linear Regression

Here we develop the theoretical results that justify the creation and use of an extra assumed data-point to improve modelling. We shall refer to this data-point as the "anchor." Consider $n-1$ ordered pairs $\{(x_i, y_i)\}$, $i \in 1, \ldots, n-1$, where y_i is some response corresponding to x_i. As per ordinary linear regression (Kutner, Nachtsheim & Neter 2004), assume that x_i and y_i have a linear relationship, meaning that for some constants β_0 and β_1, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, with independent error terms $\epsilon_i \sim N\left(0, \sigma^2\right)$. Furthermore, assume an additional observation referred to as the "anchor" given by (x_n, y_n), where y_n is some fixed constant in R.

Consider the problem of predicting a new value y_0 corresponding to a given x_0, which is typically obtained by using the OLS estimates for β_0 and β_1, denoted as a and b. Denote the resultant prediction for y_0 which utilizes the first $n-1$ coordinate pairs by $\widehat{Y}_0^{(n-1)} = a^{(n-1)} + b^{(n-1)} x_0$, and the prediction which also includes the anchor by $\widehat{Y}_0^{(n)} = a^{(n)} + b^{(n)} x_0$. Denote the errors between our prediction and the truth to be $e_0^{(n-1)} = y_0 - \widehat{Y}_0^{(n-1)}$ and $e_0^{(n)} = y_0 - \widehat{Y}_0^{(n)}$. Recall that the variance of $e_0^{(n-1)}$ (which was built from $n-1$ ordered pairs of data in standard OLS regression) is equivalent to:

$$Var\left(e_0^{(n-1)}\right) = Var\left(y_0 - \widehat{Y}_0^{(n-1)}\right) = \sigma^2\left(1 + \frac{1}{n-1} + \frac{\left(\bar{x}^{(n-1)} - x_0\right)^2}{\sum_{i=1}^{n-1}\left(x_i - \bar{x}^{(n-1)}\right)^2}\right)$$

where $Var\left(e_0^{(n)}\right) = Var\left(y_0 - \widehat{Y}_0^{(n)}\right)$represents the variance of the prediction error obtained from utilizing all n datapoints (meaning we include the anchor).

We first show that $Var\left(e_0^{(n)}\right) \leq Var\left(e_0^{(n-1)}\right)$, meaning any choice of anchor will decrease the variance component of the MSPE. We then derive an upper bound for the bias of the anchor such that the MSPE will decrease; in other words, how far away from the "true" line can the anchor be before it makes the MSPE worse.

Without loss of generality, we will assume the following for the observed data:

Assume that $x_1, \ldots, x_{n-1}$ have been normalized such that $\bar{x}^{(n-1)} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} = 0$ and $\sqrt{\sum_{i=1}^{n-1} x_i^2} = 1$. Any collection of (x_i, y_i) can be linearly transformed in the x-coordinate by subtracting from each x_i the mean of the x's and dividing by the Euclidean norm to achieve this normalization. Explicitly, each x_j is transformed by applying $g\left(x_j\right) = \dfrac{x_j - \bar{x}^{(n-1)}}{\sqrt{\sum_{i=1}^{n-1}\left(x_i - \bar{x}^{(n-1)}\right)^2}}$. It is interesting to point out that this transformation has no impact on the OLS estimators for σ^2.

Then the following hold:

$$SSX^{(n-1)} = \sum_{i=1}^{n-1} x_i^2 = 1,$$

$$\bar{x}^{(n)} = \frac{x_n}{n} + \frac{n-1}{n}\left(\bar{x}^{(n-1)}\right) = \frac{x_n}{n},$$

$$\begin{aligned} SSX^{(n)} &= \sum_{i=1}^{n-1} x_i^2 + x_n^2 - n\left(\bar{x}^{(n)}\right)^2 \\ &= 1 + x_n^2 - \frac{x_n^2}{n}. \end{aligned}$$

2.1. Utilizing an Anchor Reduces Predictive Variability

Here we show that inclusion of an anchor in an OLS regression will always reduce the variance of newly predicted responses. Intuitively, it makes sense that the variance for the slope and intercept estimates will shrink as more points are included in the OLS regression: consider a simulation where one draws two observations and obtains the OLS estimates for the slope and intercept as compared to a simulation where one draws three observations. The latter will have less variance on the OLS estimates, resulting in less variance on newly predicted responses. The variance is then reduced even further when one assumes that the additional observation is an anchor and has a variance of zero.

Theorem 1. *For any anchor point (x_n, y_n), with y_n a fixed constant, $Var\left(e_0^{(n)}\right) \leq Var\left(e_0^{(n-1)}\right)$.*

Proof. Let a, b be the OLS estimated intercept and slope through the points $(x_1, y_1), \ldots, (x_n, y_n)$. In other words, a and b are the regression estimates for β_0 and β_1. Since y_0 and $\widehat{Y}_0^{(n)}$ are independent,$\text{Var}\left(e_0^{(n)}\right) = \text{Var}\left(y_0 - \widehat{Y}_0^{(n)}\right) = \text{Var}\,(y_0) + \text{Var}\,(a + bx_0)$. Utilizing our assumptions on $x_1 \ldots, x_{n-1}$, the inequality $\text{Var}\left(e_0^{(n)}\right) \leq \text{Var}\left(e_0^{(n-1)}\right)$holds if and only if:

$$Var\,(y_0) + Var\,(a + bx_0) \leq Var\left(e_0^{(n-1)}\right) = \sigma^2\left(1 + \frac{1}{n-1} + \frac{\left(\bar{x}^{(n-1)} - x_0\right)^2}{SSX^{(n-1)}}\right).$$

