On the Hauck–Donner Effect in Wald Tests: Detection, Tipping Points, and Parameter Space Characterization

Thomas W. Yee

Department of Statistics, University of Auckland, Auckland, New Zealand

ABSTRACT
The Wald test remains ubiquitous in statistical practice despite shortcomings such as its inaccuracy in small samples and lack of invariance under reparameterization. This article develops on another but lesser-known shortcoming called the Hauck–Donner effect (HDE) whereby a Wald test statistic is no longer monotone increasing as a function of increasing distance between the parameter estimate and the null value. Resulting in an upward biased $p$-value and loss of power, the aberration can lead to very damaging consequences such as in variable selection. The HDE afflicts many types of regression models and corresponds to estimates near the boundary of the parameter space. This article presents several new results, and its main contributions are to (i) propose a very general test for detecting the HDE in the class of vector generalized linear models (VGLMs), regardless of the underlying cause; (ii) fundamentally characterize the HDE by pairwise ratios of Wald and Rao score and likelihood ratio test statistics for 1-parameter distributions with large samples; (iii) show that the parameter space may be partitioned into an interior encased by at least 5 HDE severity measures (faint, weak, moderate, strong, extreme); (iv) prove that a necessary condition for the HDE in a 2 by 2 table is a log odds ratio of at least 2; (v) give some practical guidelines about HDE-free hypothesis testing. Overall, practical post-fit tests can now be conducted potentially to any model estimated by iteratively reweighted least squares, especially the GLM and VGLM classes, the latter which encompasses many popular regression models.

1. Introduction
In classical likelihood theory three test statistics are used for general hypothesis testing and inference. They are the likelihood ratio test (LRT), Rao’s score (Lagrange multiplier) test, and the Wald test. It is well known the LRT generally performs the best, and that the Wald test suffers from shortcomings such as its lack of invariance under reparameterization and inaccuracy in small samples. Despite these, the Wald test is probably the most widespread test, as output such as Table 3 is ubiquitous in statistical practice.

A lesser known but no less pernicious problem of the Wald test is that it suffers potentially from the Hauck–Donner effect (HDE; Hauck and Donner 1977, 1980) whereby the test statistic fails to increase monotonically as a function of its distance from the null value, for example, as in Figure 1. Consequently, a truly large effect contaminated by the HDE might be construed as being nonsignificant. The loss of power problem is aggravated by variable selection procedures that are based on Wald’s test.

In general likelihood theory, much of statistical inference is dichotomized into the cases where the true parameter value is either in the interior of parameter space $\Theta$ or lies on its boundary. This article is concerned with the interface of the two. Informally, partition the closure $\Theta = \Theta^o \cup \partial \Theta$ into its interior and boundary, and further partition $\Theta^o = \Theta^{oo} \cup \Theta^{hde}$ where $\Theta^{hde}$ may be the empty set. The usual regularity conditions fully operate practically in $\Theta^{oo}$ but start breaking down in $\Theta^{hde}$ for want of sufficient Taylor series terms, which we define as the subspace where an aberration of the Wald statistic is more full-blown. Failure to distinguish between $\Theta^{oo}$ and $\Theta^{hde}$ can result in incorrect inferences, therefore, special attention should be made to identify it if it occurs. This article sheds light on $\Theta^{hde}$ (Figure 5 where $\Theta^{oo}$ is at most weak HDE and $\Theta^{hde}$ is at least moderate HDE).

One reason for choosing and defining $\Theta^{hde}$ as such is the huge popularity of Wald tests in general regression modeling, including generalized linear models (GLMs; Nelder and Wedderburn 1972). They appear almost universally as standard computer output in the form of a 4-column matrix (called a Wald table in this article) consisting of the point estimates, standard errors, Wald statistics, and $p$-values, with optional embellishments of asterisks and dots.

Despite four decades since it was first observed, there has been relatively very little general work characterizing the Hauck–Donner phenomenon, or even to detect it. This work is an attempt to address these. We use “general” because special cases such as $N\mu (1 - \mu) > 10$ for the (normal approximation to the) binomial are very well-known special cases only. An empirical approach is taken in this article and we develop new methods that can be used routinely on any GLM or potentially any model based on a weighted crossproduct matrix of the form $X^TWX$ ($= A$, say) within an iteratively reweighted least squares (IRLS) algorithm. Consequently the results are very general and widely applicable for HDE detection.
Væth (1985) studied the HDE but mainly restricted his investigation to the one-sample problem for one-parameter exponential families. His results were heavily dependent on the nice mathematical properties of exponential families and it was stated that these do not easily generalize to arbitrary probability models due to certain boundary problems, and the Wald test statistic should be used cautiously with discrete distributions. He also advised that Wald’s test should not be used cautiously in logistic regressions with many explanatory variables. This caution has been repeated by others, for example, Xing et al. (2012) in the context of genome-wide case-control associations.

It is stressed here, as does Væth (1985) for one-parameter exponential families, that it is the behavior of the Wald statistic for $\beta_0$, for fixed sample size $n$, as the maximum likelihood estimate (MLE) moves away from the null value that is of concern and not the distributional properties of $W$ as $n \to \infty$. Here, and sometimes for simplicity, we let the coefficients of a regression model such as a GLM be denoted $\beta_s$, $s = 1, 2, \ldots$. This article concerns simple null hypotheses of the form $H_0: \beta_1 = \beta_0$ for some known prespecified $\beta_0$ that is usually taken to be 0; the signed square root of the Wald statistic for the $s$th coefficient is $\sqrt{W_s} = (\hat{\beta}_s - \beta_0)/SE(\hat{\beta}_s)$ where the Wald statistic $W_s = \hat{W}_s^2$ is asymptotically $\chi^2_1$ under $H_0$. Note that while the denominator of $\hat{W}_s$ is sometimes evaluated at $\beta_0$ (in which case the HDE will be absent), the vast majority of software such as glm() in R evaluates the standard error SE at the MLE so that the HDE is an ever-present threat.

Notationally, let the $(j, s)$ element of the inverse of a matrix $A$ having $(j, s)$ element $a_{js}$ be denoted by $a^{js}$, and $A \succ 0$ and $A \succeq 0$ indicates that the symmetric matrix $A$ is positive-definite and positive semidefinite, respectively. The Hadamard (element-by-element) and Kronecker products of two matrices $A$ and $B$ are denoted by $A \odot B = [(a_{js} \cdot b_{js})]$ and $A \otimes B = [(a_{js} \cdot B)]$, respectively. If $\theta$ and $\eta$ are $M$-vectors then $\partial \theta / \partial \eta^T$ is an order-$M$ matrix whereas $\partial \theta / \partial \eta_j$ is an $M$-vector with elements $\partial \theta_j / \partial \eta_i$.

We also let $e_s$ be a vector of 0s except for a 1 in the $j$th position, whose dimension is obvious, and the indicator function $I_{[=j]}$ equals 1 if $i = j$ and 0 otherwise. We use $W_L = 2(\ell - \ell_0)$ to denote the LRT statistic.

An outline of the article is as follows. As a concrete example, we consider the original dataset of Hauck and Donner (1977) in Section 2 (as well as citing more instances of the HDE by others for additional motivation) before describing elements of a class of models called VGLMs—the detection test specifically applies to this (large) class. The general test is described for VGLMs in Section 3, and supporting asymptotic results are given in Section 4. Section 5 proposes two important refinements: finite-difference approximations for derivatives to make HDE detection applicable to all VGLMs, and categorizing the HDE into several severity measures that form a partition of $\Theta$. Some numerical examples are given in Section 6, and Section 7 discusses computational details and provides some practical guidelines. The article concludes with a discussion of the overall findings. The methodology here is implemented in the R package VGAM 1.1-2 or later (available on CRAN) for 100+ models. Some extensions for detecting the HDE in related contexts are given in Appendices A2–A5 in the supplementary materials.

