Abstract: This paper revisits the statistical specification of near-multicollinearity in the logistic regression model. We argue that the ceteris paribus clause, which assumes that the maximum likelihood estimator of $\beta$ remains constant as the correlation ($\rho$) between the regressors increases, invoked under the traditional account of near-multicollinearity is rather misleading. We derive the parameters of the logistic regression model and show that they are functions of $\rho$, indicating that the ceteris paribus clause is unattainable. Monte Carlo simulations confirm these findings and further show that: coefficient estimates and related statistics fluctuate in a non-symmetric, non-monotonic way as $|\rho| \to 1$; that the impact of near-multicollinearity is centered on the estimates of $\beta$; and that the impact on substantive inferences does not necessarily follow what the traditional account implies.

Keywords: logistic regression; model diagnostics; near-multicollinearity.

JEL Classification: C1; C2; C4.

1 Introduction

In econometrics, statistics and related disciplines, near-multicollinearity is seen as a pervasive problem that can have significant consequences for estimation and the reliability of substantive inference. Traditionally, when regressors in a regression model are perfectly or nearly perfectly correlated, estimates of the individual regression coefficients become unstable and the inferences based on the model may be misleading. This condition is known as multicollinearity (Mason, Webster, and Gunst 1975). Multicollinearity occurs when variables in the model are correlated to an extent that individual regression coefficient estimates become unreliable. When the regressors have an exact linear relationship, they are said to be perfectly collinear. When the relationship between the predictor variables is almost linear (but not exact), this results in the phenomenon known as near-multicollinearity, the problem specifically addressed in this paper and frequently encountered in the applied literature.

The problems traditionally associated with near-multicollinearity in the logistic regression model are similar to those traditionally found for the linear regression model (Menard 2002). This argument is based on the premise that the logistic regression model can be expressed linearly in terms of the log-odds as $\ln(p/(1-p))=\beta'X+e$, where $p$ is the mean of the binary dependent variable ($Y$), $\beta$ is a vector of parameters, $X$ is a $(k\times1)$ vector of explanatory variables, and $e$ is a zero mean IID random error term. Therefore, similar to the linear regression model, near-multicollinearity in the logistic regression model may lead to near-singularity among the columns of $X$, resulting in numerical instability, inflated variances of the maximum likelihood coefficient estimators, and misleading inferences. Many of the diagnostics suggested in the literature use the
linear form of the logistic regression for detecting problems with near-multicollinearity (e.g. variance inflation and tolerance factors) (Hosmer and Lemeshow 2000; Menard 2010).

Spanos and McGuirk (2002) challenge the traditional account of near-multicollinearity for the linear regression model. The traditional account relies upon a \( \text{ceteris paribus} \) clause that assumes that the estimators of \( \beta \) and the \( R^2 \) remain constant as \( |\rho| \to 1 \), where \( \rho \) represents the correlation between two regressors. However, the estimators of \( \beta \) and \( R^2 \) are in fact functions of \( \rho \) and change as \( |\rho| \to 1 \). This implies that changes in associated statistics, such as the variances \( \text{Var}(\hat{\beta}) \) and \( t \)-ratios \( t(\hat{\beta}) \) of the estimators will be different than previously thought. In fact, Spanos and McGuirk (2002) conclude that “the changes in the relevant statistics \( \text{Var}(\hat{\beta}), t(\hat{\beta}) \) are rather different than the usual account suggests; the \emph{ceteris paribus} clause is unattainable and the traditional account needs to be reconsidered (p: 369).” It is the unattainable \emph{ceteris paribus} clause that is also invoked when examining multicollinearity in the logistic regression model that is of major concern in this paper. It is the link between the linear regression model and linearized form of the logistic regression model in terms of the log odds that makes the findings from Spanos and McGuirk (2002) relevant in re-evaluating the statistical specification of near-multicollinearity in the logistic regression model. Derivations by Scrucca and Weisberg (2004) show that the parameters of the logistic regression model may be functions of \( \rho \). Thus, the \emph{ceteris paribus} clause is unlikely to hold for the logistic regression model.

The purpose of this paper is to revisit the statistical specification of near-multicollinearity in the logistic regression model under the assumption of linearity of the predictor or index function. Generalizing the derivations from Scrucca and Weisberg (2004), we derive the parameters of the logistic regression model in terms of the correlation coefficient, \( \rho \), between the regressors in the model. This parameterization allows for an investigation of near-multicollinearity in the logistic regression model and why the \emph{ceteris paribus} clause is not appropriate. Evidence from simulations confirm these findings, making the impact of near-multicollinearity different than previously thought. The results in this paper extend what Spanos and McGuirk (2002) find with regards to near-multicollinearity in the logistic regression model, but differ in that the the regression function is nonlinear and the impacts of near-multicollinearity are more related to the changes in the parameter estimates of the regressors.

The paper proceeds as follows. Section 2 reviews the specification of near-multicollinearity for the linear regression model. In Section 3, we examine the problem of near-multicollinearity in the logistic regression modeling framework. Section 4 simulates a simple logistic model to illustrate the impacts of near-multicollinearity on parameter estimates, associated standard errors and asymptotic \( t \)-ratios. Section 5 concludes and provides some general findings.

2 Background: Multicollinearity in the linear regression model

In the linear regression model \( y = X\beta + e \), perfect multicollinearity occurs when at least one of the columns of \( X \) is a linear transformation of the others. This situation occurs when the correlation (\( \rho \)) between the regressors is equal to one in absolute value. Perfect multicollinearity results in parameter identification problems since the \( (X'X) \) matrix is singular. Thus, the OLS estimators, \( \hat{\beta} = (X'X)^{-1}Xy \), and their associated variances/covariances, \( \text{cov}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} \) (where \( \hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{T-K} \)), cannot be estimated. On the other hand, if \( \text{det}(X'X) = 0 \) but \( (X'X) \) is still nondegenerate, then we have the associated problem of near-multicollinearity. Simply put, near-multicollinearity occurs when \( |\rho| = 1 \). When the regressors are highly correlated in this way, the data matrix is ill-conditioned, even though the OLS estimators still exist. However, high correlation among the regressors may lead to numerical instability in the precision and significance of the estimates of \( \beta \).