This can be simplified using our assumptions on $\bar{x}^{(n-1)}$ and $SSX^{(n-1)}$ to obtain

$$Var\,(y_0) + Var\,(a + bx_0) \leq \sigma^2\left(1 + \frac{1}{n-1} + \frac{(0 - x_0)^2}{1}\right),$$

which simplifies as follows using properties of variance:

$$\sigma^2 + Var\,(a) + x_0^2 Var\,(b) + 2x_0 Cov\,(a, b) \leq \sigma^2\left(1 + \frac{1}{n-1} + x_0^2\right),$$

$$Var\left(a\right)+x_0^2Var\left(b\right)+2x_0Cov\left(a,b\right)\leq\sigma^2\left(\frac{1}{n-1}+x_0^2\right).$$

We next consider the individual terms $Var\left(a\right),Var\left(b\right),$ and $Cov\left(a,b\right)$. For convenience SSX denotes $SSX^{(n)}$ and $\bar{x}$ denotes $\bar{x}^{(n)}$. □

Part 1: variance of slope

$$Var\left(b\right)=Var\left(\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)}{SSX}y_i\right)=\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)}{SSX^2}^2Var(y_i).$$

Recall $Var\left(y_i\right)=\sigma^2$ if $i\leq n-1$ and $Var\left(y_n\right)=0$ since y_n is a constant. Thus the nth term of the summation is zero and we can write $Var\left(b\right)$ as:

$$Var\left(b\right)=\frac{\sigma^2}{SSX^2}\sum_{i=1}^{n-1}\left(x_i-\bar{x}\right)^2=\frac{\sigma^2}{SSX^2}\sum_{i=1}^{n-1}\left(x_i^2+\bar{x}^2-2x_i\bar{x}\right).$$

Utilizing the assumption that $\sum_{i=1}^{n-1}\left(x_i^2\right)=1$ and that $\sum_{i=1}^{n-1}\left(x_i\right)=0$,

$$Var\left(b\right)=\frac{\sigma^2}{SSX^2}\left(1+\left(n-1\right)\bar{x}^2\right).$$

Or equivalently (multiply top and bottom by n)

$$Var\left(b\right)=\sigma^2\frac{n^2+nx_n^2-x_n^2}{\left(nx_n^2+n-x_n^2\right)^2}.$$

Part 2: variance of intercept

Since $Var\left(y_n\right)=0$ and $Var\left(y_i\right)=\sigma^2$ for $i\in 1,\ldots,n-1$, we use properties of the variance to find:

$$\begin{aligned}Var\left(a\right)&=Var\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{SSX}\right)y_i\right)=\sigma^2\sum_{i=1}^{n-1}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{SSX}\right)^2\\&=\sigma^2\sum_{i=1}^{n-1}\left(\frac{1}{n^2}+\frac{\bar{x}^2(x_i-\bar{x})^2}{SSX^2}-2\frac{\bar{x}\left(x_i-\bar{x}\right)}{nSSX}\right).\end{aligned}$$

Distributing the summation to each term results in

$$Var\left(a\right)=\sigma^2\left(\frac{n-1}{n^2}+\frac{\bar{x}^2\left(1+\left(n-1\right)\bar{x}^2\right)}{SSX^2}+2\frac{(n-1)\bar{x}^2}{nSSX}\right),$$

which, after multiplying as needed to get a common denominator, is equivalent to

$$Var\left(a\right)=\sigma^2\frac{nx_n^4+2nx_n^2+n-x_n^4-x_n^2-1}{\left(nx_n^2+n-x_n^2\right)^2}.$$

Part 3: covariance of intercept and slope

Consider $Cov\left(a,b\right)$. We use the property that $Cov\left(\sum c_i y_i, \sum d_i y_i\right) = \sigma^2 \sum\left(c_i d_i\right)$ and the fact that any covariance or variance term involving y_n is 0, since y_n is a constant.

$$\begin{aligned}
Cov\left(a,b\right) &= Cov\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{SSX}\right)y_i, \sum_{i=1}^{n}\frac{\left(x_i-\bar{x}\right)}{SSX}y_i\right)\\
&= Cov\left(\sum_{i=1}^{n-1}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{SSX}\right)y_i, \sum_{i=1}^{n-1}\frac{\left(x_i-\bar{x}\right)}{SSX}y_i\right)\\
&= \sigma^2\ \sum_{i=1}^{n-1}\left(\frac{1}{n}-\frac{\bar{x}\left(x_i-\bar{x}\right)}{SSX}\right)\left(\frac{\left(x_i-\bar{x}\right)}{SSX}\right)\\
&= \frac{\sigma^2}{SSX}\ \sum_{i=1}^{n-1}\left(\frac{x_i-\bar{x}}{n}-\frac{\bar{x}(x_i-\bar{x})^2}{SSX}\right)\\
&= \frac{-\sigma^2}{SSX}\ \left(\frac{n-1}{n}\left(\bar{x}\right)+\frac{\bar{x}}{SSX}\left(1+\left(n-1\right)\bar{x}^2\right)\ \right)
\end{aligned}$$

Or equivalently (after multiplying as needed to get a common denominator)

$$Cov\left(a,b\right) = -\sigma^2\frac{x_n\left(nx_n^2+2n-x_n^2-1\right)}{\left(nx_n^2+n-x_n^2\right)^2}.$$