2. The Hauck and Donner Dataset

Hauck and Donner (1977) applied logistic regression to a $2 \times 2$ table of counts to illustrate the HDE. Because of its simplicity a more complete analysis of the behavior of $\hat{W}_2$ is permitted here. We will generalize the data slightly to afford a little more flexibility (Table 1(a)). The actual data has $N = 100, n = 2N$, $R_0 = 25$, and we will generally keep these fixed but vary $R \in \{1, \ldots, N - 1\}$. The sample proportions are $\pi_0 = R_0/N = 0.25$ at $x_2 = 0$, and $\pi_1 = R/N$ at $x_2 = 1$. Fit

$$\logit \mu_i = \beta_1 + \beta_2 x_2, \quad i = 1, \ldots, n,$$

independently, (1) so that $\hat{\beta}_1 = \logit \hat{\pi}_0 \approx -1.099$ and $\hat{\beta}_2 = \log \hat{\psi}$ is the estimated log odds ratio. The HDE is defined formally as being present when the Wald statistic points downward. The HDE becomes
pronounced once past a certain threshold as \( R \to 0 \) or \( R \to N \), corresponding to sparsity in one off-diagonal cell. In Figure 2 this corresponds to \( \hat{\pi}_1 < 0.03 \) and \( \hat{\pi}_1 > 0.91 \) approximately because, for example, the larger threshold corresponds to \( R \approx 91 \). For such count tables the well-known formula

\[
\text{SE}(\log \hat{\psi}) \approx \sqrt{\frac{1}{N - R_0} + \frac{1}{R_0} + \frac{1}{N - R} + \frac{1}{R}}
\]  

(2)

is routinely used. For handling low counts, there are several popular pre-fit improvisions, for example, adding 0.5 to each cell count in the SE calculation give a bias-corrected log odds ratio, and applying the Mantel–Haenszel method. However, in the context of routine logistic regressions, it is far more convenient for some post-fit test for HDE to be applied to a fitted model—if the HDE is absent then we can conclude that the counts are sufficiently large that no adverse behavior on the SE was seen. The aberrant behavior of the Wald statistic is not well known in general by practitioners and the HDE seldom mentioned in applied textbooks for statisticians.

In this particular example the HDE can be explained by a situation already well-known to practitioners. This is not the case in general because the HDE has been observed in other regression models by various authors since. Some examples include Storer, Wacholder, and Breslow (1983) in conditional logistic regression with matched and stratified samples, Væth (1985) in one-sample problems for one-parameter exponential families and GLMs, Nelson and Savin (1990) in Tobit and nonlinear regression models, Fears, Benichou, and Gail (1996) in a balanced 1-way random effects ANOVA design, Therneau and Grambsch (2000, p. 60) in Cox proportional hazards models, and Kosmidis (2014) in cumulative link models. In general the Wald test can be expected to be valid only if a normal likelihood can be used to approximate the profile likelihood for the parameter well (Meeker and Escobar 1995) and the observed value of the sufficient statistic is away from \( \hat{\psi} \).

Regardless of its underlying cause, it would be very useful to have a procedure for detecting the HDE that could be routinely applied to a given regression model. To this end, Section 3 proposes a method for this, and the method applies very generally to models estimated by IRLS. This general purpose algorithm has been used to fit many popular regression models, and most notable is the GLM class, of which a multivariate extension called vector GLMs (VGLMs; Yee 2015) have been proposed.
Continuing with this example, when the full model is fitted by IRLS, then
\[
A^{-1} = \frac{1}{N} \times \begin{pmatrix}
\hat{\pi}_0^{-1}(1 - \hat{\pi}_0)^{-1} & -\hat{\pi}_0^{-1}(1 - \hat{\pi}_0)^{-1} \\
-\hat{\pi}_0^{-1}(1 - \hat{\pi}_0)^{-1} & \hat{\pi}_0^{-1}(1 - \hat{\pi}_0)^{-1} + \hat{\pi}_1^{-1}(1 - \hat{\pi}_1)^{-1}
\end{pmatrix}.
\]
Consequently, (2) is obtained, as is
\[
\frac{\partial^2 W_L}{\partial \beta_2 \partial \beta_2} = \frac{1}{SE(\hat{\beta}_2)} \left\{ \frac{1}{N} \left[ -\hat{\pi}_1(1 - \hat{\pi}_1) \frac{SE^2(\hat{\beta}_2)}{SE(\hat{\beta}_2)} \right] \right\}
\]
which is plotted in Figure 2(b). It appears impossible to write a closed form expression for \( R \) upon setting (4) to 0; this indicates that it is impractical to determine beforehand the threshold of how low the counts can be before observing the HDE—and likewise for more complicated situations. This supports the view that a detection test post-fit is more practical than trying to avoid it pre-fit.

Incidentally, the LRT cannot exhibit the HDE
\[
\frac{\partial^2 W_L}{\partial R^2} = 2 \left( \frac{1}{R} - \frac{1}{R + R_0} \right) + 2 \left( \frac{1}{N - R} - \frac{1}{(N - R) + N - R_0} \right) > 0
\]
for all \( R \in \{1, \ldots, N-1\} \), provided that \( R_0 > 0 \), therefore \( W_L \) is convex in \( \hat{\beta}_2 \). It will also be seen that the Rao score (Lagrange multiplier) test is also immune to the HDE as it does not depend on the third derivatives of the log-likelihood.

Somewhat similar to the Hauck and Donner dataset, it is noted that the HDE can arise as a result of data exhibiting near overlap, quasi-complete separation or complete separation (see, e.g., Albert and Anderson 1984; Lesaffre and Albert 1989). For example, starting off with a dataset comprising \((x_{i2} = (i - 1)/(2N - 2), y_i = 0), i = 1, \ldots, 2N - 1\), plus \((x_{i2} = \frac{1}{2}, y_i = 1)\), we replace each successive \( y_i = 0 \) on the RHS of \( x_{i2} > \frac{1}{2} \) by \( y_i = 1 \) (Figure 3) except for the very rightmost. As the number of \( y_i = 1 \) increases the logistic regression (1) fitted to these data has \( \beta_2 \) increasing and the HDE will become present eventually. Figure 3(b) shows the Wald statistic as a function of \( \hat{\beta}_2 \), and the HDE is evident at the RHS. Data separation is likely in Big Data situations when extraneous covariates are included in regression models of high dimensionality.

2.1. VGLMs

The detection test specifically applies to this (large) class, therefore we briefly summarize them. They can loosely be thought of as multivariate GLMs applied to parameters \( \theta \) but extending far beyond the exponential family; more details can be found in Yee (2015). The data are \((x_i, y_i), i = 1, \ldots, n\), independently, with response \( y_i \) and explanatory variables \( x_i \) usually with an intercept \( x_{i1} = 1 \). The \( j \)th linear predictor
\[
g_j(\theta_j) = \eta_j = \beta_j^T x = \sum_{k=1}^d \beta_{(j)k} x_{ik}, \quad j = 1, \ldots, M,
\]
for some parameter link function \( g_j \) satisfying the usual properties (strict monotonicity and twice-differentiable). If \( M > 1 \) then linear constraints between the regression coefficients are accommodated, as
\[
\eta(x_i) = \begin{pmatrix}
\eta_1(x_i) \\
\vdots \\
\eta_M(x_i)
\end{pmatrix} = \sum_{k=1}^d \beta_{(k)k} x_{ik} = \sum_{k=1}^d \text{H}_k \beta_{(k)k} x_{ik},
\]
for known constraint matrices \( \text{H}_k \) of full-column-rank (i.e., rank \( \text{rank}(\text{H}_k) = \text{ncol}(\text{H}_k) \)), and \( \beta_{(k)k}^* \) is a possibly reduced set of regression coefficients to be estimated. While trivial constraints are denoted by \( \text{H}_k = \text{I}_M \), other common examples include parallelism (\( \text{H}_k = \text{I}_M \)), exchangeability, and intercept-only parameters \( \eta_j = \beta_{(j)1}^* \). The overall "large" model matrix is \( \text{X}_{VLM} \), which is \( \text{X}_{LM} \otimes \text{I}_M \) with trivial constraints, while \( \text{X}_{LM} = [(x_{ik})] \) is the ‘smaller’ \( n \times d \) model matrix associated with a \( M = 1 \) model.