Spanos and McGuirk (2002) explore in detail the problem of near-multicollinearity in the linear regression model. They consider the following linear model:

\[
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)
\]

\footnote{See Greene (2011), page 256 for a more detailed discussion of the symptoms and consequences of near-multicollinearity.}
Spanos and Mcguirk (2002) show that the OLS estimators of $\beta$, $\sigma^2$ and $R^2$ are:

$$\hat{\beta}_0 = y - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2,$$  
$$\hat{\beta}_1 = \frac{\hat{\sigma}_{11} - \hat{\rho}\hat{\sigma}_{12}}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}},$$  
$$\hat{\beta}_2 = \frac{\hat{\sigma}_{11} - \hat{\rho}\hat{\sigma}_{12}}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}}.$$

$$\hat{\sigma}^2 = \hat{\sigma}_{11}^2 \frac{\hat{\beta}_{12} - 2\hat{\rho}\hat{\sigma}_{12}}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}} + \hat{\sigma}_{11}^2; \quad R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_{33}^2}; \quad \hat{\rho}^2 = \frac{\hat{\sigma}_{11}^2}{\hat{\sigma}_{33}^2},$$

where

$$\hat{\sigma}_{ij} = \frac{1}{T} \sum_{t=1}^{T} (z_{it} - \bar{z}_i)(z_{jt} - \bar{z}_j),$$  
i, j = 1, 2, 3;  
z_{it} = y_t;  
z_{jt} = x_{jt}.$$

They further show that the estimated variances of the above estimators for $\beta$ take the form:

$$\text{Var}(\hat{\beta}_1) = \frac{T\hat{\sigma}_{11}^2}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}}, \quad \text{Var}(\hat{\beta}_2) = \frac{T\hat{\sigma}_{11}^2}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}}, \quad \text{Var}(\hat{\beta}_3) = \frac{T\hat{\sigma}_{11}^2}{(1 - \hat{\rho}^2)\hat{\sigma}_{33}},$$

and the corresponding $t$-statistics, $t(\beta)$ are:

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{\sqrt{T}/(1 - \hat{\rho}^2)\hat{\sigma}_{33}}, \quad t(\hat{\beta}_2) = \frac{\hat{\beta}_2}{\sqrt{T}/(1 - \hat{\rho}^2)\hat{\sigma}_{33}}.$$

The conventional account of near-multicollinearity in the linear regression model is associated with the effects on $t(\hat{\beta}_1)$, $t(\hat{\beta}_2)$, $\text{Var}(\hat{\beta}_1)$, and $\text{Var}(\hat{\beta}_2)$ as $|\rho| \to 1$, ceteris paribus. The ceteris paribus clause assumes that $\hat{\beta}, \hat{\sigma}^2$, and $R^2$ are held constant, when examining the impact as $|\rho| \to 1$ on the statistics $t(\hat{\beta}_1), t(\hat{\beta}_2), \text{Var}(\hat{\beta}_1)$, and $\text{Var}(\hat{\beta}_2)$. However, because $\hat{\beta}, \hat{\sigma}^2$, and $R^2$ are all functions of $\hat{\rho}$, the changes in the associated variances and $t$-ratios will be different than the traditional account implies. Therefore the ceteris paribus assumption fails to hold.

The problems traditionally associated with near-multicollinearity in the logistic regression model are similar to those traditionally found for the linear regression model (Menard 2002). The analysis, however, is not directly applicable to the logistic regression model since the variance-covariance structure of the estimated parameters for the logistic regression model is different than that of the linear regression model and the regression is nonlinear. In the next section, the problem of near-multicollinearity in the logistic regression model is examined using the probabilistic reduction approach (Bergtold, Spanos, and Onukwugha 2010).

### 3 Multicollinearity in the logistic regression model

Let $Y_i$ represent a Bernoulli random variable with mean parameter $p$ and $X_i$ be a $(k \times 1)$ vector of regressors. The logistic regression model takes the form $Y_i = E(Y_i|X_i=x_i) = \frac{1}{1+\exp(-\beta'X_i)} + e_i$, Bergtold, Spanos, and Onukwugha (2010) show that a sufficient condition for the existence of the model is the compatibility between the conditional distribution $f_{Y_i|X_i}(Y_i|X_i;\beta)$ underlying the model and its associated inverse conditional distribution $f_{X_i|Y_i}(X_i|\theta)$. This condition is usually examined in logistic discriminant analysis, and has been explored by, among others Kay and Little (1987) and Cox and Snell (1989). The conditional distribution is usually modeled in logistic regression, whereas the inverse conditional distribution is modeled in discriminant analysis. While discriminant analysis is not commonly utilized in econometrics, the link between logistic regression and discriminant analysis provides a strong framework upon which to examine near-multicollinearity in the logistic regression model. The compatibility condition provides the basis for a systematic approach to model specification and ensures that the underlying joint distribution of $Y_i$ and $X_i$ exists, which is a requirement for the existence of a model (Bergtold, Spanos, and Onukwugha 2010). Thus, the functional specification of the
logistic regression is intertwined with the inverse conditional distribution that captures the needed probability information in the explanatory variables, which is often overlooked (e.g. see Kay and Little 1987; Arnold, Castillo, and Sarabria 1999).