Part 4: proving the inequality $\boldsymbol{Var\left(e_0^{(n)}\right) \leq Var\left(e_0^{(n-1)}\right)}$

Recall, $Var\left(e_0^{(n)}\right) \leq Var\left(e_0^{(n-1)}\right)$ is equivalent to the following:

$$Var\left(a\right)+x_0^2 Var\left(b\right)+2x_0 Cov\left(a,b\right) \leq \sigma^2\left(\frac{1}{n-1}+x_0^2\right).$$

Substituting the previously derived terms on the left hand side results in a statement which is trivially true if $\sigma^2 = 0$. Otherwise, this statement is equivalent to

$$\begin{aligned}
0 \leq x_0^2\left(1-\frac{n+nx_n^2-x_n^2}{\left(nx_n^2+n-x_n^2\right)^2}\right)+x_0\left(2\frac{x_n\left(nx_n^2+2n-x_n^2-1\right)}{\left(nx_n^2+n-x_n^2\right)^2}\right)\\
+\left(\frac{1}{n-1}-\frac{nx_n^4+2nx_n^2+n-x_n^4-x_n^2-1}{\left(nx_n^2+n-x_n^2\right)^2}\right).
\end{aligned}$$

The right hand side of the inequality is quadratic in x_0^2 with form $g\left(x_0\right) = Ax_0^2+Bx_0+C$. Note the coefficients A, B, C simplify to single terms in the following way:

$$A = \frac{\left(n-1\right)x_n^2\left(nx_n^2+2n-x_n^2-1\right)}{\left(nx_n^2+n-x_n^2\right)^2},$$

$$B = \frac{2x_n\left(nx_n^2 + 2n - x_n^2 - 1\right)}{\left(nx_n^2 + n - x_n^2\right)^2},$$
$$C = \frac{\left(nx_n^2 + 2n - x_n^2 - 1\right)}{(n-1)\left(nx_n^2 + n - x_n^2\right)^2}.$$

Since $A > 0$ for $n > 2$, then $g(x_0)$ is an upward-facing parabola. Also, the discriminant, given by $B^2 - 4AC$ is equal to zero:

$$\begin{aligned} B^2 - 4AC &= \frac{4x_n^2\left(nx_n^2 + 2n - x_n^2 - 1\right)^2}{\left(nx_n^2 + n - x_n^2\right)^4} \\ &\quad - 4\frac{(n-1)\,x_n^2\left(nx_n^2 + 2n - x_n^2 - 1\right)}{\left(nx_n^2 + n - x_n^2\right)^2}\frac{\left(nx_n^2 + 2n - x_n^2 - 1\right)}{(n-1)\left(nx_n^2 + n - x_n^2\right)^2} \\ &= \frac{4x_n^2\left(nx_n^2 + 2n - x_n^2 - 1\right)^2}{\left(nx_n^2 + n - x_n^2\right)^4} - \frac{4x_n^2\left(nx_n^2 + 2n - x_n^2 - 1\right)^2}{\left(nx_n^2 + n - x_n^2\right)^4} = 0, \end{aligned}$$

meaning there is exactly one root in $g(x_0)$. Therefore, it must be true that $g\left(x_0\right) \geq 0$ and $Var\left(\widehat{Y}_0^{(n)}\right) \leq Var\left(\widehat{Y}_0^{(n-1)}\right)$as desired.

Thus we see that any choice of anchor will necessarily result in a reduction in the variance of the prediction of y_0, which is equivalent to a reduction of the variance component of the MSPE. However, we still need to consider the bias. Recall that the typical OLS estimators for slope and intercept are unbiased, but this is not necessarily true when including an anchor. We next consider how much bias will be introduced by an anchor to the estimators for the slope and intercept, and the effect this has on the MSPE (compared to the MSPE that does not include an anchor). It will be shown that any choice of an anchor (x_n, y_n) such that $y_n \neq \beta_0 + \beta_1 x_n$ will introduce bias to the model. Note that the bias is a direct function of β_0 and β_1, which are rarely known in practice. Again, let $\bar{x}$ denote $\bar{x}^{(n)}$ and SSX denote $SSX^{(n)} = \sum_{i=1}^{n} x_i^2$.

2.2. Predictive Bias Caused by Utilizing an Anchor

There is no such thing a free lunch, and while using an anchor brings the benefit of predictive variance reduction, it can potentially inject bias into the predictions. Here we quantify this bias in terms of the true regression slope (Theorem 2) and intercept (Theorem 3). These biases propagate in to the prediction of $\widehat{Y}_0^{(n)}$ directly.

Theorem 2. *Using anchor point (x_n, y_n) results in biasing the slope by*

$$E\left(b - \beta_1\right) = \frac{(n-1)\,x_n\left(y_n - \beta_0\right) + \beta_1 n}{nSSX} - \beta_1.$$

This result shows that an anchor will almost always bias the estimate for the slope, however we will show that no bias is added when the anchor lies directly on the true regression line as a corollary.

