Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data

Robert G. Clark[a,*], Wade Blanchard[b], Francis K.C. Hui[a], Ran Tian[c], Haruka Woods[d]

[a] *Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia*
[b] *Fenner School of Environment & Society, The Australian National University, Canberra, Australia*
[c] *School of Information and Physical Sciences, University of Newcastle, Newcastle, Australia*
[d] *School of Literature, Language and Linguistics, The Australian National University, Canberra, Australia*

ARTICLE INFO

Keywords:
Categorical data
Complete separation
Corpus data
Logistic regression
Multi-level models
Multinomial regression
Random effects
Second language acquisition

ABSTRACT

Logistic regression is a powerful and widely used analytical tool in linguistics for modelling a binary outcome variable against a set of explanatory variables. One challenge that can arise when applying logistic regression to linguistics data is complete or quasi-complete separation, phenomena that occur when (paradoxically) the model has too much explanatory power, resulting in effectively infinite coefficient estimates and standard errors. Instead of seeing this as a drawback of the method, or naïvely removing covariates that cause separation, we demonstrate a straightforward and user-friendly modification of logistic regression, based on penalising the coefficient estimates, that is capable of systematically handling separation. We illustrate the use of penalised, multi-level logistic regression on two clustered datasets relating to second language acquisition and corpus data, showing in both cases how penalisation remedies the problem of separation and thus facilitates sensible and valid statistical conclusions to be drawn. We also show via simulation that results are not overly sensitive to the amount of penalisation employed for handling separation.

1. Introduction

Categorical data is frequently collected in linguistics research, and there exists a variety of statistical methods for analysing such data. Of these, one of the most popular is logistic regression which models the probability of a binary outcome of interest as a function of other explanatory variables or covariates. A major advantage of logistic regression models is that they are able to capture the joint effect of multiple explanatory variables on the outcome while possessing sensible interpretations in terms of the effects on the odds of the outcome. For these reasons among others, logistic regression has been widely applied to analyse data in linguistic studies (e.g., Brezina, 2018, Chapter 4; Cheung & Zhang, 2016; De Cuypere et al., 2014; Hinrichs & Szmrecsanyi, 2007; among many others). Two specific examples include that of Levshina (2015), who applied logistic regression to examine the impact of causation type, transitivity, and country on the use of a Dutch causative construction (either *doen* or *laten,* the latter is thought to denote indirect causation) measured for sentences in Netherlandic and Flemish newspapers, and Johnson (2008) in socio-linguistics, who collected data on 120 native English-speaking store clerks working in various stores in Ohio, U.S.A. and applied logistic regression to study the relationship between the use of two binary tokens of words containing the 'str' sound and age group, social class, and gender.

When applying logistic regression in linguistic analysis, a common but not well understood problem that can arise is when some of the coefficients are estimated to be effectively infinite in magnitude (Hosmer et al., 2013, Chapter 4). This occurs due to a phenomenon

[*] Corresponding author.
E-mail address: robert.clark@anu.edu.au (R.G. Clark).

https://doi.org/10.1016/j.rmal.2023.100044
Received 28 September 2022; Received in revised form 30 January 2023; Accepted 30 January 2023

called separation in the data, of which there are two main types: (1) complete separation which is the extreme case where the zeros and ones of a binary response can be separated perfectly according to a linear function of the covariates; and (2) quasi-complete separation which occurs when the binary outcome is either 100% positive or 100% negative for all observations in a cell defined by the covariates. In both situations, standard statistical software will return estimated logistic regression coefficients, which are extremely large positive or negative values, along with corresponding standard errors which are also very large and are effectively meaningless. Consequently, standard inferencing methods such as confidence intervals and hypothesis testing for these covariates may not be practically useful in the presence of separation.

Complete or quasi-complete separation sometimes occurs because the model contains an excessive number of explanatory variables or because there are categorical explanatory variables with very small sample sizes in some categories. An extreme case occurs when there is multi-collinearity in the covariates, in which case separation always occurs (Zeng & Zeng, 2019). In these cases, the solution often employed in linguistics analysis is straightforward; the model can be simplified by removing explanatory variables or combining categories (e.g., Levshina, 2015). However, separation can also arise if one or more covariates are extremely successful in predicting the binary outcome of interest; put simply, the model is a victim of its own success. In such cases, simplifying the model as described above is a poor option because it removes the most powerful explanatory variables or categories from the model.

In this article, we advocate for *penalised logistic regression (PLR)* which has been proposed by Firth (1993) and others as a more systematic and data-driven solution to handle separation in binary outcome data. As the name suggests, PLR slightly penalises or shrinks coefficients based on their large absolute values, such that coefficient estimates and standard errors (and ultimately statistical inference) become usable. We demonstrate how PLR can be straightforwardly implemented in software and numerically compare several different varieties of penalisation. Our paper provides a complement to the recent work of Kimball et al. (2019) who discussed the separation problem in the context of multi-level (i.e., when random effects are also included) logistic regression for linguistic data using the RStan package in R with two case studies (a psycholinguistic and a perception study). Specifically, we illustrate the use of the blme package in R for Bayesian PLR, which is both simpler to use and computationally more efficient, while also showing that results are generally insensitive to the precise choice of prior distribution (penalisation). Kimball et al. demonstrated the use of a different Bayesian approach but did not consider sensitivity to penalties or priors; we also include a wider review and apply methods to two new case studies.

The rest of this article is structured as follows: Section 2 provides a review of logistic regression with a focus on linguistics and the issue of complete and quasi-complete separation along with the use of PLR to remedy this problem. In Section 3, we present a simulation study to assess the performance of PLR for handling separation in multi-level logistic regression, showing that different degrees of penalisation produce practically equivalent results in many cases. In Section 4, we illustrate two case studies of PLR with a socio-linguistic dataset of Japanese language learners and logistic regression of a corpus dataset. Section 5 summarises our findings and discusses some other issues around complete and quasi-complete separation.