The model specification approach [the probabilistic reduction (PR) approach] adopted in this paper provides a more robust specification for examining near-multicollinearity. Traditional methods of simulating conditional binary discrete choice processes (e.g. Ryan 1997; Hosmer and Lemeshow 2000) may not capture or produce the true underlying parameters of the model or may impose restrictions upon the model, thereby limiting its flexibility. For example, the predictor or index function of the logistic regression model may be nonlinear in the variables. If the inverse conditional distribution is multivariate normal with heterogeneous covariance matrix, then the predictor or index function will be a quadratic function of the explanatory variables (McFadden 1976).

Following Bergtold, Spanos, and Onukwugha (2010), the logistic regression model can be expressed as:

\[ Y_i = [1 + \exp(-\eta(X_i; \beta))]^{-1} + \epsilon_i \]  

(1)

where \( \eta(X_i; \beta) = \ln \left( \frac{f_{X \mid Y}(X_i; \theta)}{f_{X \mid \theta}(X_i; \theta_0)} \right) + \kappa, \kappa = \ln \left( \frac{p}{1-p} \right), \text{ and } p = \text{Prob}(Y_i = 1). \) Equation (1) is usually reparametrized such that \( \beta = \beta(\rho), \) where \( \rho \) is the correlation between the explanatory variables when \( f_{X \mid Y}(X_i; \theta) \) is multivariate normal. Thus, one can show that the traditional account of near-multicollinearity and the invocation of the ceteris paribus clause are unattainable, as in the case for the linear model.

Following the commonly used specification in the literature of linearity in the variables and parameters of the function \( \eta(\cdot), \) assume that \( f_{X \mid Y}(X_i; \theta) \) is multivariate normal with a homogenous covariance matrix. That is:

\[ f_{X \mid Y}(X_i; \theta) = (2\pi)^{-\frac{k}{2}} |V|^\frac{1}{2} \exp \left\{ -\frac{1}{2} ((X_i - \mu_i)^\prime V^{-1} (X_i - \mu_i)) \right\} \]  

(2)

where \( \mu_i \) is the mean vector for \( X_i \) conditional on \( Y_i = j, \) and \( V \) is the covariance matrix for \( X_i. \)

Following Day and Kerridge (1967), a more general distributional functional form \( f_{X \mid Y}(X_i; \theta) = \alpha \exp \left\{ -\frac{1}{2} (X_i - \mu_i)^\prime V^{-1} (X_i - \mu_i) \right\} \delta(X_i) \) can be expressed, as well. When \( \delta(X_i) = 1, f_{X \mid Y}(X_i; \theta) \) is distributed multivariate normal. However, if \( \delta(X_i) = 1, \) this functional form can represent a wider range of alternatives, including skewed distributions (Byth and McLachlan 1980). Thus, the results from this paper should hold for more general specifications of the logistic regression model than when the inverse conditional model is multivariate normal. This flexibility may be particularly important in economics when considering cases with a skewed distribution.

With the above assumptions, the

\[ \ln \left( \frac{f_{X \mid Y}(X_i; \theta_i)}{f_{X \mid \theta}(X_i; \theta_0)} \right) = \left[ (\mu_i - \mu_0)^\prime V^{-1} \right] X_i + \left[ \frac{1}{2} \mu_0^\prime V^{-1} \mu_0 - \frac{1}{2} \mu_i^\prime V^{-1} \mu_i \right], \]  

giving:

\[ \eta(X_i; \beta) = \beta_0 + \beta X_i, \]  

(3)

where \( \beta_0 = \ln \left( \frac{p}{1-p} \right) + \left[ \frac{1}{2} \mu_0^\prime V^{-1} \mu_0 - \frac{1}{2} \mu_i^\prime V^{-1} \mu_i \right]; \) and \( \beta = (\mu_i - \mu_0)(V^{-1}). \)

Without loss of generality, standardize \( X \) by dividing each element by its respective standard deviation, so that \( \text{Var}(X_{m}) = 1 \) for \( m = 1, \ldots, k. \) Then repartition the covariance matrix, \( V \) so that \( V = \begin{bmatrix} 1 & \rho_m^- \rho_m^- \rho_m^- P_{m} \end{bmatrix} \), where \( \text{cov}(X_m, X) = \rho_m \) is the correlation between \( X_m \) and \( X; \rho_m \) is the \((k-1 \times 1)\) vector of correlation coefficients.

Arnold, Castillo, and Sarabria (1999) note that linearity in the parameters and variables for the predictor or index function requires severe restrictions and that many of the applied logistic regression models in the literature may be questionable in light of this information. Furthermore, note that the covariance matrix \( V \) is not conditional on \( Y \) for this case.
between \(X_m\) and all the other explanatory variables; and \(P_m\) is the remainder of the covariance matrix \(V\), with 1 s along the diagonal and correlation coefficients on the off-diagonal elements between the remaining regressors. The repartition of \(V\) allows us to isolate the parameterization of the coefficient \(\beta_m\) for \(m=1, \ldots, k\). Invoking the Schur lemma following Spanos and McGuirk (2002):

\[
V^{-1} = \begin{pmatrix}
\gamma^{-1} & -\gamma^{-1} \rho_m' P_m^{-1} \\
-\gamma^{-1} \rho_m P_m^{-1} & P_m^{-1} \rho_m' P_m^{-1} + P_m^{-1}
\end{pmatrix}
\]

(4)

where \(\gamma = 1 - \rho_m' P_m^{-1} \rho_m\). Plugging equation (4) into the formula for \(\beta\) in equation (3), gives:

\[
\beta = \beta_{\text{MLE}} = \begin{pmatrix}
\gamma^{-1} (\mu_{1,m} - \mu_{0,m}) - \gamma^{-1} \rho_m' P_m^{-1} (\mu_{1,m} - \mu_{0,m}) \\
-\gamma^{-1} \rho_m P_m^{-1} (\mu_{1,m} - \mu_{0,m}) + [P_m^{-1} \rho_m' P_m^{-1} + P_m^{-1}] (\mu_{1,m} - \mu_{0,m})
\end{pmatrix}
\]

(5)

where \(\mu_{j,m}\) is the \(m\)th element of \(\mu_j\) for \(j=1, 0, 1\); and \(\mu_{0,m}\) represents the \((k-1 \times 1)\) vector of \(\mu_j\) without \(\mu_{0,m}\) for \(j=0, 1\). Equation (5) illustrates that \(\beta_{\text{MLE}}\) for \(m=1, \ldots, k\) are functions of \(\rho_m\). Furthermore, \(\beta_{\text{MLE}}\) is affected by the correlation between other explanatory variables via \(P_m^{-1}\). Thus, the ceteris paribus clause is unattainable in the logistic regression model because \(\beta\) is a function of \(\rho\).