***Proof*.** Recall $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and that the OLS estimate for β_1, denoted by b, is given by

$$b = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right) y_i}{SSX} = \frac{\sum_{i=1}^{n} \left(x_i - \frac{x_n}{n}\right) y_i}{SSX}.$$

We first derive $E(b)$:

$$\begin{aligned} E\left(b\right) &= E\left(\frac{\sum_{i=1}^{n} \left(x_i - \frac{x_n}{n}\right) y_i}{SSX}\right) \\ &= E\left(\frac{\sum_{i=1}^{n-1} \left(x_i - \frac{x_n}{n}\right) y_i}{SSX} + \frac{\left(x_n - \frac{x_n}{n}\right) y_n}{SSX}\right). \end{aligned}$$

Recall that y_n is a nonrandom constant, and hence $E\left(y_n\right) = y_n$. Then we can partition the expectation of the summation:

$$\begin{aligned} E\left(b\right) &= \frac{1}{SSX} E\left(\sum_{i=1}^{n-1} \left(x_i y_i - \frac{y_i x_n}{n}\right)\right) + \frac{1}{SSX}\left(x_n - \frac{x_n}{n}\right) y_n \\ &= \frac{1}{SSX} E\left(\sum_{i=1}^{n-1} \left\{x_i y_i\ \right\} - \bar{y}_{n-1} x_n \left(\frac{n-1}{n}\right)\right) + \frac{1}{SSX}\left(x_n - \frac{x_n}{n}\right) y_n. \end{aligned}$$

Note that the OLS estimate for β_0 when not using the anchor point is given by $a^{(n-1)} = \bar{y}^{(n-1)} - b^{(n-1)}\bar{x}^{(n-1)} = \bar{y}^{(n-1)}$ since $\bar{x}^{(n-1)} = 0$. Similarly, $b^{(n-1)} = \sum_{i=1}^{n-1} \left\{x_i y_i\right\}$. Since these are the unbiased OLS estimators for β_0 and β_1 when ignoring the anchor point, then it must be that $E\left(\bar{y}^{(n-1)}\right) = \beta_0$ and $E\left(\sum_{i=1}^{n-1} \left\{x_i y_i\right\}\right) = \beta_1$. Using these values and the linearity of expectation, we then have

$$\begin{aligned} E(b) &= \frac{1}{SSX}\left(\beta_1 - \beta_0 x_n \left(\frac{n-1}{n}\right)\right) + \frac{1}{SSX}\left(x_n - \frac{x_n}{n}\right) y_n \\ &= \frac{1}{SSX}\left(\beta_1 + x_n y_n - \beta_0 x_n \left(\frac{n-1}{n}\right) - \frac{x_n y_n}{n}\right) \\ &= \frac{1}{SSX}\left(\beta_1 + \left(\frac{n-1}{n}\right) x_n y_n - \beta_0 x_n \left(\frac{n-1}{n}\right)\right) \\ &= \frac{1}{SSX}\left(\beta_1 + x_n \left(\frac{n-1}{n}\right)\left(y_n - \beta_0\right)\right). \end{aligned}$$

Or equivalently

$$E\left(b\right) = \frac{\left(n-1\right) x_n \left(y_n - \beta_0\right) + \beta_1 n}{nSSX},$$

which means the bias of b is given by

$$E\left(b - \beta_1\right) = \frac{\left(n-1\right) x_n \left(y_n - \beta_0\right) + \beta_1 n}{nSSX} - \beta_1.$$

□

We next quantify the bias added to the estimate of the intercept parameter when using an anchor.

Theorem 3. *Using anchor point (x_n, y_n) results in biasing the intercept by*

$$E\left(a-\beta_0\right)=\frac{\beta_0\left(n-1\right)\left(x_n^2+1\right)-\beta_1 x_n^2+y_n}{nSSX}-\beta_0.$$

Similarly to Theorem 2, this result shows that an anchor will almost always bias the estimate for the intercept. Again, this bias is minimized when choosing an anchor that is closer to being on the true regression line. No bias is added when the anchor lies directly on the true regression line.

Proof. Recall that the OLS estimate for β_0, denoted by a, is given by

$$a=\bar{y}^{(n)}-b\bar{x}=\bar{y}^{(n)}-\frac{bx_n}{n}.$$

We first calculate $E\left(a\right)$:

$$\begin{aligned}E(a)&=E\left(\bar{y}^{(n)}\right)-\frac{x_n}{n}E\left(b\right)\\&=\frac{1}{n}E\left(\left(n-1\right)\bar{y}^{(n-1)}+y_n\right)-\frac{x_n}{n}E\left(b\right).\end{aligned}$$

Again, recall that $E\left(\bar{y}^{(n-1)}\right)=\beta_0$ and that $E\left(y_n\right)=y_n$. We derived $E(b)$ in Theorem 2. Thus:

$$E\left(a\right)=\frac{\left(n-1\right)}{n}\beta_0+\frac{y_n}{n}-\frac{x_n}{n^2SSX}\left(\left(n-1\right)x_n\left(y_n-\beta_0\right)+\beta_1 n\right),$$

which can be reduced to

$$E\left(a\right)=\frac{\beta_0\left(n-1\right)\left(x_n^2+1\right)-\beta_1 x_n+y_n}{nSSX}.$$

Therefore the bias of the intercept is

$$E\left(a-\beta_0\right)=\frac{\beta_0\left(n-1\right)\left(x_n^2+1\right)-\beta_1 x_n+y_n}{nSSX}-\beta_0.$$

□

With Theorems 2 and 3, we can combine these using the linearity of expected values to determine the bias when predicting a new response $\widehat{Y}_0^{(n)}$.