Some models have \( \eta_j \)-specific explanatory variables, such as a time-varying covariate, then (6) extends to
\[
\eta_i = \alpha_i + \sum_{k=1}^d \text{diag}(x_{i1}, \ldots, x_{id}) \text{H}_k \beta_{(k)k}^*,
\]
with provision for offsets \( \alpha_i \). The results of this article apply most generally to this case.

The \( \text{W}^{(a)} = (\text{W}_{1}^{(a)}, \ldots, \text{W}_{n}^{(a)}) \) are the working weight matrices, comprising \( \text{W}_{i}^{(a)} = -E[\partial^2 \ell_i/(\partial \eta_i \partial \eta_i^T)] \) at iteration \( a \), for
each log-likelihood component \( \ell_i \). Here, \( \ell = \sum_{i=1}^n \ell_i \) is the log-likelihood, and Fisher scoring is adopted as opposed to Newton-Raphson. Usually the individual expected information matrices (EIMs) are closely related to the working weight matrices \( W_i \), as
\[
\mathbf{I}_{E,i} \delta \theta = E \left[ \frac{\partial^2 \ell_i}{\partial \theta^T \partial \theta} \right],
\]
that is,
\[
(W_i)_{uv} = -E \left[ \frac{\partial^2 \ell_i}{\partial \theta_u \partial \theta_v} \right] = E \left[ \frac{\partial^2 \ell_i}{\partial \theta_u \partial \theta_v} \right] \partial \theta_u \partial \theta_v \quad (8)
\]
for \( u, v \in \{1, \ldots, M\} \). In particular, (8) holds for 1-parameter link functions \( g_w \).

For VGLMs the estimated variance-covariance matrix is
\[
\text{var} (\hat{\beta}^*) = (X_{VLM}^T W(a) X_{VLM})^{-1}, \quad (9)
\]
evaluated at the final iteration, where \( \hat{\beta}^* = (\beta_s^T, \ldots, \beta_d^T)^T \) are all the regression coefficients to be estimated. The iteration number \( a \) will be suppressed henceforth. One reason for the widespread use of the Wald test is their computationally convenience: the estimated variance-covariance matrix (9) is a by-product of the IRLS algorithm, and importantly it is evaluated at the MLE \( \hat{\beta}^* \).

### 3. HDE Detection for VGLMs

For VGLMs the overall model matrix has form \( X_{VLM} = (X_1^T, \ldots, X_d^T)^T \) so that \( A = \sum_{i=1}^n X_i^T W_i X_i \) and
\[
\frac{\partial A}{\partial \beta_r} = \sum_{i=1}^n X_i^T \frac{\partial W_i}{\partial \beta_r} X_i, \quad (10)
\]
for \( v = 1, \ldots, d \), and \( r = \{1, \ldots, R_s\} \) and \( t \in \{1, \ldots, d\} \). For simplicity map all the coefficients from (5) to \( (\hat{\beta}_1, \hat{\beta}_2, \ldots) \) so that for the \( s \)th coefficient \( 0 < a^s \) because \( A \) is positive-definite. Then \( (a^s)^T \) can be computed by
\[
\frac{\partial A}{\partial \beta_s} = -A^{-1} \frac{\partial A}{\partial \beta_s} A^{-1}. \quad (11)
\]
Of central interest for testing \( H_0 : \beta_s = \beta_{s0} \) versus \( H_1 : \beta_s \neq \beta_{s0} \) is
\[
\partial \hat{\beta}_s = \frac{\partial \hat{\beta}_s - \beta_{s0}}{\sqrt{a^s}} = \frac{1}{\sqrt{a^s}} \left[ 1 - \frac{\hat{\beta}_s - \beta_{s0}}{a^s} \right] (a^s)^T. \quad (12)
\]
This equation furnishes a first-derivative detection test: we define the HDE as being present in VGLMs for \( \hat{\beta}_s \) if (12) is negative-valued. Consequently that coefficient's SE and \( p \)-value should be flagged as unreliably biased upward. Only a quadratic form needs to be computed in (11) for each \( a^s \).

With provision to handle constraint matrices and \( x_{ij} \) as in (7),
\[
\frac{\partial W_i}{\partial \beta_r} = \sum_{j=1}^M \frac{\partial W_i}{\partial \beta_j} \frac{\partial \theta_j}{\partial \eta_j} \frac{\partial \eta_j}{\partial \beta_r} \quad (13)
\]
because the working weight matrices have the simple outer product form (8).

#### 3.1. Some Remarks and Properties

Several remarks are in order at this stage, which mainly pertain to VGLMs with \( M = 1 \) parameter.

1. From (12) we say the Wald test statistic for VGLMs has HDE present if and only if
\[
\frac{1}{2} \left( \frac{\partial \hat{\beta}_s - \beta_{s0}}{d \log a^s} \right)^2 - 1 > 0. \quad (14)
\]
For 1-parameter models involving large samples this criteria simplifies to an interesting expression involving the ratio of the Wald and LRT statistics being less than the constant 3/5 (Section 4.1).

2. The following are some sufficient conditions for \( \hat{\beta}_i \) to be monotonic as \( \hat{\beta}_i \to \infty \).
   (i) If
   \[
   \frac{\partial A}{\partial \beta_i} > 0. \quad (15)
   \]
   The result follows from a property of positive-definite matrices (e.g., Seber (2008, eq. (10.46))) and (10)–(12).

One really needs \( (a^s)^T < 0 \) in (12) and this is provided if the inner matrix in (11) is positive-definite. For \( M = 1 \) models, (15) entails that \( \partial w_i / \partial \beta_i > 0 \) in (10).

(ii) If \( (a^s)^T > 0 \). For example, the full-likelihood LM
\[
\eta_1 = \mu, \quad \eta_2 = \log \sigma, \quad Y \sim N(\mu, \sigma^2), \quad (16)
\]
whose EIM is diagonal so that each parameter can be treated separately, the regression coefficients corresponding to \( \mu \) do not suffer from the HDE because the (1, 1) element of the EIM is not a function of \( \mu \). For \( \eta_2 \) the choice of link function does matter, for example, using an identity link then \( \partial w_i / \partial \beta_i = -4 x_{is} \sigma^{-3} \), so that if \( x_{is} > 0 \), then \( \partial A / \partial \beta_i < 0 \) and \( \partial A^{-1} / \partial \beta_i > 0 \) so that \( (a^s)^T > 0 \); thus \( \partial \hat{\beta}_i / \partial \beta_i < 0 \) if \( \hat{\beta}_i \gg 0 \).

3. For 1-parameter VGLMs, (13) simplifies to
\[
\frac{\partial w_i}{\partial \beta_i} = \frac{\partial w_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_i} = \left[ \frac{\partial E_i}{\partial \theta} \right] \frac{\partial \theta_i}{\partial \eta} \frac{\partial \eta_i}{\partial \beta_i} \frac{\partial \eta_i}{\partial x_{is}}. \quad (17)
\]
This makes availability for a variety of link functions, for example, probit and complementary log-log links for binary regression.