Following Schaefer, Roi, and Wolfe (1984) and Gourieroux (2000), let \(\overline{\text{Prob}}(y_j=1|X; \beta) = F(\beta_{\text{MLE}} X_j) = F\), so that the asymptotic variance estimator of \(\beta\) can be written as \(\text{Var}(\beta_{\text{MLE}}) = (X \Omega X)\)\(^{-1}\), where \(\Omega = \text{diag} [F(1 - F)]\)\(^n\) and \(X\) is the \((n \times k)\) dimensional matrix of regressors. Then, \(\text{Var}(\beta_{\text{MLE}})\) is a function of \(\rho\) through \(F\). Consider the case where \(|\rho_{m,n}|=1\) for some \(s=1, \ldots, k\). If \(|\rho_{m,n}|\rightarrow\infty\), then \(F\) will approach 0 or 1, depending on the sign of \(\beta_{\text{MLE}}\). This implies that \(F(1 - F)\rightarrow0\) as \(|\rho_{m,n}|\rightarrow\infty\), suggesting that the \(\text{Var}(\beta_{\text{MLE}})\) may follow the shifts in \(\beta_{\text{MLE}}\) as \(|\rho_{m,n}|\rightarrow1\) for some \(s=1, \ldots, k\). While changes in \(\beta\) with respect to \(\rho_m\) can be plotted using equation (5), \(\text{Var}(\beta_{\text{MLE}})\) and asymptotic t-ratios \(|t(\beta_{\text{MLE}})|\) need to be examined using simulation methods as no closed form expressions exist for these statistics (Gourieroux 2000). \(^3\\)

4 Simulation

This section simulates a simple logistic regression model with two covariates to illustrate the impacts of near multicollinearity on parameter estimates, associated standard errors, and asymptotic t-ratios.

4.1 Logistic regression model simulated

Assume that \(X_i=(x_{i1}, x_{i2})'\) conditional on \(Y_i=j\) is bivariate normal with homogenous covariance matrix. That is:

\[
\begin{pmatrix}
x_{i1} \\
x_{i2}
\end{pmatrix} | Y_i=j \sim N \left( \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \right)
\]

where \(\rho = \text{corr}(x_{i1}, x_{i2})\). The inverse conditional distribution function can be expressed as (see Spanos 1986: p 119–121 and Bergtold, Spanos, and Onukwugha 2010):

\[
f_{X_i|Y_i=j}(x_i, x_i; \theta) = \left(1 - \rho^2\right)^{\frac{1}{2}} \exp \left\{ \frac{1}{2} \frac{1}{1 - \rho^2} \left[ \frac{x_{i1} - \mu_{i1}}{\sigma_1} \right]^2 + \left( \frac{x_{i2} - \mu_{i2}}{\sigma_2} \right)^2 - 2 \rho \frac{x_{i1} - \mu_{i1}}{\sigma_1} \frac{x_{i2} - \mu_{i2}}{\sigma_2} \right\}
\]

(6)

\(^3\\) We examine asymptotic t-ratios in this paper, however, similar results hold for other asymptotic test statistics, such as the likelihood ratio or Wald test statistics.
Let \( \theta_j = \frac{1}{2} \left( 1 - \rho^2 \right) \left[ \frac{(x_j - \mu_{1,j})^2}{\sigma_1^2} + \frac{(x_j - \mu_{2,j})^2}{\sigma_2^2} - 2\rho \frac{(x_j - \mu_{1,j})(x_j - \mu_{2,j})}{\sigma_1 \sigma_2} \right] \), so that equation (6) can be rewritten as:

\[
 f_{X',y}(x_1, x_2; \theta_j) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp\{ \theta_j \}
\]

Then:

\[
 \ln \left( \frac{f_{X',y}(x_1, x_2; \theta_j)}{f_{X',y}(x_1, x_2; \theta_0)} \right) = \frac{\exp(\theta_j)}{\exp(\theta_0)} = \theta_j - \theta_0
\]

Combining equations (6) and (8) for \( j = 0, 1 \) simplifying, and rearranging terms gives:

\[
 \begin{align*}
 \theta_j - \theta_0 &= \frac{1}{2(1 - \rho^2)} \left[ \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma_1^2} + \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma_2^2} - 2\rho \frac{(\mu_{1,j} - \mu_{2,j})}{\sigma_1 \sigma_2} \right] x_1 \\
 &+ \frac{2\rho(\mu_{1,j} - \mu_{2,j})}{\sigma_1^2} - \frac{(\mu_{1,j} - \mu_{2,j})}{\sigma_1^2} \right] x_2
\end{align*}
\]

Using the relationships in equations (1), (6), (8) and (9), the corresponding logistic regression model can be written as:

\[
 Y = [1 + \exp\{ -\beta_0 - \beta_1 x_1 - \beta_2 x_2 \}]^{-1} + u,
\]