Corollary 1. *The overall bias induced by using anchor point (x_n, y_n) is given by*

$$\begin{aligned}E\left(\widehat{Y}_0^{(n)}-y_0\right)&=E\left(a+bx_0-\beta_0-\beta_1 x_0\right)\\&=\frac{\beta_0\left(n-1\right)\left(x_n^2+1\right)-\beta_1 x_n+y_n}{nSSX}\\&\quad-\beta_0+x_0\left\{\frac{\left(n-1\right)x_n\left(y_n-\beta_0\right)+\beta_1 n}{nSSX}-\beta_1\right\},\end{aligned}$$

which can be reduced algebraically to

$$E\left(\widehat{Y}_0^{(n)} - y_0\right) = \frac{1+(n-1)\,x_n x_0}{nSSX}\,(y_n - \beta_0 - \beta_1 x_n).$$

We next show when no bias is added when using an anchor. As mentioned previously, this typically happens only when the anchor lies directly on the true regression line; in other words when the anchor (x_n, y_n) is such that $y_n = \beta_0 + \beta_1 x_n$.

Corollary 2. *When using an anchor to predict y_0 for any given $x_0 \neq \frac{-1}{(n-1)x_n}$, no bias is introduced by the anchor, meaning $E\left(\widehat{Y}_0^{(n)}\right) = y_0$, if and only if $y_n = \beta_0 + \beta_1 x_n$.*

Proof. Recall the overall bias is given by $E\left(\widehat{Y}_0^{(n)} - y_0\right) = \frac{1+(n-1)x_n x_0}{nSSX}$ $(y_n - \beta_0 - \beta_1 x_n)$. This is zero if and only if either $y_n - \beta_0 - \beta_1 x_n = 0$ or $1 + (n-1)\,x_n x_0 = 0$, which is equivalent to $y_n = \beta_0 + \beta_1 x_n$ or $x_0 = \frac{-1}{(n-1)x_n}$. Thus the overall bias is zero if and only if the anchor point is on the true regression line, given that you are not predicting where $x_0 = \frac{-1}{(n-1)x_n}$. □

2.3. Using an Anchor to Reduce the Mean Square Predictive Error

We combine the previous theorems to deduce exactly when using an anchor will improve the MSPE of predicting a new response. If the variability is reduced more than the square of the bias is increased, the MSPE will shrink, which is desired. In Theorem 4 we derive an exact bound for when this occurs.

Theorem 4. *Utilizing anchor point (x_n, y_n) reduces the overall MSPE when the following inequality holds:*

$$\left(\frac{\beta_0\,(n-1)\left(x_n^2+1\right) - \beta_1 x_n + y_n}{nSSX} - \beta_0 + x_0\left\{\frac{(n-1)\,x_n\,(y_n-\beta_0) + \beta_1 n}{nSSX} - \beta_1\right\}\right)^2 \leq$$

$$x_0^2\left(\frac{(n-1)\,x_n^2\left(nx_n^2+2n-x_n^2-1\right)}{n^2SSX^2}\right) + x_0\left(\frac{2x_n\left(nx_n^2+2n-x_n^2-1\right)}{n^2SSX^2}\right) + \frac{\left(nx_n^2+2n-x_n^2-1\right)}{(n-1)\,n^2SSX^2}.$$

Note that this bound is a function of the true regression slope and intercept, β_1 and β_0. In practice, these are rarely known, which makes this bound difficult to use as a decision rule for including the use of an anchor (at least, outside of a simulation).

Proof. Consider the following inequality

$$MSPE^{(n)} \leq MSPE^{(n-1)}.$$

This is equivalent to

$$Bias^2\left(e_0^{(n)}\right)+Var\left(e_0^{(n)}\right)\leq Bias^2\left(e_0^{(n-1)}\right)+Var\left(e_0^{(n-1)}\right)$$

$$Bias^2\left(e_0^{(n)}\right)\leq Var\left(e_0^{(n-1)}\right)-Var\left(e_0^{(n)}\right).$$

But recall $Bias^2\left(e_0^{(n-1)}\right)=0$, and that $Var\left(e_0^{(n-1)}\right)=\sigma^2\left(1+\frac{1}{n-1}+x_0^2\right)$. The remaining pieces, $Bias^2\left(e_0^{(n)}\right)$ and $Var\left(e_0^{(n)}\right)$, were derived in Theorem 2 and Theorem 3, and substituting them in to this inequality results in the formula given in the statement of Theorem 4.

□

Thus we see that any choice of anchor point will reduce the variance of prediction, but will almost always increase the bias of the prediction depending on how far away the anchor point is from the "true" regression line. To see why the bias increases based on how far the anchor is from the true regression line, observe that the square of the total bias is quadratic in y_n, which must have exactly one root at the vertex. Given that $x_0 \neq \frac{-1}{(n-1)x_n}$, this root occurs only when $y_n=\beta_0+\beta_1 x_n$. Because it is quadratic in y_n, the square of the bias will increase as (x_n, y_n) moves further away from the true regression line. Therefore, using an anchor may be beneficial or not, depending on how much bias is added.

The bound calculated in Theorem 4 could potentially be used as a decision rule for determining if using an anchor is beneficial or not. Unfortunately, one needs to know the true values of β_0 and β_1 in order to use Theorem 4's result. In practice, one tends to not know the true regression parameters (which would result in no need of including an anchor), although with sufficiently informed prior knowledge, precise estimates may exist. Thus, when deciding whether to use an anchor or not, we suggest comparing the anchor model to a standard model by validating the model in some way (perhaps via a cross-validation approach). We show an example of this in Section 3.