4. For standard logistic regression \( \partial w_i / \partial \beta_i = (1 - 2 \mu_i) \mu_i (1 - \mu_i) x_{is} \), which is an odd function about \( \mu_i = 0.5 \). Thus, the model becomes more susceptible to the HDE as \( \mu_i \) approaches a boundary (observed in Figure 2 as \( \beta_2 \)) becomes large.) It is shown that \( |\beta| > 2 \) is a necessary condition for the HDE in Section 4.2. In fact, if \( \tau_0 = \frac{1}{2} \) then it is shown in Appendix A6 in the supplementary materials that approximately \( |\beta_2| > 2.40 \) is needed in order for the HDE to occur, for example, this corresponds to an odds ratio of about 11.0 or higher.
5. For the standard Poisson regression model, applying a similar derivation as logistic regression to a dataset comprising \( N \) points at \( (x = 0, y = \mu_0) \) and \( N \) points at \( (x = 1, y = \mu_1) \), yields

\[
\frac{\partial \hat{\eta}_s}{\partial \beta_i} = \sqrt{\frac{N \mu_0 \mu_1}{\hat{\mu}_0 + \hat{\mu}_1}} \left[ 1 + \frac{\hat{\mu}_2}{\mu_0 - \mu_1} \right].
\] (17)

Thus, conducive conditions for the HDE are when \( \hat{\beta}_2 < 0 \) and \( \hat{\mu}_0 > \hat{\mu}_1 \). Figure 4 shows this for \( N = 1, \mu_0 = 20 \) and \( \mu_1 \), taking on successive values in \( \{1, \ldots, 20\} \). It can be seen that if \( H_0 \) were rejected when \( |\hat{\eta}_s| > 3 \) then it would do so only for \( \mu_1 = 2 \) or \( 3 \) but not 1.

6. Given that \( \hat{\eta}_s \) is computable, it can be of interest to determine the rate at which the \( p \)-value decreases as a function of the effect size. Assuming that \( p_s \approx 2 \Phi(-|\hat{\eta}_s|) \) for a two-sided alternative, called \( p_s \), say, then

\[
\frac{\partial p_s}{\partial \beta_i} = -2 \phi(\hat{\eta}_s) \text{sgn}(\hat{\eta}_s) \frac{\partial \hat{\eta}_s}{\partial \beta_i},
\] (18)

for example, it is plotted in Figure 2(d) for the Hauck and Donner data. An application might be in designing simple experiments to determine how the \( p \)-value will decrease given an increasing treatment effect.

Prompted by the example of Section 6.1, a reviewer raised the question of whether it would be a good idea to measure HDE severity (5.2) through the \( p \)-value instead of \( \hat{\eta}_s \). It is theoretically possible because

\[
\frac{\partial^2 p_s}{\partial \beta_i^2} = 2 \phi(\hat{\eta}_s) \text{sgn}(\hat{\eta}_s) \left[ \frac{\partial \hat{\eta}_s}{\partial \beta_i} \frac{\partial^2 \hat{\eta}_s}{\partial \beta_i^2} - \frac{\partial^2 \hat{\eta}_s}{\partial \beta_i^2} \right],
\] (19)

however this quantity is more complicated and the normality assumption needed to convert the Wald statistic into a \( p \)-value makes this approach a little more fraught. Nevertheless this question deserves further investigation.

7. The basic technique (10)–(13) carries naturally over to a variety of setting such as sandwich estimators, multiple tests, and profile likelihoods (Appendices A2–A5 in the supplementary materials).

8. If \( \theta_u \) and \( \theta_v \) are orthogonal then the \((u, v)\) element of the information matrix remains 0 upon differentiation, therefore agrees intuitively that the Wald statistic for \( \hat{\beta}_u \) is minimally affected by \( \hat{\beta}_v \).

9. Second derivatives for the Wald statistic follow as before; these can be used, for example, to determine whether \( \hat{\eta}_s \) is a convex function, that is, is a particular model impervious to the HDE? Use

\[
\frac{\partial^2 \hat{\eta}_s}{\partial \beta_i^2} = \frac{1}{(a^{as})^{3/2}} \times \left[ -(a^{as})' + \frac{\hat{\beta}_i - \hat{\beta}_0}{2} \left[ \frac{3}{2} \left( \frac{a^{as}}{a^{as}} - (a^{as})' \right) \right] \right]
\] (20)

and

\[
\frac{\partial^2 A^{-1}}{\partial \beta_i^2} = A^{-1} \left[ 2 \frac{\partial A}{\partial \beta_i} A^{-1} \frac{\partial A}{\partial \beta_i} - \frac{\partial^2 A}{\partial \beta_i^2} \right] A^{-1}
\] (21)

to allow for the computation of \((a^{as})''\). To compute (21) requires

\[
\frac{\partial^2 w_i}{\partial \beta_i^2} = \left[ \frac{\partial^2 w_i}{\partial \beta_i^2} - \frac{\partial w_i}{\partial \beta_i} \frac{\partial^2 \theta}{\partial \beta_i \partial \eta} + \frac{\partial w_i}{\partial \beta_i} \frac{\partial^2 \theta}{\partial \beta_i \partial \eta} \right] \frac{\partial \theta}{\partial \beta_i} \frac{\partial \theta}{\partial \beta_i}
\] (22)

This in turn requires the third derivatives of the link function in its inverse form. The ordinary form \( \frac{\partial^3 \theta}{\partial \eta^3} \) is straightforward while the inverse form can be reexpressed as

\[
\frac{\partial^3 \theta}{\partial \eta^3} = \left( \frac{\partial \theta}{\partial \eta} \right)^4 \left[ 3 \frac{\partial^2 \theta}{\partial \eta^2} - \frac{\partial \theta}{\partial \eta} \frac{\partial^3 \theta}{\partial \eta^3} \right],
\] (23)
for example, for logistic regression $\theta^3/\mu^3 = 2[1 - 3\mu(1 - \mu)]/[\mu(1 - \mu)^2]$ and (23) is $\mu(1 - \mu)[1 - 6\mu(1 - \mu)]$.

4. Inference for a 1-Parameter Regular Model

4.1. Large Sample Results

The results for $M = 1$ VGLMs can be investigated further. Consider the special case of $\beta_1$ being the sole parameter $\theta$ of a regular distribution. For simplicity use the observed information here. We have the following large sample result concerning two tipping points.

**Theorem 1.** For $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, and where the observed information is evaluated at $\hat{\theta}$, if the HDE is present then

(a) the ratio of the Wald and LRT statistics satisfies

$$\frac{W(\hat{\theta}, \theta_0)}{W_L(\hat{\theta}, \theta_0)} < \frac{3}{5} + Op(n^{-1}), \quad (24)$$

(b) and the ratio of the Wald and Rao score test statistics satisfies

$$\frac{W(\hat{\theta}, \theta_0)}{W_S(\hat{\theta}, \theta_0)} < \frac{1}{4} + Op(n^{-3/2}). \quad (25)$$

**Proof.** (a) As $a^{11} = [-\ell''(\theta)]^{-1}$ the HDE is present iff (14):

$$1 < \frac{\hat{\theta} - \theta_0}{2} \frac{d}{d\theta} \left[ -\ell''(\theta) \right]^{-1} = \frac{\hat{\theta} - \theta_0}{2} \left[ -\ell''(\hat{\theta}) \right]. \quad (26)$$

A Taylor series expansion of $\ell(\theta_0)$ about $\hat{\theta}$ gives

$$W_L(\hat{\theta}, \theta_0) = -\ell'(\hat{\theta})(\hat{\theta} - \theta_0)^2 + \frac{1}{3} \ell''(\theta)(\hat{\theta} - \theta_0)^3 + Op(n^{-1})$$

$$= \left[ -\ell''(\hat{\theta}) \right] (\hat{\theta} - \theta_0)^2 \left\{ 1 + \frac{2}{3} \cdot \frac{1}{2} \left[ -\ell''(\hat{\theta}) \right] \right\} + Op(n^{-1})$$

$$< \left[ -\ell''(\hat{\theta}) \right] (\hat{\theta} - \theta_0)^2 \left\{ 1 + \frac{2}{3} \cdot \frac{1}{1} \right\} + Op(n^{-1}) \text{ by (26)}$$

$$= \frac{5}{3} W(\hat{\theta}, \theta_0) + Op(n^{-1}).$$

The inequality (24) follows from the property $W_L = Op(1)$.