From equations (3) and (9), it follows that:

\[
 \begin{align*}
 \beta_0 &= \ln \left( \frac{p}{1-p} \right) \frac{1}{2(1 - \rho^2)} \left[ \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma_1^2} + \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma_2^2} - 2\rho \frac{(\mu_{1,j} - \mu_{2,j})}{\sigma_1 \sigma_2} \right] \\
 \beta_1 &= \frac{1}{2(1 - \rho^2)} \left[ \frac{2(\mu_{1,j} - \mu_{2,j})}{\sigma_1^2} - \frac{2\rho(\mu_{1,j} - \mu_{2,j})}{\sigma_1 \sigma_2} \right] \\
 \beta_2 &= \frac{1}{2(1 - \rho^2)} \left[ \frac{2(\mu_{1,j} - \mu_{2,j})}{\sigma_2^2} - \frac{2\rho(\mu_{1,j} - \mu_{2,j})}{\sigma_1 \sigma_2} \right]
\end{align*}
\]

### 4.2 Simplifications

The simplifications below reduce the dimensionality of the problem being simulated without necessarily limiting the generalizability of the analysis, only affecting scaling of the results.

#### 4.2.1 Admissible parameter values

In light of the statistical reparametizations shown above, it is important to determine the range of admissible values for which the model exists. The importance of this is emphasized by Spanos and McGuirk (2002). Let \( \Sigma \)
denote the the variance-covariance matrix of the inverse conditional distribution, then \( \Sigma = \begin{pmatrix} \sigma^2_1 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma^2_2 \end{pmatrix} \).

Following this parametrization, the above parameters \( \{ \beta_0, \beta_1, \beta_2 \} \) are valid when (Spanos 1986):

\[
det(\Sigma) = \sigma^2_1 \sigma^2_2 (1-\rho^2) > 0
\]

Condition (14) implies that the inverse conditional distribution given by equation (6) will not exist when \( \Sigma \preceq 0 \). Thus, during simulations if \( \det(\Sigma) = 0 \) then zeros are assigned for all the parameter values and associated statistics for that simulation run.

4.2.2 Simplification 1

Standardize the variables \( x_1 \) and \( x_2 \) by dividing them by \( \sigma_m \) for \( m=1, 2 \), so that \( \sigma_1 = \sigma_2 = 1 \). It follows that equation (14) becomes \( \det(\Sigma) = (1-\rho^2) > 0 \). Thus, the admissible parameter range for \( \rho \) is \( \rho \in (-1, 1) \). Following this simplification, the parameters \( \{ \beta_0, \beta_1, \beta_2 \} \) from equations (11) to (13) reduce to:

\[
\begin{align*}
\beta_0 &= \ln \left( \frac{p}{1-p} \right) - \frac{1}{2(1-\rho^2)} \left[ \mu^2_{1,0} - 2 \rho (\mu_{1,0} \mu_{2,0}) + \mu^2_{2,0} \right] \\
\beta_1 &= \frac{1}{2(1-\rho^2)} \left[ 2 (\rho_{1,2} - \mu_{1,0}) - 2 \rho (\mu_{2,0} - \mu_{1,0}) \right] \\
\beta_2 &= \frac{1}{2(1-\rho^2)} \left[ 2 (\mu_{2,1} - \mu_{1,0}) - 2 \rho (\mu_{1,1} - \mu_{1,0}) \right]
\end{align*}
\]

4.2.3 Simplification 2

Further, consider the following mean deviation forms of \( x_{1,i} \) and \( x_{2,i} \) by dividing them by \( \mu_{m,0} \) for \( m=1, 2 \), where \( \mu_{m,0} = E(x_m \mid Y=0) \) for \( m=1, 2 \). Then \( \tilde{\mu}_{m,0} = E(\bar{x}_m \mid Y=0) = 0 \) for \( m=1, 2 \), making \( \mu_{m,1} = E(\bar{x}_m \mid Y=1) = \mu_{m,1} - \mu_{m,0} \) for \( m=1, 2 \). This further reduces the dimensionality of the parameters \( \{ \beta_0, \beta_1, \beta_2 \} \) from equations (15) to:

\[
\begin{align*}
\beta_0 &= \ln \left( \frac{p}{1-p} \right) - \frac{1}{2(1-\rho^2)} \left[ \bar{\mu}^2_{1,1} - 2 \rho (\bar{\mu}_{1,1} \bar{\mu}_{2,1}) + \bar{\mu}^2_{2,1} \right] \\
\beta_1 &= \frac{1}{2(1-\rho^2)} \left[ 2 \bar{\mu}_{2,1} - 2 \rho \bar{\mu}_{1,1} \right] \\
\beta_2 &= \frac{1}{2(1-\rho^2)} \left[ 2 \bar{\mu}_{1,2} - 2 \rho \bar{\mu}_{1,1} \right]
\end{align*}
\]

4.3 Monte Carlo simulation procedure

To generate binary choice data, we use a two stage process following Bergtold, Spanos, and Onukwugha (2010) and Scrucca and Weisberg (2004). First, using a binomial random number generator, realizations of the vector stochastic process \( \{ Y_i, i=1, \ldots, n \} \) are generated. Second, using \( Y \) as the conditioning variable, the vector stochastic process of predictors \( \{ X_i, i=1, \ldots, n \} \) is generated using the bivariate normal inverse conditional distribution \( f_{X_i \mid Y_j} (X_i, \theta) \) using appropriate random number generators. Given these simplifications, Monte Carlo simulations are conducted for different mean pair combinations for each value of \( \rho \) examined, where \( \rho \) was varied between

\[4\] Monte Carlo simulations were conducted for a large number of different mean pairs. All simulations and graphics were conducted in MATLAB. These results are available upon request from the authors.
–1 and 1 by increments of 0.005 for the plots in Figure 1, with specific values reported in Tables 1 and 2. The means of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, their associated standard errors [se($\hat{\beta}_0$), se($\hat{\beta}_1$), se($\hat{\beta}_2$)], and asymptotic t-ratios [t($\hat{\beta}_0$), t($\hat{\beta}_1$), t($\hat{\beta}_2$)] were calculated across runs.