Before moving to the application section, we note that many of the ideas in this paper have Bayesian connections. For example, consider performing a Bayesian analysis of classical regression. When utilizing the standard noninformative prior distribution, the posterior mean estimates for the slope and intercept terms (and their standard errors) are equivalent to those obtained under frequentist OLS (Gelman 2014). It follows that Theorems 1-4 still hold under the Bayesian paradigm, meaning that an anchor can be utilized to reduce the variance of posterior predictions.

3. Application to ALS Prediction

We next consider using an anchor to improve linear models that pertain to predicting disease progression in patients with ALS. Note that the theory developed in part (2) applies to a single OLS regression (prediction for the individual). The following example expands on this by showing how utilizing an anchor can improve the average prediction error across several OLS regressions (prediction for each of several individuals).

Our data comes from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database (Atassi, Berry, Shui, Zach, Sherman, Sinani, Walker, Katsovskiy, Schoenfeld, Cudkowicz & Leitner 2014). In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data available in the PRO-ACT Database has been volunteered by PRO-ACT Consortium members (`https://nctu.partners.org/PRO-ACT/`).

Recall ALS disease progression is tracked by the ALSFRS-R (see Section 1), our outcome variable, which is an integer value between 0 and 48, where 48 represents the minimal amount of disease progression and 0 represents the maximal progression. For each patient, we model the ALSFRS-R versus time (in days). Specifically, time is measured in days from trial baseline, meaning $x = 0$ corresponds to the beginning of the trial and $x = 365$ corresponds to the 365^{th} day after the trial began. On this scale, a patient's disease onset time is typically negative, as it happened before the trial began. We required patients to have the following: (1) at least two recorded ALSFRS-R scores before 3 months, for model building purposes; (2) non-missing value for time of disease onset; (3) at least one year between the baseline and last ALSFRS-R score for MSPE-validation purposes. This resulted in 1606 patients, with an average $\pm$ standard error of 12 ± 4.54 time-points per patient (and 3 ± 0.96 visits in the first three months).

Note that we are now considering data on several distinct patients, each with their own ALSFRS-R trajectory. To demonstrate how utilizing the anchor-point improves OLS regression, we will simply model each patient independently with (1) a standard OLS regression model and (2) with an OLS regression model utilizing an anchor. Note that the ALSFRS-R follows a fairly linear decline, with wildly varying between-patient progression rates, justifying using linear models (Figure 1). The assumed data-point, or anchor, utilized in the anchor model comes from assuming minimal disease progression at the time of disease onset. In other words, each patient's data is augmented with the additional data point given by the ordered pair $(x_{onset}, 48)$, since 48 is the ALSFRS-R corresponding to minimal progression.

Our validation method is as follows: we will compare the standard model versus the anchor model by comparing their ability to predict each patient $k's$ first ALSFRS-R score after 365 days (1 year), observed at time $x_{k,0}$, using only ALSFRS-R scores measured before 92 days (3 months). Specifically for both models we calculate (for the 1606 patients)

$$\sqrt{MSPE} = \sqrt{\frac{\sum_{k=1}^{1606} \left(\widehat{Y}_k - Y_k\right)^2}{1606}},$$

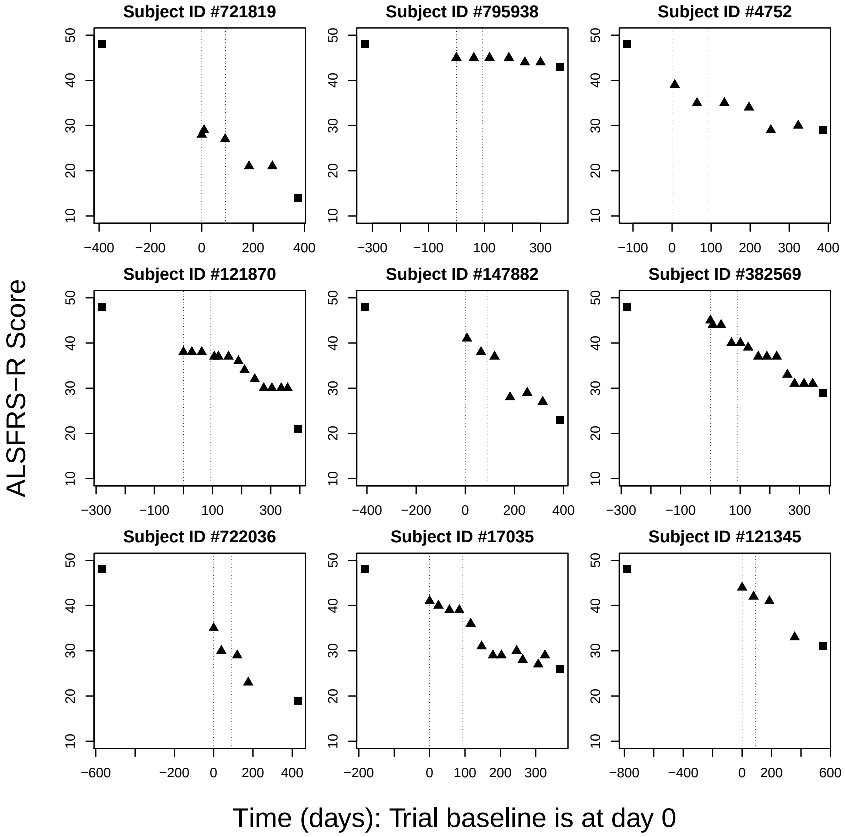

Figure 1: For nine randomly selected subjects, we plotted their ALSFRS-R versus time. The leftmost black square is the anchor, the rightmost square is the observed value Y_k, and the gray triangles are observed scores. The dashed black lines denote days 0 and 92 of the trial, meaning observations between the two dashed lines were used for model fitting.

where $\widehat{Y}_k$ is the predicted ALSFRS-R score for patient k at time $x_{k,0}$ and Y_k is the true ALSFRS-R score at time $x_{k,0}$. Because we know the ALSFRS-R is bounded between 0 and 48, any model prediction that falls outside these bounds will be truncated to the closest boundary value before evaluating the MSPE. To assist in visualizing this data, Figure 1 shows the progression of the ALSFRS-R versus time for nine subjects (simple random sample without replacement).