(b) Using $\tilde{W_S} = \ell'(\theta_0)/[-\ell'(\theta)]^{1/2}$, then expanding the numerator with a Taylor series expansion about $\hat{\theta}$ yields

$$\tilde{W_S}(\hat{\theta}, \theta_0)$$

$$= \frac{\hat{\theta} - \theta_0}{\sqrt{-\ell''(\hat{\theta})}} \left[ -\ell''(\hat{\theta}) + \frac{1}{2} (\hat{\theta} - \theta_0) \ell'''(\hat{\theta}) + Op(n^{-1/2}) \right]$$

$$= \frac{(\hat{\theta} - \theta_0)}{-\ell''(\hat{\theta})} \left[ -\ell''(\hat{\theta}) + \frac{1}{2} (\hat{\theta} - \theta_0) \ell'''(\hat{\theta}) \right] + Op(n^{-3/2})$$

$$= \tilde{W} \left[ 1 + \frac{1}{2} (\hat{\theta} - \theta_0) \frac{\ell'''(\hat{\theta})}{-\ell''(\hat{\theta})} \right] + Op(n^{-3/2}).$$

As $W_S = Op(1)$ and

$$\tilde{W} = \sqrt{-\ell''(\hat{\theta})} \left[ 1 + \frac{\hat{\theta} - \theta_0}{2} \ell'''(\hat{\theta}) \right] \quad (27)$$

then

$$\frac{\tilde{W}_S(\hat{\theta}, \theta_0)}{W(\hat{\theta}, \theta_0)} = 1 + 1 - \sqrt{-\ell''(\hat{\theta})}$$

$$+ Op(n^{-3/2}) > 2 + Op(n^{-3/2})$$

when the HDE is present. Taking the squared reciprocal gives (25).

Equation (24) can be interpreted by saying that with large samples, if the Wald statistic becomes too small relative to the LRT statistic (which is likely to be more accurate) then the HDE will become present. Likewise we can interpret (26) by saying that as $|\hat{\theta} - \theta_0| \to \infty$, if the negative second derivative of $\ell(\hat{\theta})$ does not grow fast enough compared to the third derivative of $\ell(\hat{\theta})$ then the HDE will become present. The accuracy of the $3/5$ bound depends upon the fourth and higher derivatives of $\ell$.

Equations (24) and (25) suggest that $W_L \approx \frac{5}{12} W_S$ when the HDE first starts to occur. In fact empirical findings indicate that $W_L/W_S < 5/12$ in the presence of a strong HDE.

Extending the result to the two-parameter case is wieldy. However, applying this result to the 2-parameter Hauck and Donner (1977) data, the 3/5 threshold lies between $R = 93$ and 94 successes, whereas the HDE becomes present for $\beta_2$ at $R \geq 92$ (Figures 2(a) and (b)). This suggests that the method can work well when the number of parameters is low or are orthogonal. As another numerical example, when fitting a Poisson regression to the data described for (17) one obtains a perfect match because the two cases of HDE present corresponds to a ratio $< 3/5$ and the other cases have a ratio $> 3/5$.

Unfortunately the LRT and Wald statistics are not independent; if they were then their ratio would have a $F_{1,1}$ distribution whose mean is infinite. The lower tail probability at the 3/5 quantile of this distribution is 0.42, indicating that their positive correlation creates a bias away from the null.

As $W_L - W = Op(1/\sqrt{n})$, it follows that $W/W_L = 1 + Op(1/\sqrt{n})$. Indeed, $W/W_L$ has approximate asymptotic expectation $3 - \text{cov}(W_L, W)$, where

$$\text{cov}(W_L, W) 
\approx \text{cov} \left( W_L, W_L - \frac{1}{3} \ell'''(\theta_0) (\hat{\theta} - \theta_0)^3 + O(n^{-2}) \right)$$

$$\approx \text{var} W_L - \frac{1}{3} \text{cov} \left( 2 [\ell(\hat{\theta}) - \ell(\theta_0)], \ell'''(\theta_0) (\hat{\theta} - \theta_0)^3 \right)$$

$$\approx 2 - \frac{2}{3} \ell'''(\theta_0) \text{cov} (\ell(\hat{\theta} - \theta_0), (\hat{\theta} - \theta_0)^3)$$

$$\approx 2 - \frac{2}{3} \ell'''(\theta_0) \ell'(\theta_0) E \left[ (\hat{\theta} - \theta_0)^4 \right]$$

$$\approx 2 \left[ 1 - \frac{1}{3} \ell'''(\theta_0) \ell'(\theta_0) 3 \text{var} (\theta_0)^2 \right],$$

as $\hat{\theta} \sim N(\theta_0, [\ell'(\theta_0)]^{-1})$ under $H_0$, and $\mu_4 = 3\sigma^4$ for a normal distribution. Hence

$$\text{cov}(W_L, W) \approx 2 \left[ 1 - \ell'(\theta_0) \cdot [\ell'''(\theta_0)]^{-1} \cdot \ell'(\theta_0) \right]. \quad (28)$$
Thus
\[ E(W/W_L) \approx E(W_L/W) \]
\approx 1 + 2 \ell'(\theta_0) \cdot (\ell''(\theta_0))^{-2} \cdot \ell'''(\theta_0), \tag{29} \]
and similarly
\[ \text{corr}(W_L, W) \approx 1 - \ell'(\theta_0) \cdot (\ell''(\theta_0))^{-2} \cdot \ell'''(\theta_0), \]
\[ \text{var}(W/W_L) \approx \text{var}(W_L/W) \approx 4 \ell'(\theta_0) \cdot (\ell''(\theta_0))^{-2} \cdot \ell'''(\theta_0). \]
An approximate upper bound from Chebyshev's inequality shows that
\[ \Pr(|W/W_L - 1| \geq \frac{2}{\ell''(\theta_0)} \cdot \frac{\text{var}(W/W_L)}{(2/5)^2}) = 25 \ell'(\theta_0) \cdot (\ell''(\theta_0))^{-2} \cdot \ell'''(\theta_0), \]
as the probability of the HDE occurring by chance, given \( H_0 \), however the bound is not very sharp.

A closing remark is that the asymptotic normality of the MLE can be augmented with an additional regularity condition to prevent the HDE, by restricting the parameter space. Under \( H_0 \), the extra condition is
\[ \Theta_0 = \left\{ \theta : \theta - \theta_0, \ell'''(\theta) < 1 \right\}, \tag{30} \]
so that \( \sqrt{n}(\theta - \theta_0) \to \mathcal{N}(0, \mathbf{I}_{E1}^{-1}(\theta_0)) \) in \( \Theta_0 \) (see, e.g., Cox and Hinkley 1974, pp. 294–295). Here, \( \mathbf{I}_{E1} \) is the expected information of one observation, and \( \to \) denotes convergence in distribution.