4.4 Simulation results and discussion

Simulation results for the mean pairs $(\bar{\mu}_1, \bar{\mu}_2)=(-1, -0.5), (-1, 1), (-0.5, -0.25), (0.25, 0.5), (0.5, 1), \text{ and } (0.5, 2)$ for 15 different values of $\rho$ ranging from –0.99 to 0.99 for $\hat{\beta}_i$ and se($\hat{\beta}_i$) for $i=1, 2$ are provided in

A  $(\bar{\mu}_1, \bar{\mu}_2)=(-1, -0.5)$

B  $(\bar{\mu}_1, \bar{\mu}_2)=(-1, 1)$

C  $(\bar{\mu}_1, \bar{\mu}_2)=(-0.5, -0.25)$

D  $(\bar{\mu}_1, \bar{\mu}_2)=(0.25, 0.5)$

Figure 1 (continued)
Tables 1 and 2. Figure 1 provides plots of the mean estimates of $\hat{\beta}$, $\hat{\beta}_1$, $\hat{\beta}_2$, se($\hat{\beta}$), se($\hat{\beta}_1$), se($\hat{\beta}_2$), and $\tau(\hat{\beta})$ as $|\rho| \to 1$ for the mean pairs mentioned above. The mean pairs presented here were selected as representative of the general findings of the large number of runs conducted. Many of the results from the simulation for alternative mean pairs not shown follow the general patterns presented.

In general, contrary to the traditional account, the plots of $\hat{\beta}$, $\hat{\beta}_1$, $\hat{\beta}_2$, (and $\hat{\beta}_\tau$) for all the mean pairs simulated were not constant as $|\rho| \to 1$. Furthermore, as illustrated by the standard error estimates in Figure 1, se($\hat{\beta}$) did not always substantially inflate as $|\rho| \to 1$. Simulation studies by Ryan (1997) support this finding. If se($\hat{\beta}$) did become inflated, it was not necessarily distinguishable if it would increase as $\rho \to -1$, $\rho \to 1$ or both. For example, se($\hat{\beta}_1$) and se($\hat{\beta}_2$) for the mean pair $(\mu_1, \mu_2) = (0.5, 2)$ in Figure 1(F) inflate as $|\rho| \to 1$ as the traditional account suggests. However, for the mean pair $(\mu_1, \mu_2) = (-1, -0.5)$ in Figure 1(A), se($\hat{\beta}_1$) and se($\hat{\beta}_2$) increases only as $\rho \to -1$; and for the mean pair $(\mu_1, \mu_2) = (-1, 1)$ in Figure 1(B), se($\hat{\beta}_1$) and se($\hat{\beta}_2$) only increase as $\rho \to 1$. These results are supported by examining the corresponding changes in the the standard error estimates in Tables 1 and 2 for the associated mean pairs, as well. This is contrary to the traditional account of near-multicollinearity identified by Schaefer, Roi, and Wolfe (1984). A possible explanation for the difference is that traditionally, it is assumed $\beta$ remains constant as $|\rho| \to 1$ (i.e. the ceteris paribus clause) when in fact this assumption is not attainable since $\beta$ is a function of $\rho$. Of particular interest is that in general, the numerical results in Tables 1 and 2, as well as the plots in Figure 1 provide evidence that the estimates of se($\hat{\beta}$) tend to follow the changes in $\hat{\beta}$. This is further corroborated by examining changes in $\tau(\hat{\beta}) = \hat{\beta} / se(\hat{\beta})$ as $|\rho| \to 1$. In each of the plots in Figure 1, when $|\hat{\beta}|$ increases, the $|se(\hat{\beta})|$ increases; and for most of these cases, $\tau(\hat{\beta})$ was still statistically significant. Thus, $\tau(\hat{\beta}) > 1$ in many cases, providing evidence that $\hat{\beta}$ may be increasing at a faster rate than the se($\hat{\beta}$) as $|\rho| \to 1$. The implications of multicollinearity on statistical significance are unclear a priori.

The ceteris paribus clause has a strong impact on the implications from near-multicollinearity for the logistic regression model. When it is shown that $\beta$ and other related statistics are functions of $\rho$, the results seem to indicate that the traditional implications may not hold. Figure 1 provides evidence that se($\hat{\beta}$) may not become inflated and if they do it may not have an appreciable impact on statistical significance. For many of the plots examined using different mean pairs, $\hat{\beta}$ remained statistically significant at the 5% level as $|\rho| \to 1$, while the magnitude of the test statistic did fluctuate as $|\rho| \to 1$. We expect the same results would
hold for other asymptotic test statistics, as well. The impact of the results was dependent upon the mean pair combination simulated and may be somewhat uncertain for a given dataset a priori, even if high correlation among the regressors is expected.

We conjecture from these findings that the systematic implications from near-multicollinearity may result from changes in $\hat{\beta}$ as $|\rho|\to 1$ in the logistic regression model. For many of the simulations of different mean pairs, as $|\rho|\to 1$, $|\hat{\beta}|\to\infty$, or got very large in magnitude. As seen in Tables 1 and 2, the inflation of $\hat{\beta}$ tended to be more severe as the absolute difference between $\bar{\mu}_{11}$ and $\bar{\mu}_{21}$ increased, indicating the scale of the data plays a role on the impact of near multicollinearity. Scaling tends to be more of a numerical issue and did give rise to problems during estimation when the absolute difference between $\bar{\mu}_{11}$ and $\bar{\mu}_{21}$ was larger [e.g. the simulation results were more erratic for mean pair $(\bar{\mu}_{11}, \bar{\mu}_{21})=(-3, -1)$ in Tables 1 and 2 as $|\rho|\to 1$]. Given equation (10), the inflation in $\hat{\beta}$ is expected as the denominator is a function of $(1-\rho^2)$, while the numerator is only a function of $\rho$. The findings here are different than the findings for the linear regression model presented by Spanos and McGuirk (2002) in that the implications of near-multicollinearity for the logistic regression model seem to be tied up with $\hat{\beta}$. Near-multicollinearity affects both the standard errors of $\hat{\beta}$ and the associated $t$-ratios as in the linear regression model, but changes in the standard errors seem to follow the changes in $\hat{\beta}$.