2. A review of logistic regression and separation

2.1. Overview of logistic regression

Consider a set of observations on a sample of units $i = 1, \dots, n$ (e.g. people or tokens) such that for the i-th unit we record a binary outcome variable Y_i (taking values 0 and 1 while noting it is arbitrary which of these is defined as "positive" or "success") and a set of p covariates $x_{i1}, \dots, x_{ip.}$ The covariates may be continuous variables, such as a person's age, although commonly in linguistics the covariates arise from categorical factors with two or more levels such as gender, nationality, or language proficiency. In the latter case, the covariates represent a set of 0/1 indicator or dummy variables for each level of the factor with reference to a baseline level. Interactions are similarly defined as indicator variables for combinations of levels in multiple factors (Johnson et al., 2008).

In linguistics, among other disciplines, logistic regression is a common statistical method for modelling the probability of a positive outcome (i.e., $Y_i = 1$) as a function of the covariates. Assuming the observations are independent, we use the following model for the distribution of Y_i conditional on a vector of p covariates $x_i = (x_{i1}, \dots, x_{ip})^T$:

$$P\left(Y_i = 1|x_i\right) = \frac{e^{\eta_i}}{(1 + e^{\eta_i})}, \tag{1}$$

where $\eta_i = \beta_0 + x_{i1}\beta_1 + \dots x_{ip}\beta_p$ is the linear predictor and $\beta_1, \dots, \beta_p$ denote a set of regression coefficients with the first element (β_0) representing the intercept. Eq. (1) may be equivalently written as:

$$log\left(\frac{P\left(Y_i = 1|x_i\right)}{1 - P\left(Y_i = 1|x_i\right)}\right) = \eta_i = \beta_0 + x_{i1}\beta_1 + \dots x_{ip}\beta_p, \tag{2}$$

where $\frac{P(Y_i=1|x_i)}{1-P(Y_i=1|x_i)}$ is referred to as the odds of a positive outcome. In turn, the coefficients β_k for $k = 1, \dots, p$ in logistic regression possess a straightforward interpretation as the amount by which the log-odds increase (or decrease) when covariate k is incremented by one unit. The exponentials of β_k are referred to as odds ratios and are the factor by which the odds of a positive outcome are multiplied when covariate k is incremented by one unit. We refer the reader to Hosmer et al. (2013) for a general discussion of logistic regression, including a discussion of model fitting by maximum likelihood (where the likelihood of the observed data is maximised with respect to the parameters) and Bayesian (where prior distributions are assumed to represent the state of knowledge or uncertainty about the parameters prior to data collection before methods such as Markov Chain Monte Carlo sampling are employed) estimation methods.

There are a number of ways in which the logistic regression model in Eqs. (1) and (2) can be extended. Data in linguistics are often multi-levelled such that units represent exam questions, for instance, with multiple questions for each person. One example of this is the second language acquisition (SLA) study described in Section 4; see also Peters et al. (2016) and Godfroid et al. (2015). Another example comes from the corpus case study obtained from Lozano (2016) and referenced in Section 4 where units represent tokens and multiple tokens are measured for each person. In either case, the observations are clustered in that multiple outcome variables are measured within each unit (person) and applying the standard logistic regression model in such settings can produce misleading conclusions by not taking the multi-level nature of data into account. To overcome this, multi-level logistic regression can be used where, in conjunction with the covariates, one or more random effects are included to model the dependencies due to clustering. Let Y_{ij} be the binary outcome variable taking on values of 0 or 1 for unit j (e.g., a token or question) in cluster i (e.g., a person). Similarly, let $x_{ij1}, \ldots, x_{ijp}$ be the covariates associated with unit j in cluster i. A multi-level logistic regression model is then defined by:

$$log\left(\frac{P\left(Y_{ij}=1|x_i\right)}{1-P\left(Y_{ij}=1|x_i\right)}\right) = \eta_{ij} = u_i + \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_p x_{ijp}, \tag{3}$$

where u_i for $i = 1, \ldots, n$ is a set of random effects (random intercepts) assumed to be independently normally distributed variables with a mean of zero and unknown standard deviation. These values are latent variables reflecting each cluster's (e.g., a person's) different propensity towards a positive or negative outcome. Eq. (3) can be further extended to include random slopes and more complex random effects as appropriate (e.g., McCulloch at al., 2008). For ease of presentation, in this paper we restrict our attention to random intercepts to illustrate the handling of separation in multi-level models because random slopes are considerably more complex to interpret. The main conclusions will still be relevant to more general cases of random slope models. It should be noted that ignoring random slopes when they are present can lead to misleading conclusions (e.g., Baird & Maxwell, 2016).

Another extension of Eqs. (1) and (2) is multinomial logistic regression which allows for outcome variables with more than two levels (Hosmer et al., 2013, Chapter 8). These have been applied to corpus data (Granvik et al., 2014; Rosemeyer & Enrique-Arias, 2016), and an analysis of native and non-native English speakers' grammatical judgements (Godfroid et al., 2015). We also refer to Levshina (2015, Chapter 13) who described a corpus analysis of the use of "let", "allow", and "permit" in American English and used two separate (binary) logistic regression analyses rather than a single multinomial logistic regression. This is less coherent and statistically less efficient than the multinomial option, but it is a common (and sensible) strategy due to its easier interpretation—indeed, we adopt this approach in the case study in Section 4.2. Random effects can also be included in multinomial logistic regression.