The anchor model results in slightly more biased predictions compared to those of the standard model, as expected. However, as demonstrated in the methods section, the variance of these errors is much smaller for the anchor model (Figure 2). The resulting root MSPE of the anchor model is 7.8 while the standard model's MSPE is 13.0: we observe a large drop in prediction error when including the anchor.

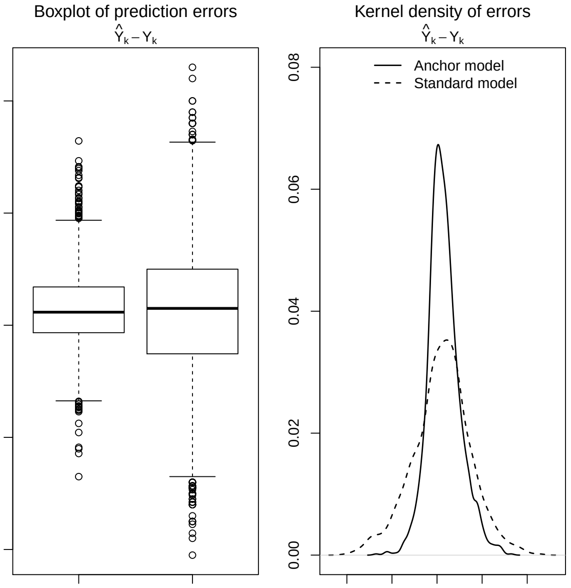

FIGURE 2: The raw prediction error for the anchor and standard model. The models' mean error as measured by $\widehat{Y}_k - Y_k$ was 3.1 and 2.1 respectively, with standard deviations of 7.1 and 12.7.

It can be shown that for some patients, the prediction from the standard model is closer to the truth than the prediction from the anchor model. Perhaps we should only use the anchor model when the increase in bias is negligible? We could explore this by taking the difference between the prediction from the anchor model $\widehat{Y}^{(a)}$ and the standard model $\widehat{Y}^{(s)}$; if this difference is sufficiently small in magnitude then the increase in bias from using the anchor model is negligible on average. In other words, for each patient consider calculating $T_k = \widehat{Y}_k^{(a)} - \widehat{Y}_k^{(s)}$, and then defining the prediction for patient k as

$$\widehat{Y}_k = \begin{cases} \widehat{Y}_k^{(a)} & if \ |T_k| \leq \Gamma \\ \widehat{Y}_k^{(s)} & otherwise \end{cases}$$

for some constant Γ. Figure 3 shows how this changes the MSPE for various values of Γ, as well the result from changing the rule to be $\widehat{Y}_k = \widehat{Y}_k^{(a)}$ if $|T_k| \geq \Gamma$ instead (meaning choose the anchor model if the difference in the model predictions is large). From Figure 3 we see that naively using the anchor model for all patients outperforms any of the Γ and T_k decision-rule hybrids for this dataset.

Finally, we compare the anchor model to that of a logit transform model. The logit transform is a model which is more advanced, yet also more difficult to calculate and interpret. The logit transform model was chosen because it is one of the easier-to-understand models for modelling bounded data. We fit the logit transform model by taking each ALSFRS-R score, dividing it by its maximum of 48, and fitting the resultant data (which is bounded between 0 and 1) with a regression model. In other words, for a given patient we fit the following model: $logit\left(\frac{y_i}{48}\right) = \beta_0 + \beta_1 x_i + \epsilon_{ij}$, where ϵ_{ij} are independent errors that follow $N(0, \sigma^2)$, β_0 and β_1 are the intercept and slope parameters, x_i is the time-point associated with ALSFRS-R score y_i. The MSPE of this model comes to be 14.65, significantly higher than the MSPE for either the standard OLS model (12.95) or the anchor model (7.78).

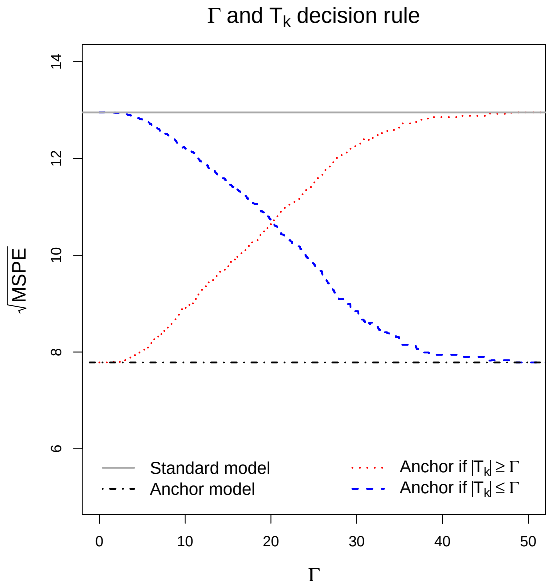

Figure 3: Shows the resulting MSPE for various cutoffs of Γ. Note that since the MSPE is bounded below by the anchor model ($\sqrt{MSPE} = 7.78$), this shows that the anchor model is uniformly better than the linear model ($\sqrt{MSPE} = 12.95$) for this data.