Table 1: Simulation results for $\hat{\beta}$ and $se(\hat{\beta})$ for different mean pairs $(\bar{\mu}_{11}, \bar{\mu}_{21})$ as $\rho$ changes.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$(\bar{\mu}<em>{11}, \bar{\mu}</em>{21})$</th>
<th>$\hat{\beta}$</th>
<th>$se(\hat{\beta})$</th>
<th>$(\bar{\mu}<em>{11}, \bar{\mu}</em>{21})$</th>
<th>$\hat{\beta}$</th>
<th>$se(\hat{\beta})$</th>
<th>$(\bar{\mu}<em>{11}, \bar{\mu}</em>{21})$</th>
<th>$\hat{\beta}$</th>
<th>$se(\hat{\beta})$</th>
<th>$(\bar{\mu}<em>{11}, \bar{\mu}</em>{21})$</th>
<th>$\hat{\beta}$</th>
<th>$se(\hat{\beta})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.99</td>
<td>(-3, -1)</td>
<td>-17.07</td>
<td>1.326</td>
<td>(-3, -1)</td>
<td>-41.63</td>
<td>1601.12</td>
<td>(-1, -1)</td>
<td>-0.50</td>
<td>0.51</td>
<td>(-0.5, -0.25)</td>
<td>-51.23</td>
<td>20.84</td>
</tr>
<tr>
<td>-0.95</td>
<td>15.96</td>
<td>2999.90</td>
<td>-17.41</td>
<td>3.68</td>
<td>-0.51</td>
<td>0.23</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>-0.90</td>
<td>-16.66</td>
<td>2376</td>
<td>-7.84</td>
<td>0.67</td>
<td>-0.53</td>
<td>0.17</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>-0.85</td>
<td>-31.89</td>
<td>137e10</td>
<td>-3.92</td>
<td>0.26</td>
<td>-0.56</td>
<td>0.12</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>-0.65</td>
<td>-7.51</td>
<td>1.88</td>
<td>-2.31</td>
<td>0.15</td>
<td>-0.60</td>
<td>0.10</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>-0.50</td>
<td>-4.93</td>
<td>0.63</td>
<td>-1.68</td>
<td>0.11</td>
<td>-0.67</td>
<td>0.089</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>-0.25</td>
<td>-3.54</td>
<td>0.32</td>
<td>-1.21</td>
<td>0.90</td>
<td>-0.81</td>
<td>0.084</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0</td>
<td>-3.05</td>
<td>0.25</td>
<td>1.0001</td>
<td>0.082</td>
<td>-1.01</td>
<td>0.088</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.25</td>
<td>-2.97</td>
<td>0.23</td>
<td>-0.94</td>
<td>0.082</td>
<td>-1.34</td>
<td>0.10</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.50</td>
<td>-3.38</td>
<td>0.26</td>
<td>-1.00</td>
<td>0.090</td>
<td>-2.01</td>
<td>0.14</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.65</td>
<td>-4.15</td>
<td>0.34</td>
<td>-1.18</td>
<td>0.10</td>
<td>-2.89</td>
<td>0.20</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.80</td>
<td>-6.30</td>
<td>0.65</td>
<td>-1.69</td>
<td>0.14</td>
<td>-5.07</td>
<td>0.41</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.90</td>
<td>-13.45</td>
<td>3.48</td>
<td>-2.91</td>
<td>0.21</td>
<td>-10.94</td>
<td>1.74</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.95</td>
<td>-82.44</td>
<td>1.03e14</td>
<td>-5.44</td>
<td>0.35</td>
<td>-50.11</td>
<td>5.32e10</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
<tr>
<td>0.99</td>
<td>-31.86</td>
<td>1.01e4</td>
<td>-26.16</td>
<td>2.48</td>
<td>-32.10</td>
<td>9507.56</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
<td>-17.07</td>
<td>0.51</td>
<td>-0.50</td>
</tr>
</tbody>
</table>

We conjecture from these findings that the systematic implications from near-multicollinearity may result from changes in $\hat{\beta}$ as $|\rho|\to 1$ in the logistic regression model. For many of the simulations of different mean pairs, as $|\rho|\to 1$, $|\hat{\beta}|\to\infty$, or got very large in magnitude. As seen in Tables 1 and 2, the inflation of $\hat{\beta}$ tended to be more severe as the absolute difference between $\bar{\mu}_{11}$ and $\bar{\mu}_{21}$ increased, indicating the scale of the data plays a role on the impact of near multicollinearity. Scaling tends to be more of a numerical issue and did give rise to problems during estimation when the absolute difference between $\bar{\mu}_{11}$ and $\bar{\mu}_{21}$ was larger [e.g. the simulation results were more erratic for mean pair $(\bar{\mu}_{11}, \bar{\mu}_{21})=(-3, -1)$ in Tables 1 and 2 as $|\rho|\to 1$]. Given equation (10), the inflation in $\hat{\beta}$ is expected as the denominator is a function of $(1-\rho^2)$, while the numerator is only a function of $\rho$. The findings here are different than the findings for the linear regression model presented by Spanos and McGuirk (2002) in that the implications of near-multicollinearity for the logistic regression model seem to be tied up with $\hat{\beta}$. Near-multicollinearity affects both the standard errors of $\hat{\beta}$ and the associated $t$-ratios as in the linear regression model, but changes in the standard errors seem to follow the changes in $\hat{\beta}$. 