2.2. Complete and quasi-complete separation

When applying logistic regression to the analysis of linguistics data, one problem that can often arise is that of complete or quasi-complete separation. Complete separation is the phenomenon where there exists a linear combination:

$z_i = x_{i1}a_1 + \ldots x_{ip}\,a_p$of the covariates such that $y_i = 1$ whenever $z_i > k$, and $y_i = 0$ whenever $z_i < k$, where $k, a_1, \ldots, a_p$ are constants Albert & Anderson, 1984). Quasi-complete separation is the case where $y_i = 1$ whenever $z_i \geq k$, and $y_i = 0$ whenever $z_i \leq k$. Collectively, these two cases are referred to as separation. In complete separation, the outcome variable is perfectly predicted by the covariates because the data can be completely divided by a linear function of the covariates with all positive outcomes on one side and all negative outcomes on the other. In both complete and quasi-complete separation, one or more estimated coefficients in equations ((2) or (3) tend toward (positive or negative) infinity. This results in the linear predictor itself being equal to (positive or negative) infinity, meaning the probability of a positive outcome is equal to 1 (or 0) for a subset of the data where the observed values of Y_i are all 1 (or all zero). Essentially, the set of covariates fit the data *too* well, such that we can achieve a perfect classification for a subset of binary outcomes. A simple example of this is when a particular covariate (say x_{li}) is a two-level factor variable (e.g., gender where $x_{li} = 0$ for male and $x_{li} = 1$ for female). Separation occurs when this factor is highly positively associated with a positive outcome, to the extent that the response Y_i is equal to 1 for every unit where x_{li} equals 1. In this case, the optimal estimator of the corresponding coefficient is positive infinity because this will give a perfect classification for those units where $x_{li} = 1$ while not affecting the fit for other units in the sample.

In practice, software that fits (multi-level) logistic regression models will not give infinite values. However, the offending coefficients will be estimated to be very large in magnitude while their standard errors are equally large or larger. Importantly, it means that the resulting p-values and confidence intervals produced by most software will not be useful for drawing valid conclusions. Indeed, when separation occurs, the corresponding coefficients are generally found to not be statistically significantly different from zero, even though they are very large, and it is intuitive that they are important and should be included. Note: it is normally obvious from the table of estimated coefficients when separation has occurred, but formal methods and software also exist for diagnosing this problem (e.g., Kosmidis & Schumacher, 2021).

Kimball et al. (2019) suggested that separation is especially likely to surface for certain types of linguistic data including (1) second language research where native populations and highly advanced learners often exhibit ceiling performance, (2) socio-linguistic studies where the likelihood of linguistic variables for some conditioning factors is low and the few tokens that do appear only occur with one variant, and (3) corpus data where linguistic forms may occur infrequently in some conditions of interest, among others. Regardless of the specific setting where it occurs, as discussed above, the standard results that arise from applying (multi-level) logistic regression may be misleading, and it is important to remedy this in some manner before drawing statistical conclusions.

2.3. Remedies for separation

A number of remedies for separation in logistic regression have been proposed in the literature. One of these remedies is to combine the categories of the offending explanatory factor or to drop the factor altogether. However, this is often one of the worst actions to take as it may remove the most interesting and informative variables from the model itself. In this article, we instead focus on one particular class of solutions based on penalised likelihood. Specifically, PLR augments the likelihood function of the standard logistic regression model with a penalty term that shrinks the magnitude of one or more of the coefficients β_k, for $k = 1, \dots, p$ toward zero. Model fitting then occurs by maximising the resulting penalised log-likelihood function. A number of penalties are available, with a well-established one being the Firth penalty (Firth, 1993; Heinze & Schemper, 2002; Kosmidis & Firth, 2009) which is designed to minimise large sample statistical bias and is implemented in the logistf (Heinz et al. 2020) package in the R statistical programming language (R Core Team 2020; see Levshina, 2015 for discussion of the use of R in linguistics). In this approach, the log of the penalised likelihood (denoted L_P) is:

$$log\left(L_P\right) = log(L) + \frac{1}{2} log\left\{ det\left(\sum_i p_i\left(1 - p_i\right) x_i x_i^T \right) \right\} \tag{4}$$

where L denotes the (unpenalised) likelihood, $p_i = P(Y_i = 1|x_i)$, and det(.) denotes the determinant of a matrix (this follows directly from Firth, 1993, Section 3.1 and Hosmer et al. 2013, p. 38). The second term becomes a large negative value when any fitted probabilities (p_i) are close to zero or 1, which in turn occurs if some coefficients are large in magnitude.

Other penalties are possible, with another common choice being a ridge penalty (le Cessie & Van Houwelingen, 1992) which penalises according to the sum of the squared magnitudes of the coefficients and the amount of penalisation/shrinkage overall is controlled by a single tuning parameter.

From the viewpoint of Bayesian inference, PLR is equivalent to assuming a (weakly) informative prior distribution for one or more of the coefficients, where the choice of a specific prior corresponds to a particular penalty function. Typical priors include the improper uniform distribution, the normal distribution, and the t-distribution (e.g., Gelman et al., 1995; Gelman et al., 2008). Gelman et al. (1995, Chapter 16) suggest using t-distributions with either 1 or 7 degrees of freedom and scale parameters of 2.5 where binary covariates are scaled to have a mean of zero and range of one. The argument given for a scale of 2.5 is that this is large enough to allow for the largest odds ratios likely to occur in most applications and small enough to sufficiently penalise large coefficient values, and hence, to avoid infinite parameter estimates. Other possible priors include the horseshoe, the regularized horseshoe, and discrete normal mixtures (e.g., van Erp et al., 2019). In the case of a ridge penalty, it is equivalent to assuming independent normal prior distributions for the coefficients where the variance (i.e., the square of the standard deviation) parameter is set to the inverse of the tuning parameter mentioned above. A general implementation of this is available via the R package blme (Bürkner 2017) and we refer interested readers to Gelman et al. (2008); Discacciati et al. (2015); Greenland (2003, 2007); Greenland and Mansournia (2015), and Sullivan and Greenland (2012), among others, for discussion of such weakly informative prior distributions. With Bayesian PLR, after fitting, the posterior distribution of the coefficients is obtained from which point estimates such as the posterior mode or mean can be obtained.