### 4.2. Disproportional Sampling

In the case of a 2 \( \times \) 2 table it is now shown that, for a fixed size effect \( \beta_2 \), disproportionally sampling can be used to circumvent the HDE.

**Table 1** is a modification of the HD data to allow for disproportional sampling. Here, \( R \to N \) so that the (2, 1) cell becomes low, hence individuals with \( x_2 = 1 \) need to be sampled with greater intensity. This is achieved by having \( N_1^* = c^*N_1 \) where the parameter \( c^* \geq 1 \). With \( c^* = 1 \), the usual scenario, increasing \( N \) had no effect on the HDE as the sign of (12) is unchanged. The quantity \( f_0 = N_0/N_1^* \) is then used to measure the relative sampling intensity. The sample proportions \( \pi_0 = R_0/N_0 \) and \( \pi_1 = R/N_1 \) remain unchanged.

With a logit link the new sampling scheme affects the intercept only. The inverse crossproduct matrix is
\[ A^{-1} = \frac{1}{N_1^*} \times \left( \begin{array}{ccc} f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & -f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} \\ -f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & -f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} \\ f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & -f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} & f_0^{-1} \pi_0^{-1} (1 - \pi_0)^{-1} \end{array} \right), \]
and then by (11), \( (a^2)^* = (2\pi_1 - 1)/(N_1^* \pi_1 (1 - \pi_1)) \). The HDE will be present for \( \beta_2 \) if
\[ 1 < \beta_2 \cdot \left( \frac{\pi_1 - \frac{1}{2}}{f_0 \pi_0 (1 - \pi_0) + \pi_1 (1 - \pi_1)} \right). \tag{31} \]
This shows that \( \beta_2 > 2 \) is needed before the HDE is possible, provided that \( \pi_1 \approx 1 \) and \( \pi_0 \) is away from the boundaries. This corresponds to an odds ratio of about 7.4 or higher. If \( f_0 \approx 0 \) then the HDE is unlikely, in particular, the quantity
\[ f_0 \pi_0 (1 - \pi_0) \]
\[ \pi_1 (1 - \pi_1) \]
measures the sampling effect on the HDE: small/large values of \( \gamma \) implies HDE is unlikely/likely, respectively. This make intuitive sense as the (2, 1) cell increases as a function of the total sample size by choosing \( c^* \to \infty \) so that \( f_0 \to 0^+ \).

Appendix A6 extends the results of this section to a more general setting. In particular, when \( \pi_0 = \frac{1}{2} \) in **Table 1** so that \( \beta_1 \) in (1) need not be estimated, then it is shown that approximately \( |\beta_2| > 2.40 \) is needed in order for the HDE to occur. This corresponds to an odds ratio of about 11.0 or more, or about 0.091 or less.

### 5. Refinements

#### 5.1. Finite-Differences

Unfortunately implementing HDE detection for any particular VGLM based on (13) is labor-intensive, for example, Appendix A1, which has two consequences. First, this work suggests that the EIM is to be preferred over the observed information matrix (see Efron and Hinkley (1978) who preferred the latter), because terms often vanish upon expectation and therefore lead to simplification, for example, (16). Second, numerical computation of the first two derivatives of \( \hat{W} \) circumvents this problem and provides a general method that empirical testing has shown to work well. In particular, simplify (13) to
\[ \frac{\partial W_i}{\partial \beta_{(r)s}} = \sum_{j=1}^{M} \frac{\partial W_i}{\partial \eta_j} (H_s)_{jr} x_{ij}, \tag{32} \]
where \( \partial W_i/\partial \eta_j \) may be approximated by, for example, central finite differences. Similarly,
\[ \frac{\partial^2 W_i}{\partial \beta_{(r)s}^2} = \sum_{j=1}^{M} \sum_{k=1}^{M} \frac{\partial^2 W_i}{\partial \eta_j \partial \eta_k} (H_s)_{tr} (H_s)_{jr} x_{ij} x_{kj} \tag{33} \]
with separate formulas to handle the diagonal and off-diagonal elements. A step value of \( h_j \approx 0.005 \) has been found to be reasonable for most VGLMs, being on the \( \eta_j \)-scale.

### 5.2. Severity Measures and Parameter Space Partitioning

So far, the above establishes whether a particular VGLM suffers from the HDE, however, a small negative derivative could indicate a mild or extreme effect. It could also be argued that once the Wald statistic starts decelerating but without having
Table 2. HDE severity measures: the 0s define 5 cutpoints that are used to define 6 categories of HDE.

| Severity       | Key property | $\hat{V}'_i$ | $\text{sgn}(\hat{\beta}_1) \cdot \hat{V}'_i$ | $\xi'$ | $R$ |
|----------------|--------------|--------------|---------------------------------|-------|-----|
| None [boundary]| Convex       | +            | 0                               |       |     |
| Faint (very mild)| Concave    | +            | 0                               | +     |     |
| Weak (mild) [boundary]| Concave | +            | 0                               | -     | 10.5, 69.5 |
| Moderate [boundary]| Concave | -            | 0                               | -     | 2.5, 91.5 |
| Strong (severe) [boundary]| Concave | -            | 0                               | +     | 1.5, 97.5 |
| Extreme (very severe)| Convex | -            | 0                               | +     | 0.5, 98.5 |

Note: $\xi = 0$ is unused here. The sign of various quantities are given where possible, and the key property of $\hat{V}_i$ is given. Approximate boundary values of $R$ are for the dataset of Table 1(a) with $R_0 = 25$.

A negative first derivative, then the HDE has already started to occur. Thus, it can be asserted that the Wald statistic may be aberrant without the HDE being formally present in a regression coefficient. Fortunately, it is possible to gauge how severe the HDE is based the first two derivatives of $\hat{V}_i$. As $\hat{V}_i$ is asymptotically locally quadratic about the origin, $\hat{V}_i(\hat{\beta})$ is assumed piecewise convex–concave–convex for $\hat{\beta} > 0$, and ditto for $\hat{\beta} < 0$. Table 2 is a summary.

Using the simple notation $\hat{\alpha}$ to denote the $x$-axis, let $\hat{\xi}(\hat{\alpha})$ be the intersection of the normal line at $(\hat{\beta}, \hat{V})$ with the $x$-axis (purple dashed line of Figure 6). The movement of $\hat{\xi}$ as the coefficient changes allows further properties of the curve to be determined additional to the location of inflection points. Further partitioning using $\hat{\xi} = 0$ is work in progress (Yee 2021). The use of $\hat{\xi}$ is a more convenient alternative to using the tangent because, using Figure 6 as an example, the normal line keeps $\hat{\xi}_2$ finite and avoids a discontinuity as one moves from left to right. In contrast, if we had used the point where the tangent intersects with the $x$-axis, then it would approach $-\infty$ a little beyond $\hat{\beta}_2 \approx 3$ and then jump to $+\infty$ a little before $\hat{\beta}_2 \approx 4$, that is, about the location of the peak. This discontinuous behavior is less tractable and so $\hat{\xi}$ is preferable. Denoting the cutpoints as $\hat{\beta}_{nf}, \hat{\beta}_{wm}, \hat{\beta}_{ms}, \hat{\beta}_{se}$ for none, faint, weak (mild), moderate, severe (strong), extreme, they are defined by $\hat{V}''(\hat{\beta}_{nf}) = 0$, $\hat{\xi}'(\hat{\beta}_{wm}) = 0$, $\hat{\xi}'(\hat{\beta}_{ms}) = 0$, $\hat{V}''(\hat{\beta}_{se}) = 0$, so that $\hat{\beta}_{nf}$ and $\hat{\beta}_{se}$ are inflection points. For positive estimates

$$0 \leq \hat{\beta}_{nf} \leq \hat{\beta}_{wm} \leq \hat{\beta}_{ms} \leq \hat{\beta}_{se} < \infty,$$

and

$$\hat{\xi}(\hat{\beta}) = \hat{\beta} + \hat{V}''(\hat{\beta}) \cdot \hat{V}'(\hat{\beta}),$$

$$\hat{\xi}'(\hat{\beta}) = 1 + \left(\frac{\hat{V}''(\hat{\beta})}{\hat{V}'(\hat{\beta})}\right)^2 + \hat{V}'(\hat{\beta}) \cdot \hat{V}''(\hat{\beta}).$$

For the dataset of Table 1(a) the scheme classifies no HDE for $R = 26, \ldots, 40$, faint HDE for $R = 11, \ldots, 25, 41, \ldots, 69$, mild HDE for $R = 3, \ldots, 10, 70, \ldots, 91$, moderate HDE for $R = 2, 92, \ldots, 97$, severe HDE for $R = 1, 98$, and extreme HDE for $R = 99$ (Figure 6).