Furthermore, near-multicollinearity may have a direct effect on the magnitude of $\hat{\beta}$, related to the difference in scale of the explanatory variables involved. These differences highlight the fact that multicollinearity in nonlinear models, such as the logistic regression model, may need to be further examined, as they may differ from the consequences of near-multicollinearity for the linear regression model.

In addition, as $|\rho| \rightarrow 1$, changes in $\hat{\beta}$ will have subsequent impacts on odds ratios and marginal effects. Odd ratios, which may be calculated as $\exp(\hat{\beta})$ may grow exponentially large as $|\rho| \rightarrow 1$, because they are a direct transformation of $\hat{\beta}$. Marginal effects of the logistic regression model in equation (1) can be expressed as (see Bergtold, Spanos, and Onukwugha 2010):

$$\frac{\partial Y}{\partial X_i} = F_i(1 - F_i) \frac{\partial \eta(X_i; \beta)}{\partial X_i}$$

where all variables and functions are as previously defined. As seen in equation (17), the marginal effects are also functions of $\rho$, since they are functions of $\beta$ through $F_i$ and $\eta(X_i; \beta)$. This implies that changes in the estimates of $\hat{\beta}$ from near-multicollinearity impact marginal effect estimates, as well.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$(\mu_{11}, \mu_{21}) = (-3, -1)$</th>
<th>$(\mu_{11}, \mu_{21}) = (-1, -0.5)$</th>
<th>$(\mu_{11}, \mu_{21}) = (-1, 1)$</th>
<th>$(\mu_{11}, \mu_{21}) = (-0.5, -0.25)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_1$</td>
<td>$s(\hat{\beta}_1)$</td>
<td>$\hat{\beta}_2$</td>
<td>$s(\hat{\beta}_2)$</td>
</tr>
<tr>
<td>-0.99</td>
<td>-16.97</td>
<td>2.01e5</td>
<td>-41.44</td>
<td>1656.25</td>
</tr>
<tr>
<td>-0.95</td>
<td>-15.58</td>
<td>3172.13</td>
<td>-17.11</td>
<td>3.62</td>
</tr>
<tr>
<td>-0.90</td>
<td>-15.81</td>
<td>3.06e4</td>
<td>-7.57</td>
<td>0.66</td>
</tr>
<tr>
<td>-0.80</td>
<td>-28.53</td>
<td>1.31e10</td>
<td>-3.64</td>
<td>0.25</td>
</tr>
<tr>
<td>-0.65</td>
<td>-6.06</td>
<td>1.58</td>
<td>-2.01</td>
<td>0.14</td>
</tr>
<tr>
<td>-0.50</td>
<td>-3.52</td>
<td>0.50</td>
<td>-1.34</td>
<td>0.11</td>
</tr>
<tr>
<td>-0.25</td>
<td>-1.91</td>
<td>0.24</td>
<td>-0.81</td>
<td>0.084</td>
</tr>
<tr>
<td>0</td>
<td>-1.02</td>
<td>0.17</td>
<td>-0.50</td>
<td>0.076</td>
</tr>
<tr>
<td>0.25</td>
<td>-0.27</td>
<td>0.15</td>
<td>-0.27</td>
<td>0.076</td>
</tr>
<tr>
<td>0.50</td>
<td>0.68</td>
<td>0.18</td>
<td>-0.0038</td>
<td>0.084</td>
</tr>
<tr>
<td>0.65</td>
<td>1.67</td>
<td>0.24</td>
<td>0.26</td>
<td>0.096</td>
</tr>
<tr>
<td>0.80</td>
<td>4.01</td>
<td>0.51</td>
<td>0.85</td>
<td>0.13</td>
</tr>
<tr>
<td>0.90</td>
<td>10.89</td>
<td>2.93</td>
<td>2.12</td>
<td>0.20</td>
</tr>
<tr>
<td>0.95</td>
<td>74.06</td>
<td>9.30e13</td>
<td>4.66</td>
<td>0.33</td>
</tr>
<tr>
<td>0.99</td>
<td>31.21</td>
<td>1.42e6</td>
<td>25.38</td>
<td>2.43</td>
</tr>
</tbody>
</table>

Table 2: Simulation results for $\hat{\beta}_1$ and $s(\hat{\beta}_1)$ for different mean pairs $(\mu_{11}, \mu_{21})$ as $\rho$ changes.
5 Conclusion

In this paper, we derive and show that the parameters of the logistic regression model, where the predictor is linear in both the variables and parameters, are functions of $\rho$. Thus, our analysis calls into question the ceteris paribus clause in the traditional account of near-multicollinearity for the logistic regression model and shows that it is inappropriate. Monte Carlo simulations provide evidence that the parameters and their associated variances and $t$-ratios may be different than the traditional account implies. The implications of our findings are that $\hat{\beta}$, $\text{se}(\hat{\beta})$, and $r(\hat{\beta})$ can fluctuate in a non-systematic, non-monotonic way as $|\rho|\rightarrow1$; that changes to relevant statistics are data specific and uncertain; that the impact of near-multicollinearity is likely centered on the estimates of $\beta$; and the impact on substantive inferences does not follow what the traditional account would imply. As the probit model may seen as a close approximation of the logistic regression model (Amemiya 1981), the results here may extend to that model, as well, though this requires further exploration. In addition, future research needs to extend the analysis to logistic regression models with predictors that are nonlinear in the regressors and linear predictors that incorporate binary covariates.

References


Supplemental Material: The online version of this article (DOI: 10.1515/snde-2013-0052) offers supplementary material, available to authorized users.