Another popular class of penalties to consider are those which induce sparsity and facilitate feature selection, including the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), the elastic-net penalty which combines the LASSO and ridge penalty (Zou & Hastie, 2005; van Erp et al., 2019), and many variations thereof depending on the precise data and problem (see Hui et al., 2017; Simon et al., 2013; Tibshirani et al., 2015 and references therein for some specific examples). While such techniques may also work for handling separation in logistic regression, in this article we choose not to focus on sparsity-inducing penalties, as variable selection is not our aim for penalisation here.

In the case of complete or quasi-complete separation arising in multi-level logistic regression models, the same remedy (i.e., multi-level PLR) can in principle be applied although there are some complications for practitioners. For example, to our knowledge, extensions of the Firth penalty above to multi-level logistic regression have not been implemented in any R packages, and it is unclear whether such penalties work in a way that systematically reduces the bias as intended (see also Abrahantes & Aerts, 2012, who suggested a two-step procedure for remedying separation in multi-level logistic regression using the Firth penalty, which works well empirically but is ad hoc). For this reason, we advocate instead for the use of the ridge penalty discussed above when it comes to multi-level PLR, and in both the simulation study (Section 3) and case studies (Section 4) we employ this approach for remedying separation in linguistics data, as available in the blme package. We take this approach for simplicity in order to demonstrate the usefulness of penalization and also because of availability of existing software for applying penalisation to single and multi-level logistic regression which are commonly used in linguistic analysis. It is certainly possible that other penalties may do better in particular cases. More importantly, our simulation suggests that results are not very sensitive to the tuning parameter, which in turn suggests that they will also not be overly sensitive to the precise choice of penalty or prior. See for example Sauter and Held (2016) for an implementation of Bayesian multi-level PLR using other weakly informative priors.

Finally, another approach worth mentioning here is that of penalised generalized estimating equation (GEE) as implemented in the R package geefirth (Mondol & Rahman, 2019). This method employs the Firth penalty for logistic regression but uses a GEE instead of maximum likelihood estimation technique for model fitting.

3. Simulation study

We performed a simulation study to investigate the performance of penalised multi-level logistic regression under various settings and differing degrees of separation, including when there is no separation. The design of the study was loosely motivated by the second language (L2) acquisition case study described in Section 4. Specifically, multi-level binary data were simulated with 4 observations for each of 150 units (individuals), leading to a total of 600 observations. Two ($p = 2$) independent binary covariates were generated: the first at the unit-level and the second at the observation-level, both with a 0.5 probability of equalling one. The binary response was then simulated from the random intercept logistic regression model in Eq. (3) where the unit-level random effect was generated from a normal distribution with a mean of 0 and standard deviation of 0.6. Selection of the value of 0.6 was motivated by the estimated random effects standard deviation in the SLA study that will be discussed in Section 4.1 (rounded up marginally to make clustering slightly more important in the simulation) which corresponds to an intra-cluster correlation of 0.10 (calculated using method D from Goldstein et al., 2002). The Supplementary Materials include results for two other values of the random effects standard deviation (given by halving and doubling the value of 0.6 assumed here). The coefficients of the covariates were set such that the exponentials of the coefficients (i.e., the odds ratios) equalled six possible combinations: 1 and 1 (i.e., no effects); 2.5 and 2.5 (moderate effects for both covariates); 25 and 25 (large effects for both covariates); 50 and 50; 50 and 1; and 1 and 50 (very large effects for one or both covariates). For each of these six scenarios, we simulated 1,000 datasets. The largest odds ratio of 50 was chosen such that complete separation occurred reasonably often (about 10% of the time when both odds ratios were 50).

For each simulated dataset, we considered eight variations of multi-level logistic regression. The first method is an unpenalised (i.e., standard) method fitted using maximum likelihood estimation via the R package glmmTMB. The remaining seven methods were based on multi-level PLR fitted using the blme package and employing a ridge penalty with various degrees of penalisation. Specifically, the ridge penalties were chosen to correspond to normal prior distributions with a mean of zero and standard deviation parameter set to one of {0.5, 1, 1.5, 2, 3, 4 and 5}. A smaller standard deviation implies a stronger amount of penalisation, that is, a greater degree of correction for potential separation in the data.

3.1. Simulation results

Fig. 1 presents comparative boxplots of the estimates of the first coefficient (corresponding to the person-level covariate) for each of the four scenarios. It should be noted that different scales are used for the different panels of the figure, due to the odds ratios being of different order of magnitude in different cases. The same information is also plotted on a common scale in the Supplementary Materials.

When both odds ratios were equal to 1, all eight methods performed similarly to one another. All of the boxplots centred around the true value and the variability was very similar, except in the most severe case of penalisation (i.e., standard deviation equal to 0.5). When the odds ratios were both 2.5, all the methods performed fairly similarly, except for penalisation based on prior standard deviation of 0.5 which clearly resulted in negatively biased estimators. It is apparent that this setting penalises large coefficient values too strongly.

When both odds ratios were 25, we started to see empirical evidence of separation occurring as exemplified through some very large estimated odds ratios produced by maximum likelihood estimation (labelled as a penalty equal to "none"). (Separation was deemed to occur when one or more estimated coefficients was greater than 10 in magnitude on the log-odds scale. This simple rule worked well for this simulation where the model was relatively simple, containing only two covariates. In more complex scenarios, software can be used; see for example Kosmidis & Schumacher, 2021.) When both odds ratios were 50, these were more common (approximately 10% of simulated datasets). When the first odds ratio was 50 and the second was 1, there was only occasional complete separation. In all three of these subplots, we see that penalised likelihood is over-penalising when the standard deviation parameter is 0.5 or 1 (and to a lesser extent, 1.5). Larger standard deviation parameters (2, 3, 4 and 5) all exhibited similar results—they handle complete separation well with negligible bias. When the first odds ratio was 1 and the second was 50, separation did not occur relative to the first coefficient.