4. Discussion

In this paper, we discussed a simple and computationally inexpensive technique that may improve the predictive power in linear models. This method consists of creating an additional assumed data-point, referred to as an anchor, and including it in the OLS regression. This is different than fixed-intercept regression, as it allows the more weight to be put on the data with respect to parameter estimation. It has been shown in this paper that including an anchor theoretically decreases prediction variance at the cost of potentially increased bias. We demonstrated how using an anchor can improve linear predictions from modelling disease progression in ALS patients.

Fitting the anchor model can be performed as easily and efficiently as a standard OLS regression, yet has the potential to be a much stronger predictive model. Furthermore, the interpretations of the anchor model's parameters remain largely unchanged from that of OLS regression, which is a huge advantage over other models: the interpretability of the parameters is arguably one of the most attractive parts of linear models.

We imagine that utilizing an anchor in the way we have demonstrated will be of particular use when modelling a bounded linear process when one can obtain a measure of when the process first began and/or ended. The bounds give a justification for choosing the y-value of the anchor; without bounds it may be difficult to justify a value without first looking at the data, potentially leading to overfitting. However, as long as monotonicity approximately holds, one could still use something such as the 99^{th} percentile of the response in lieu if no bound exists.

While the results in this paper are done under the assumptions of frequentist OLS regression, it is in no way limited to this. The idea of utilizing this additional data-point can easily extend to other families of models such as generalized linear models, hierarchical models, and mixed models. For example, one can dramatically improve the model performance in the ALS example by switching from independent linear regressions for each patient to a Bayesian hierarchical model; this allows patients to borrow information from one another and results in improved estimators due to shrinkage (Morris & Lysy 2012). This model is improved even further when it becomes a Bayesian hierarchical model that utilizes an anchor for each patient (Karanevich, Statland, Gajewski & He 2018). While less straightforward than OLS regression, one could also include additional random error associated with the anchor (either on the x_n, y_n, or both) in their model to allow for more flexibility to the approach.

Deciding when to include an anchor for modelling is not straightforward. If the goal is estimation, the induced bias may not be worth the reduced variability in estimates. While we developed a theoretical bound for when an anchor will improve the MSPE, it depends on having theoretical knowledge of the underlying linear process, which is rarely possible in practice. Thus we recommend using some sort of prediction validation (such as cross-validation) to compare utilizing an anchor versus a more standard approach. The validation scheme used in our ALS example is one way of doing this. Other models might benefit more from cross-validation, which builds a model on a training set and tests the model on a withheld validation set. Because some sort of validation is good standard practice when evaluating predictive models, we feel that this is a very small price to pay for a potentially dramatic improvement in predictive error.

Acknowledgment

Study Funded by the Mabel A. Woodyard Fellowship in Neurodegenerative Disorders and the Roofe Fellowship in Neuroscience Research. This work was supported by a CTSA grant from NCRR and NCATS awarded to the University of Kansas Medical Center for Frontiers: The Heartland Institute for Clinical and

Translational Research # UL1TR000001 (formerly #UL1RR033179). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH, NCRR, or NCATS.

[Received: October 2017 — Accepted: March 2018]

References

Amemiya, T. (1973), 'Regression analysis when the dependent variable is truncated normal', *Econometrica* **41**(6), 997–1016.

Armon, C., Graves, M., Moses, D., Forté, D., Sepulveda, L., Darby, S. & Smith, R. (2000), 'Linear estimates of disease progression predict survival in patients with amyotrophic lateral sclerosis', *Muscle & Nerve* **23**(6), 874–882.

Atassi, N., Berry, J., Shui, A., Zach, N., Sherman, A., Sinani, E., Walker, J., Katsovskiy, I., Schoenfeld, D., Cudkowicz, M. & Leitner, M. (2014), 'The pro-act database: design, initial analyses, and predictive features.', *Neurology* **83**(19), 1719–1725.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015), Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, *in* 'Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 1721–1730.

Cedarbaum, J., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B. & Nakanishi, A. (1999), 'The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function', *Journal of the Neurological Sciences* **169**(1), 13–21.

Gelman, A. (2014), *Bayesian data analysis*, tercera edn, CRC Press, Boca Raton, FL.

Hoerl, A. & Kennard, R. (2000), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **42**(1), 80–86.

Karanevich, A., Statland, J., Gajewski, B. & He, J. (2018), 'Using an onset-anchored bayesian hierarchical model to improve predictions for amyotrophic lateral sclerosis disease progression', *BMC Medical Research Methodology* **18**(1), 19.

Kutner, M., Nachtsheim, C. & Neter, J. (2004), *Applied linear regression models*, cuarta edn, McGraw-Hill, New York.

Lesaffre, E., Rizopoulos, D. & Tsonaka, R. (2007), 'The logistic transform for bounded outcome scores', *Biostatistics* **8**(1), 72–85.

Magnus, T., Beck, M., Giess, R., Puls, I., Naumann, M. & Toyka, K. (2002), 'Disease progression in amyotrophic lateral sclerosis: Predictors of survival', *Muscle & Nerve* **25**(5), 709–714.

Morris, C. & Lysy, M. (2012), 'Shrinkage estimation in multilevel normal models', *Statistical Science* **27**(1), 115–134.