As a consequence, the HDE severity measures allow the parameter space $\Theta$ to be partitioned into an interior $\Theta^{oo}$ where the regularity conditions hold practically, which is encased by layers at the boundary of increasing HDE severity. As mentioned in Section 1, $\Theta^{oo}$ is at most weak HDE and $\Theta^{hde}$ is at least moderate HDE. Figure 5 displays this schematically. For a given dataset and regression model, not all the layers may be present and $\Theta$ may be discrete. The descriptors should be interpreted relatively rather than absolutely— their purpose is to provide an ordinal categorization of the HDE severity.
6. Examples

6.1. Birds and Logistic Regression

Mangiafico (2015) applied a variable selection algorithm based on the AIC to choose covariates in a logistic regression applied to a bird dataset with 67 cases after missing values have been removed. Six variables plus an intercept were chosen out of a possible 13 variables (Table 3) for the response Status.

For this model, three variables display the HDE: Upland, Migr, and Indiv (Table 3), and Mass is weakly affected. According to the usual Wald test, the former two have p-values of moderate strength: 2% and 3%, while Indiv is more strongly significant (0.005). However, LRT p-values for the 3 variables are 0.00135, 0.0076, 5 × 10⁻¹¹, respectively. This suggests that these variables are considerably more statistically significant than would naively appear; the ratios of the p-values are approximately 15, 4, and 1 × 10⁻⁴. The latter is huge and shows that a moderate HDE can be associated with an enormous relative effect on the p-value (although qualitatively similar); that variable is actually skewed and a log-transformation would be recommended.

6.2. Wine and the Partial Proportional Odds Model

Table 1 of Kosmidis (2014) reports the Wald table of a partial proportional odds model (PPOM) fitted to a wine tasting dataset where the response bitterness was measured on a 5-level ordinal scale. The two binary variables x₂ = “temperature” and x₃ = “contact” were measured in this experiment, and the parallelism assumption was applied to x₂ only. Three of the SEs are inflated. The data were first analyzed in Randall (1989).

Table 4 provides additional information to the Wald table, viz. the first two derivatives of the signed root Wald statistics, and the HDE severity.

| Parameter | MLE | SE  | \( \hat{\eta} \) | \( \hat{\eta}'' \) | HDE severity |
|-----------|-----|-----|----------------|----------------|--------------|
| µ         |     |     | -0.56          | 0.18            |              |
| π         |     |     | 0.47           | 3.13            |              |

The latter is huge and shows that a moderate HDE can be associated with an enormous relative effect on the p-value (although qualitatively similar); that variable is actually skewed and a log-transformation would be recommended.

The tests are \( H_k : \beta_k = \beta_k \) versus \( H_{1k} : \beta_k \neq \beta_k \) where usually \( \beta_k = 0 \). By an “HDE-free Wald test,” it is meant that \( \beta_k \) is replaced by its hypothesized value \( \beta_k \) when computing the test statistic’s SE so that its derivative with respect to \( \beta_k \) vanishes.

Two options open up: whether or not to perform IRLS iterations for the other coefficients. If so, then this is equivalent to fitting the LRT model under \( H_{0k} \), so it is not surprising that the time cost is similar. Notationally, let \( ( \hat{\beta}_{0k}, \hat{\beta}_{-k} ) \) correspond to no (further) iteration and \( ( \hat{\beta}_{0k}, \hat{\beta}_{-k} ) \) be with iteration.

Table 3 of Kosmidis (2014) discussed adjustments specifically for cumulative link models and proposes bias reduction methods in order to safeguard against infinite parameter estimates. In this example, the number of coefficients manifesting the HDE ought to alert the practitioner of a nonstandard situation. Furthermore, suspicion for something gone awry should have been raised by monitoring convergence and observing that this took 19 IRLS iterations—this is considerably more than the usual 6–8 iterations—and that the decrease in deviance during the last 10 iterations was only slight.

7. Computational Details and Recommendations

One can obtain an HDE-free Wald test by evaluating the denominator of the Wald statistic at \( \hat{\theta}_0 \) rather than \( \hat{\theta}_M \). For simplicity of notation, this section also enumerates the regression coefficients to be estimated of a general VGLM as \( \beta_k \) for \( k = 1, \ldots, p_{VLM} \).

Tests are \( H_k : \beta_k = \beta_k \) versus \( H_{1k} : \beta_k \neq \beta_k \) where usually \( \beta_k = 0 \). By an “HDE-free Wald test,” it is meant that \( \beta_k \) is replaced by its hypothesized value \( \beta_k \) when computing the test statistic’s SE so that its derivative with respect to \( \beta_k \) vanishes.

Then two options open up: whether or not to perform IRLS iterations for the other coefficients. If so, then this is equivalent to fitting the LRT model under \( H_{0k} \), so it is not surprising that the time cost is similar. Notationally, let \( ( \hat{\beta}_{0k}, \hat{\beta}_{-k} ) \) correspond to no (further) iteration and \( ( \hat{\beta}_{0k}, \hat{\beta}_{-k} ) \) be with iteration.

Computational, the use of \( \beta_{0k} \) instead of \( \hat{\beta}_{0k} \) is implemented by deleting the \( k \)th column of \( X_{0k} \) and adding \( x_k \beta_{0k} \) to the matrix of offsets. If iterating for the other coefficients, equivalent values can be obtained from the original model and then convergence is usually rapid (often only 1 or 2 IRLS iterations are needed).

To obtain the SE for \( \beta_k \) for an HDE-free Wald test, the main steps are as follows.

(i) Optionally iterate (i.e., use the MLEs \( \hat{\beta}_{-k} \))

(ii) Compute \( \eta_i \) using \( ( \hat{\beta}_{0k}, \hat{\beta}_{-k} ) \) or \( ( \beta_{0k}, \hat{\beta}_{-k} ) \). Compute the Cholesky decompositions \( U_i \) of the \( W_i \).
(v) Compute \( \text{diag}(U_1, \ldots, U_n) X_{\text{VLM}} \) and its QR-decomposition.  
(vi) Compute \( R^{-1} \) and then \( (R^{-1} R^{-1/2})_{kk} \) to obtain \( \text{SE}_k \).

These steps are performed for each \( \beta_k \).

A small numerical study involving timing various VGLMs fitted with the author’s software gave the following results.

- The cost of conducting an HDE test on all the regression coefficients is typically about \( \frac{1}{4} \) the cost of obtaining the HDE-free iterated Wald statistics.
- Without iterations, HDE-free Wald tests can be about 25% less costly time-wise compared to LRTs.
- With iteration, HDE-free Wald tests can be about 30% more costly time-wise than LRTs.
- The cost of a score test is similar to an iterated Wald test.