The simulation results for the second coefficient are presented in Figure S1 in the Supplementary Materials and offer similar conclusions. The Supplementary Materials also show more detailed results relating to confidence interval coverage and widths along with the proportion of simulations where separation occurred. Overall, the simulation study suggests that penalisation works well to handle complete separation in multi-level logistic regression, and results are not overly sensitive to the amount of penalisation. Good results are achieved by using ridge penalties based on standard deviation parameters of 2, 3, 4 and 5. This parameter is set to 3 in the case studies in the next section which is broadly consistent with the advice of Gelman et al. (1995, Chapter 16) since the t-distribution with 7 degrees of freedom and scale parameter 2.5 has a standard deviation of approximately 3. Moreover, a value of 3 gave a reasonable trade-off between confidence interval width and coverage rate (see supplementary Tables S1 and S2).

4. Applications of penalised logistic regression

4.1. Case Study I: Japanese SLA survey

We illustrate the use of penalised logistic regression to remedy separation of binary outcomes in two linguistic case studies. The first of these comes from Woods (2021) who aimed to examine how L2 learners of Japanese chose when to use the particles *wa* and *ga* under the influence of various factors that reside at different linguistic domains and the interactions between these factors.

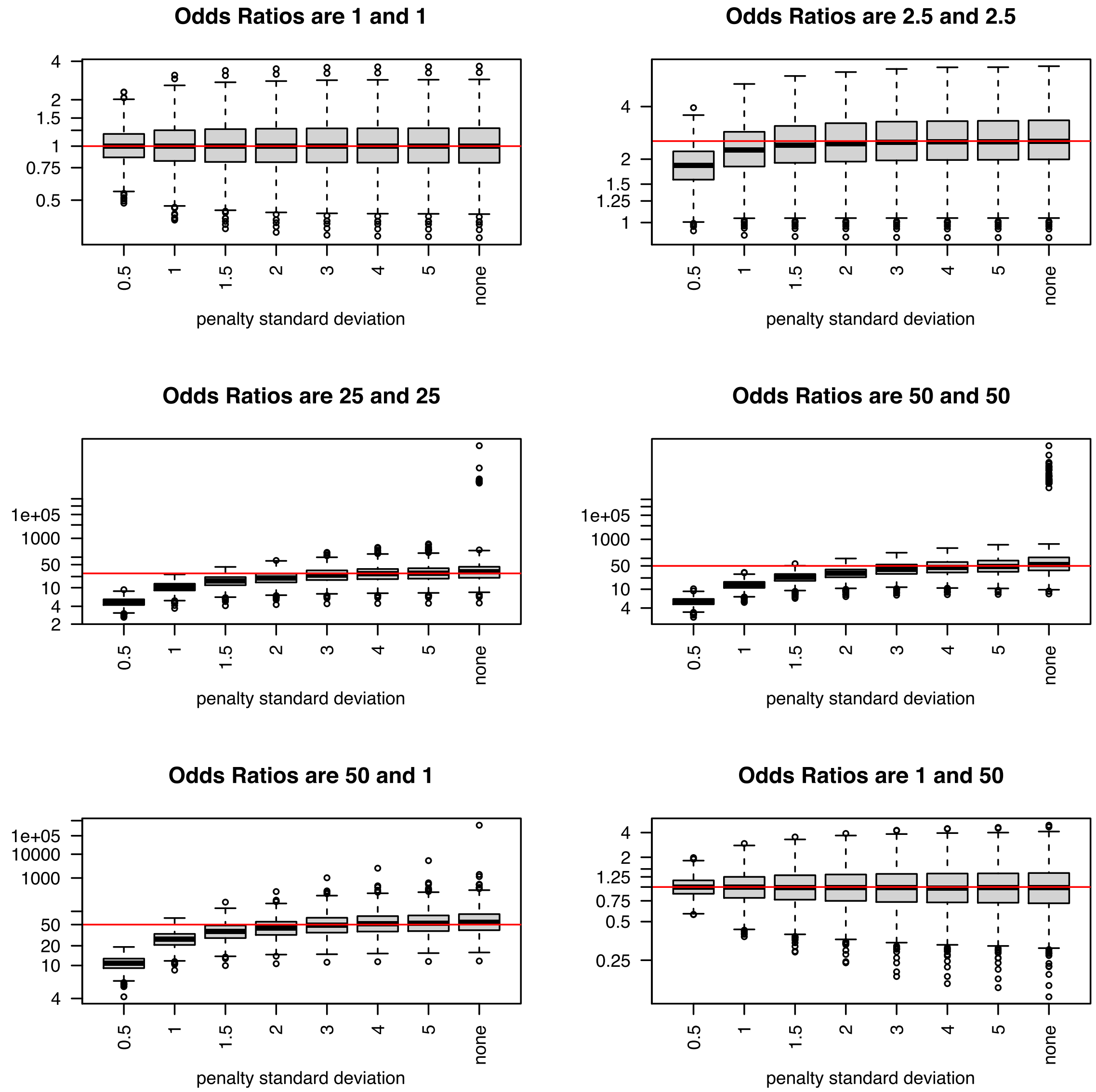

Fig. 1. Distributions of the estimates for the first odds ratio (person-level covariate) for six different true models.
Note: Estimators are either based on PLR with a ridge penalty chosen to correspond to normal prior distributions with a mean of zero and standard deviation parameters set to one of {0.5, 1, 1.5, 2, 3, 4 and 5} or are based on maximum likelihood estimation with no penalty (labelled as "none"). The red horizontal line indicates the true value of the odds ratio (i.e., the value from which the data were simulated). Note the scaling of the y-axis to enable plotting of the extremely large odds ratio estimates occurring under complete separation.

The study also addressed the question of whether the ability to choose appropriately between *wa* and *ga* can be fully acquired by highly advanced L2 learners, and thereby contributes Japanese language examples to the ongoing debate of the Interface Hypothesis (Sorace, 2005; Montrul, 2011).

Ga is a case marker that marks the nominative case, while *wa* is a discourse marker that marks the topic of a sentence (e.g., Kuno, 1975; Noda, 1996). As such, these particles reside on different linguistic levels. However, as the nominative case is frequently topicalised in a sentence, Japanese speakers often need to make a choice between these particles. It is predicted that the choice becomes particularly problematic when *ga* sentences are not overtly distinguished from *wa* sentences in the speaker's first language, such as in English. The difficulty in L2 acquisition of these particles is predicted to stem from their characteristics as an interface property (i.e., a syntactic representation that is governed by different domains of information; Sorace, 2005), and we refer to Woods (2021, and references therein) for details about the existing literature and theory regarding such topics.

We applied multi-level logistic regression, provided in Eq. (3), to model how the probability of answering each question appropriately (choice between *wa* and *ga*) depends on the sentence context (topic or exhaustive listing) and on the speaker's acquisition status (L1, L2a and L2b representing native speakers, advanced L2 speakers and beginner L2 speakers, respectively). We included

Table 1
Results from applying standard unpenalised versus penalised multi-level logistic regression to the Japanese SLA survey data.

Effect	Coefficients (standard errors)	
	Unpenalised	Penalised
Intercept	1.53 (0.28)**	1.52 (0.27)**
Topic (*wa*)	20.53 (>1000)	3.33 (0.93)**
Exhaustive (*ga*): L2b	-1.75 (0.33)**	-1.73 (0.32)**
Topic (*wa*): L2b	-19.95 (>1000)	-2.70 (0.93)**
Exhaustive (*ga*): L2a	-1.17 (0.33)**	-1.15 (0.33) **
Topic (*wa*): L2a	-19.54 (>1000)	-2.28 (0.95)*

Note. Model is of Appropriate Choice of *wa* or *ga*, regressed against Acquisition Status (L1, L2a, and L2b) and Context (topic or exhaustive listing, or *wa* and *ga*, respectively, in Japanese). Estimates of the coefficients, with corresponding standard errors in parentheses, are shown. Statistical significance at the 5% or 1% level is indicated by one (*) or two (**) asterisks, respectively.

Table 2
Tests of differences between L2a and L2b speakers within context.

Contrast	Coefficients (standard errors)	
	Unpenalised	Penalised
L2a-L2b for Exhaustive (*ga*)	0.58 (0.25)*	0.58 (0.26)*
L2a-L2b for Topic (*wa*)	0.41 (0.43)	0.42 (0.42)

Note. Tests are based on applying standard unpenalised and penalised multi-level logistic regressions to the Japanese SLA survey data. In each case, the difference in coefficients is shown with the standard error of the difference in parentheses. Positive values indicate higher probabilities of answering appropriately for L2a speakers compared to L2b speakers. Statistical significance at the 5% or 1% level is indicated by one or two asterisks, respectively.

person-level random intercepts to reflect individual variation in Japanese language use and proficiency. As in the simulation study, unpenalised (i.e., standard multi-level) logistic regression was fitted using the glmmTMB package while multi-level PLR models were fitted via the blme package. For the latter, we used a ridge penalty that corresponded to a normal prior distribution with a mean of zero and standard deviation parameter equal to 3 for all coefficients in the model.

We set up all the logistic regressions so that Acquisition Status had a baseline category of L1 and was nested within Context. This was done to enable L2a vs L1 and L2b vs L1 comparisons within each Context to be obtained directly from the model output. We also made comparisons between L2a and L2b speakers within Context, using the emmeans package (Lenth, 2021).

Results from the unpenalised model clearly show that complete separation has occurred with very large coefficient estimates for three of the five covariates in the model (see Table 1). The corresponding standard errors are even greater in magnitude and, as such, the corresponding hypothesis tests are invalid in these cases. By contrast, the penalised model gives finite (though still large) estimates, along with standard errors, that facilitate proper statistical inferences to be made. Specifically, the multi-level PLRs show clear statistical evidence that both L2a and L2b speakers were less likely to answer appropriately compared to L1 speakers in both contexts. The unpenalised logistic regression fails to identify statistical significance for two of these four comparisons due to the complete separation problem.

There was statistically clear evidence that L2a speakers were more accurate than L2b speakers in exhaustive (*ga*) contexts (see Table 2), while no difference was found between the same set of groups of speakers in the topic (*wa*) contexts. This perhaps reflected the relative lack of data available for this context (and hence higher standard errors). The L2a versus L2b comparisons were not affected by complete separation and, consequently, the penalised and unpenalised results were very similar. The standard error of the person-level random effect was estimated to be 0.52 in the unpenalized model fit and 0.55 in the penalized fit.

4.2. Case Study II: Corpus analysis

The second case study we examined comes from Lozano (2016) who investigated a series of research questions with the one we focus on being "How can we account for the high production of noun phrases (NP) anaphors in topic-shift contexts?" (p. 243). To address this question, corpus data were obtained from *Corpus Escrito del Español L2 (CEDEL2)* which consisted of an L1 English/L2 Spanish written learner corpus and a comparable Spanish native corpus of nearly 800,000 words from over 1,500 participants. The data were collected online. Proficiency levels were measured by the standard University of Wisconsin Placement Test and the learners' self-proficiency ratings on a 6-point scale for each of the four language skills. Attributes of the learners including gender and age, along with task variables such as composition title, were also recorded.