These results suggest that the cost of a LRT and a HDE-free Wald test is roughly comparable.

For the practitioner, the above suggests that a reasonable strategy is to first apply an HDE test to all the regression coefficients. If any are affirmative, then the next step is applied to those coefficients and depends on whether SEs are sought, for example, as a rough measure of statistical uncertainty. Thus, second, when SEs are required, HDE-free Wald tests should be conducted for those coefficients, and should only \( p \)-values be necessary, then LRTs should be computed instead. As for deciding whether iterations are required, if the computational expense is of concern, then the noniterated variant is suggested as it can be approximately half the cost of the iterated variant.

It should be noted that conducting hypothesis tests may be difficult for some models, such as a nonparallel cumulative link model because of intrinsic order restrictions such as \( \beta_{11k} < \cdots < \beta_{j1k} < \cdots < \beta_{M1k} \). Circumventing the difficulties involved is an area for future research. Another note is that the score test lacks the intuitive appeal of the Wald test and may be inconsistent (Freedman 2007), therefore it is suggested that the order of preference of the tests be, in decreasing order, LRTs followed by HDE-free Wald tests followed by score tests.

8. Discussion

In an era of high-dimensional statistics and Big Data the \( p \)-value remains a chief centerpiece of classical frequentist statistical inference (Meinshausen, Meier, and Bühlmann 2009; Fan 2014; Dezeure et al. 2015) despite recent statements about their misuse (Wasserstein and Lazar 2016). Indeed, Siegfried (2010) wrote: “It’s science’s dirtiest secret: The ‘scientific method’ of testing hypotheses by statistical analysis stands on a flimsy foundation.” Although most of the weakness is interpretative this article highlights another deficiency in the form of the HDE.

This work was motivated by practical Wald testing in a general regression setting, and has developed methods for testing whether an estimate \( \hat{\beta} \) in a VGLM gives rise to parameters so close to the parameter space boundary that some Wald statistics are aberrant. The practical implication is clear: for SEs computed at the MLE establishing whether the HDE is manifest in a fitted model should be determined where possible, and if so then the LRT or other large-sample tests be conducted instead. Ideally HDE testing should be carried out whenever a Wald table is presented, and statistical software should be upgraded to make this practical and automatic. Software for variable selection based on the usual Wald statistics need modification too. With business intelligence software companies being a major driver of Big Data, it is plausible that litigation could occur if they fail to respond adequately to implement new methodologies such as here to address known flaws such as the HDE. As a minimum, disclaimer statements that the HDE is not detected could be issued as interim measures.

This article is an important first step toward shedding light on the structure of the parameter space in a practical sense. There is further work to be done, such as refining the severity measures of Section 5.2, exploring consequences under a Bayesian framework, and seeing if adjustments are needed for multiplicity if we switch to another test.

A parting remark is that it has been disappointing to see that most texts on statistical inference and practice do not even mention the HDE, especially those making heavy use of Wald tables. Of the few that do, various authors have described the HDE as “the major statistical problem of \( \mathcal{W} \)” and “certainly disturbing.” It is hoped that this work will help make the HDE a more recognized problem and provide a practical solution.

Supplementary Materials

The supplementary materials include specific HDE details needed for the zero-inflated Poisson and cumulative logit model as examples, sandwich estimators, multiple testing, the Cox proportional hazards model, and profile likelihoods. It is shown that it is possible to determine whether the HDE will occur for a binary covariate in a logistic regression. An R script file is also included to run the examples using the \texttt{VGAM} package.

Acknowledgments

Thanks are extended to the Centre for Applied Statistics and School of Mathematics and Statistics at the University of Western Australia for hospitality during a workshop given there and for valuable feedback from participants that helped lead to this work. Constructive comments from the editor, an associate editor, reviewers, George Seber and Elbert Chia are gratefully acknowledged. This article is dedicated to the memory of Prof. Alastair J. Scott FASA FRSNZ (1939–2017), a real statistics professor at the University of Auckland. SDG.

ORCID

Thomas W. Yee  
http://orcid.org/0000-0002-9970-3907

References

Albert, A., and Anderson, J. A. (1984), “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models,” Biometrika, 71, 1–10. [1766]

Cox, D. R., and Hinkley, D. V. (1974), Theoretical Statistics, London: Chapman & Hall. [1770]

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), “High-Dimensional Inference: Confidence Intervals, \( p \)-Values and R-Software hdi,” Statistical Science, 30, 533–558. [1773]

Efron, B., and Hinkley, D. V. (1978), “Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information” (with discussion), Biometrika, 65, 457–487. [1770]

Fan, J. (2014), “Features of Big Data and Sparsest Solution in High Confidence Set,” in Past, Present, and Future of Statistical Science, eds. X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang, Boca Raton, FL: Chapman and Hall/CRC, pp. 507–523. [1773]
Fears, T. R., Benichou, J., and Gail, M. H. (1996), "A Reminder of the Fallibility of the Wald Statistic," *The American Statistician*, 50, 226–227. [1765]

Freedman, D. A. (2007), "How Can the Score Test Be Inconsistent?," *The American Statistician*, 61, 291–295. [1773]

Hauck, W. W., and Donner, A. (1977), "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association*, 72, 851–853. [1763,1764,1769]

(1980), "Corrigenda: Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association*, 75, 482. [1763]

Kosmidis, I. (2014), "Improved Estimation in Cumulative Link Models," *Journal of the Royal Statistical Society*, Series B, 76, 169–196. [1765,1772]

Lesaffre, E., and Albert, A. (1989), "Partial Separation in Logistic Discrimination," *Journal of the Royal Statistical Society*, Series B, 51, 109–116. [1766]

Mangiafico, S. S. (2015), "An R Companion for the Handbook of Biological Statistics," Version 1.3.1, available at [http://rcompanion.org/documents/RCompanionBioStatistics.pdf](http://rcompanion.org/documents/RCompanionBioStatistics.pdf) [1772]

Meeker, W. Q., and Escobar, L. A. (1995), "Teaching About Approximate Confidence Regions Based on Maximum Likelihood Estimation," *The American Statistician*, 49, 48–53. [1765]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [1773]

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Series A, 135, 370–384. [1763]

Nelson, F. D., and Savin, N. E. (1990), "The Danger of Extrapolating Asymptotic Local Power," *Econometrica*, 58, 977–981. [1765]

Randall, J. H. (1989), "The Analysis of Sensory Data by Generalized Linear Model," *Biometrical Journal*, 31, 781–793. [1772]

Seber, G. A. F. (2008), *A Matrix Handbook for Statisticians*, Hoboken, NJ: Wiley. [1767]

Siegfried, T. (2010), "Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics," *Science News*, 177, 26–29. [1773]

Storer, B. E., Wacholder, S., and Breslow, N. E. (1983), "Maximum Likelihood Fitting of General Risk Models to Stratified Data," *Journal of the Royal Statistical Society*, Series C, 32, 172–181. [1765]

Therneau, T. M., and Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, New York: Springer. [1765]

Vaeth, M. (1985), "On the Use of Wald's Test in Exponential Families," *Statistical Review/Revue Internationale de Statistique*, 53, 199–214. [1764,1765]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA’s Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133. [1773]

Xing, G., Lin, C.-Y., Wooding, S. P., and Xing, C. (2012), "Blindly Using Wald’s Test Can Miss Rare Disease-Causal Variants in Case-Control Association Studies," *Annals of Human Genetics*, 76, 168–177. [1764]

Yee, T. W. (2015), *Vector Generalized Linear and Additive Models: With an Implementation in R*, New York: Springer. [1765,1766]

(2021), "Some New Results Concerning the Hauck–Donner Effect" (in preparation). [1771]