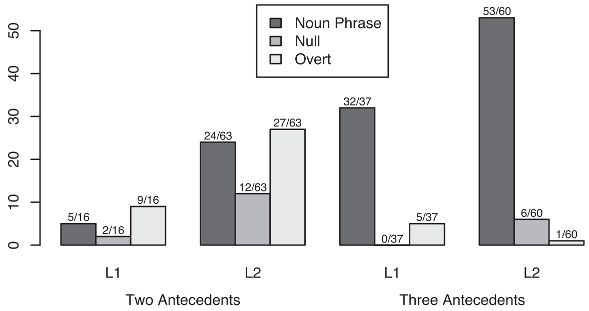

Fig. 2. Summary of data from Lozano (2016) Showing anaphor production by L1/L2 speakers and number of antecedents in topic shift contexts.

Table 3
Logistic regression results for four modelling approaches and three binary dependent variables.

Dependent Variable	Effect	Coefficients (standard errors)	
		Unpenalised	Penalised
Noun Phrase	Intercept	-0.90 (0.64)	-0.74 (0.59)
	Antecedents3	2.90 (0.83)**	2.66 (0.75)**
	L2:antecedents2	0.55 (0.75)	0.38 (0.69)
	L2:antecedents3	0.23 (0.74)	0.28 (0.71)
Null	Intercept	-1.95 (0.76)**	-1.94 (0.70)**
	Antecedents3	-20.63 (>1000)	-2.36 (1.33)
	L2:antecedents2	0.50 (0.82)	0.49 (0.76)
	L2:antecedents3	20.38 (>1000)	2.07 (1.25)
Overt	Intercept	0.27 (0.62)	0.15 (0.58)
	Antecedents3	-2.27 (0.78)**	-2.15 (0.74)**
	L2:antecedents2	-0.87 (0.75)	-0.71 (0.70)
	L2:antecedents3	-2.23 (1.20)	-2.00 (1.05)

Note: Multi-level logistic regression models with random effects for the speaker were fitted using both unpenalised maximum likelihood estimation and PLR. In each case, coefficients (standard errors) are shown. Significance at the 5% or 1% level is indicated by one or two asterisks, respectively.

Lozano (2016) extracted 20 texts produced by very advanced L1 English/L2 Spanish learners and Spanish native speakers. Only third-person human (singular/plural) subjects were analysed, with a total of 498 subjects tagged for 11 properties. Each tagged subject contained a minimum of 14 and a maximum of 20 terminal tags. The total number of terminal tags was 8,850.

Among other topics, Lozano (2016) examined how the choice of anaphor (from NP, null and overt) depends on the number of antecedents and how this differs between L1 and L2 speakers. It would be expected that when there are more potential antecedents, a noun phrase would be commonly used in order to avoid possible ambiguity. When there are few potential antecedents, an overt or perhaps null anaphor would be more likely. Data on this choice are summarised in Lozano (2016, Figure 10), which we reproduce below in a slightly different form in Fig. 2, emphasising that anaphor choice is a function of the number of antecedents and L1/L2 status. Only sentences with two or three antecedents were included as anaphor choice was much more straightforward when there was only one antecedent.

Lozano (2016) noted that both groups of speakers mostly produce a NP when there are three antecedents, which is appropriate to avoid ambiguity about the subject. This is the case for both L1 and L2 speakers, but particularly for the latter. L2 speakers sometimes even produce a null-gender pronoun in this case, which is likely to be infelicitous. When there are only two antecedents, the subject is less ambiguous and a pronoun (either overt or null) is more common, particularly for L2 speakers. It is unclear whether Lozano formally tested the above hypotheses although hypotheses were tested elsewhere using chi-squared tests of contingency tables, assuming independence across all tokens.

We applied random effects logistic regression with and without penalisation to re-examine the above data to deal with separation in the binary outcome. Unpenalised multi-level logistic regression was applied as defined by Eq. (3) and fitted using the glmmTMB package with person included as a random effect. Multi-level PLR was fitted using the blme package, with a ridge penalty that corresponds to a normal prior distribution with a mean of zero and standard deviation parameter equal to 3 for all coefficients. For both modelling approaches, we fit three separate logistic regressions where the binary outcomes were NP (versus the other two types of anaphor), overt pronoun (versus the other two), and null pronoun (versus the other two); this is similar to Levshina (2015, Chapter 13).

Table 3 shows that complete separation occurred for the null dependent variable which is unsurprising given the zero count that can be seen in Fig. 1. The coefficient estimates of antecedents3, L2:antecedents3, and their respective standard errors were effectively infinite in the unpenalised model. As such, as discussed previously, inferences such as hypothesis testing and confidence intervals produced from these values would be inappropriate. By contrast, PLR provides meaningful information—the coefficient of antecedent3 is substantially negative (below -2 is substantial, as it implies an odds ratio less than exp(-2) or approximately 0.14) while the coefficient of L2:antecedent3 is substantially positive. Although these results are not statistically significant, they nevertheless suggest that for L1 speakers are very unlikely to produce null pronouns when there are three antecedents, while this becomes much more likely for L2 speakers. Note, these qualitative findings are also visible in Fig. 2, but PLR quantifies them in a principled manner and provides further evidence that the observed difference between L1 and L2 is not statistically significant.

5. Conclusion

Logistic regression is a powerful and widely applicable method for analysing binary outcomes in linguistics with applications ranging from corpus data and socio-linguistic surveys to second language acquisition studies. A complication that can arise from applying logistic regression is complete or quasi-complete separation which can occur when powerful explanatory variables lead to effectively infinite parameter estimates, rendering standard analysis and hypothesis testing invalid. The naive but arguably common solution of removing and grouping these covariates is often poor statistical practice because it effectively eliminates the most interesting and important components of the data. In this article, we have demonstrated how penalised logistic regression, potentially with random effects included to account for clustering, overcomes this separation. Although it requires the setting of tuning parameters or prior distributions controlling the level of penalisation, which may disincentivise linguists from its use, we have shown via simulation that, at least with a simple ridge penalty (corresponding to a normal prior distribution), results are not overly sensitive to this choice. Through two case studies on second language acquisition and corpus datasets, we demonstrate that useful conclusions from PLR can be extracted while standard unpenalised methods may lead to misleading and inaccurate conclusions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

R code to reproduce the simulation study in Section 3 is included as supplementary material.

Acknowledgements

We thank Associate Professor Cristóbal Lozano for generously providing the tagged corpus data which we re-analysed in Section 4.2.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2023.100044.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*(1), 1–10.

Abrahantes, J. C., & Aerts, M. (2012). A solution to separation for clustered binary data. *Statistical Modelling, 12*(1), 3–27.

Baird, R., & Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods, 21*(2), 175–188. 10.1037/met0000070.

Cheung, L., & Zhang, L. (2016). Determinants of the synthetic–Analytic variation across English comparatives and superlatives. *English Language and Linguistics, 20*(3), 559–583. 10.1017/S1360674316000368.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. 10.1017/9781316410899.

le Cessie, S, & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 41*(1), 191–201.

De Cuypere, L., Baten, K., & Rawoens, G. (2014). A corpus-based analysis of the Swedish passive alternation. *Nordic Journal of Linguistics, 37*(2), 199–223.

Discacciati, A., Orsini, N., & Greenland, S. (2015). Approximate Bayesian logistic regression via penalized likelihood by data augmentation. *The Stata Journal, 15*(3), 712–736.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*(1), 27–38.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman and Hall/CRC.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360–1383.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*(4), 223–231. 10.1207/S15328031US0104_02.

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition, 37*(2), 269–297.

Granvik, A., & Taimitarha, S. (2014). Topic-marking prepositions in Swedish: A corpus-based analysis of adpositional synonymy. *Nordic Journal of Linguistics, 37*(2), 257–296.

Greenland, S. (2003). Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics, 59*(1), 92–99.

Greenland, S. (2007). Bayesian perspectives for epidemiological research. *International Journal of Epidemiology, 36*(1), 195–202.

Greenland, S., & Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine, 34*(23), 3133–3143.

Heinz, G., Ploner, M., & Jiricka, L. (2020). *logistf: Firth's bias-reduced logistic regression* R package version 1.2.4. https://CRAN.R-project.org/package=logistf .

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine, 21*(16), 2409–2419.

Hinrichs, L., & Szmrecsanyi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language & Linguistics, 11*(3), 437–474.

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken: John Wiley & Sons. 10.1002/9781118548387.

Hui, F. K. C., Müller, S., & Welsh, A. H. (2017). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica, 27*(2), 501–518. 10.5705/ss.202015.0329.

Johnson, K. (2008). *Quantitative methods in linguistics*. Malden: Blackwell.

Kimball, A. E., Shantz, K., Eager, C., & Roy, J. (2019). Confronting quasi-separation in logistic mixed effects for linguistic data: A Bayesian approach. *Journal of Quantitative Linguistics, 26*(3), 231–255. 10.1080/09296174.2018.1499457.

Kosmidis, I., & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika, 96*(4), 793–804.

Kosmidis, I., & Schumacher, D. (2021). *detectseparation: Detect and check for separation and infinite maximum likelihood estimates* R package version 0.2. https://CRAN.R-project.org/package=detectseparation .

Kuno, Susumu. (1975). *The structure of the Japanese language*. Cambridge: MIT Press.

Lenth, R. V. (2021). *emmeans: Estimated marginal means aka least squares means* R package version 1.5.5-1. https://CRAN.R-project.org/package=emmeans .

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Lozano, C. (2016). Pragmatic principles in anaphora resolution at the syntax-discourse interface: Advanced English learners in the CEDEL2 corpus. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 235–266). John Benjamins Publishing Company.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear and mixed models* (2nd ed.). Hoboken: Wiley.

Mondol, M. H., & Rahman, M. S. (2019). Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data. *Statistics in Medicine, 38*(14), 2544–2560.

Montrul, S. (2011). Multiple interfaces and incomplete acquisition. *Lingua, 121*(4), 591–604. 10.1016/j.lingua.2010.05.006.

Noda, H. (1996). *Wa to Ga (Wa and Ga),* Shin Nihongo Bunpoo Sensho. Tokyo: Kurosio Publishers.

Peters, S., Wilson, K., Boiteau, T. W., Gelormini-Lezama, C., & Almor, A. (2016). Do you hear it now? A native advantage for sarcasm processing. *Bilingualism: Language and Cognition, 19*(2), 400–414.

R Core Team. (2020). *R: A language and environment for statistical computing* R Core Team (2020). Vienna, Austria: R Foundation for Statistical Computing https://www.R-project.org/.

Rosemeyer, M., & Enrique-Arias, A. (2016). A match made in heaven: using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change, 28*(3), 307–334.

Sauter, R., & Held, L. (2016). Quasi-complete separation in random effects of binary response mixed models. *Journal of Statistical Computation and Simulation, 86*(14), 2781–2796.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics, 22*(2), 231–245. 10.1080/10618600.2012.681250.

Sorace, A. (2005). Selective optionality in language development. In Leonie Cornips, & Karen P Corrigan (Eds.), *Syntax and variation: Reconciling the biological and the social* (pp. 55–80). Cambridge: Cambridge University Press.

Sullivan, S. G., & Greenland, S. (2012). Bayesian regression in SAS software. *International Journal of Epidemiology, 42*(1), 308–317.

Tibshirani, R. (1996). Regression Shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*, 267–288. 10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R., Hastie, T., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. United Kingdom: Taylor & Francis. 10.1201/b18401.

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology, 89*, 31–50.

Woods, H. (2021). *L2 acquisition of interface properties: Case of Japanese WA and GA*. Australian National University PhD Thesis. 10.25911/5QZ2-6D61.

Zeng, G., & Zeng, E. (2019). On the relationship between multicollinearity and separation in logistic regression. *Communications in Statistics - Simulation and Computation, 50*(7), 1989–1997.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320. 10.1111/j.1467-9868.2005.00503.x.