Weak Identification in Nonlinear Econometric Models

Lee C. Adkins

August 16, 2022

Abstract

The Belsley, Kuh, and Welch Belsley et al. (1980) diagnostics are applied to the precision estimator in nonlinear models and the properties are explored via simulation. The performance of the diagnostics is examined specifically with the ordered probit model. The performance is examined under several parameterizations and under different degrees of collinearity. The results suggest that the identification of the model's parameters is indeed related to collinearity of the data and by the parametric functional form of the model itself. The BKW diagnostics appear to be useful in the sense that serious problems with identification can be detected quite easily using the built in vif function of gretl.

Introduction

Adkins et al. (2015) provide gretl functions to compute the Belsley et al. (1980) collinearity diagnostics that could be used to detect weak identification in linear and nonlinear models. In this companion, I explore how well the BKW measures perform in nonlinear contexts. In linear models the rules-of-thumb suggested by BKW indicate how severe the collinearity problem is and can tell the user whether the variance of a particular coefficient estimator is adversely affected by collinear data. The diagnostics are now available in the base versions of GRETL, and preliminary results on the performance of the rules of thumb, which also include another example, sample selection, can be found in Adkins (2017).

Hill and Adkins (2001) Consider collinearity in linear and nonlinear context and I follow salient points from their discussion below. In nonlinear models the BKW diagnostics are computed based on a rescaled variance covariance matrix, and are comparable to what is currently used in linear models estimated by least squares. In nonlinear models the covariance

matrix depends on parameters as well as data and the interaction of the two can affect the identification of parameters.

In nonlinear models weak identification can lead to excessive numbers of iterations, to non-convergence of the algorithm, or can cause the covariance estimate (e.g., negative of the inverse Hessian) to be negative definite; if this happens the numerical algorithm fails to locate the optimum. When any of these happens, one would like to know whether the problem happens due to data or because of its interaction with parameters. A corrective strategy may materialize; get better data, adopt a simpler model, or impose additional constraints on the model.

The generalizability of the diagnostics to nonlinear models is studied using simulations of two nonlinear models. The average condition numbers and variance decompositions for each model are presented. The output also includes standard deviations from the Monte Carlo for each scenario. In this way, the relative variability due to parameter estimation can be measured. If the estimates play a fairly modest role, then much of the weak identification of the nonlinear estimator can be attributed to the data, which is similar to the situation in linear models. The goal is to form a foundation for the expanded use of the BKW condition number and variance decomposition analysis to the nonlinear world.

The paper is organized as follows. First, I review collinearity and the BKW diagnostics in linear models. In section 3 the extension of nonlinear models is reconsidered. In section 4 I examine how the diagnostics perform using an example: ordered probit. Within this section a set of simulations is conducted and the performance of the BKW diagnostics are studied. I show that the diagnostics can be quite useful in detecting problems of identification in ordered probit that are due to either collinearity of the data or to parameterization of the model.

1 Linear Model

Denote the linear regression model as

$$y = X\beta + u$$

where y is a $n \times 1$ vector of observations on the dependent variable, X is a $n \times k$ non-stochastic matrix of observations on k explanatory variables, β is a $k \times 1$ vector of unknown parameters, and u is the $n \times 1$ vector of uncorrelated random errors, with zero means and constant variances,σ^2. In the general linear model exact, or perfect, collinearity exists when the columns of X, denoted x_i, $i = 1, \ldots, K$, are linearly dependent. That is, if there is at least one relation of the form $c_1x_1 + c_2x_2 + \cdots + c_Kx_k = 0$, where the c_i are constants, not all equal to zero. In this case the column rank of X is less than k, and the normal equations

$X^TX\beta = X^Ty$ do not have a unique solution, and least squares estimation breaks down. Unique best linear unbiased estimators do not exist for all K parameters. However, even in this most severe of cases, all is not lost.

Exact collinearity is rare, and easily recognized. More frequently, one or more linear combinations of explanatory variables are nearly exact, so that $c_1x_1 + c_2x_2 + \cdots + c_Kx_k \approx 0$. We now examine the consequences of such near exact linear dependencies. The collinearity problem can more broadly be viewed as an identification problem. If collinearity is exact, then identification fails. If collinearity is strong, then the parameters are identified, but they may be estimated imprecisely given the data on hand.

1.1 Diagnosing Collinearity using the Eigenvalues and Eigenvectors of X^TX

The $k \times k$ matrix X^TX is symmetric. For symmetric matrices their exists an orthonormal $k \times k$ matrix C such that

$$C^TX^TXC = \Lambda \tag{1}$$

where Λ is a diagonal matrix with the real values $\lambda_1, \lambda_2, \ldots, \lambda_k$ on the diagonal. An orthonormal matrix, sometimes also called an orthogonal matrix, has the property that $C^T = C^{-1}$, so that $C^TC = CC^T = I_k$, where I_k is a $k \times k$ identity matrix. The columns of the matrix C, denoted c_i, are the eigenvectors (or characteristic vectors) of the matrix, and the real values λ_i are the corresponding eigenvalues (or characteristic roots). It is customary to assume that the columns of C are arranged so that the eigenvalues are ordered by magnitude, $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_k$.

Silvey (1969) popularized the use of eigenvalues to diagnose collinearity, and Belsley et al. (1980) [hereinafter BKW] refined, and improved, the analysis. The $n \times k$ matrix $Z = XC$ is called the matrix of principal components of X. The i^{th} column of Z, z_i, is called the i^{th} principal component. From equation (1) z_i has the property that $z_i^Tz_i = \lambda_i$. If the characteristic root $\lambda_i = 0$, then $z_i = Xc_i = 0$; we have an exact linear relation among the columns of X, and thus exact collinearity. If $\texttt{rank}(X) = \ell < k$, then we will find $k - \ell$ eigenvalues that are zero.

If X is of full column rank k, so that there are no exact linear dependencies among the columns of X, then is a positive definite and symmetric matrix, and all its eigenvalues are not only real but also positive. If we find a "small" eigenvalue, $\lambda_i \approx 0$, then

$$\lambda_i = z_i^Tz_i = c_i^TX^TXc_i \approx 0$$

and there is a near exact linear dependency among the columns of X. If there is a single small eigenvalue, then the linear relation indicates the form of the linear dependency, and

which of the explanatory variables are involved in the relationship. If there are two (or more) small eigenvalues, then we have two (or more) near exact linear relations. Multiple linear relationships do not necessarily indicate the form of the linear dependencies. The eigenvectors associated with the near zero eigenvalues define a 2-dimensional vector space in which the two near exact linear dependencies exist. While we may not be able to identify the individual relationships among the explanatory variables that are causing the collinearity, we can identify the variables that appear in the two (or more) relations.

The singular-value decomposition of X is an alternative technique that achieves the same goals as the analysis of eigenvalues. For computational reasons there are reasons to prefer the singular-value decomposition, and the literature on collinearity is divided between the two approaches. The matrix X may be decomposed as $X = UDV^T$, where $U^TU = V^TV = I_k$ and D is a diagonal matrix with non-negative diagonal values $\mu_1, \mu_2, \ldots, \mu_k$, called the singular values of X.

The relation to eigen analysis is that the singular values are the positive square roots of the eigenvalues of X^TX, and the matrix $V = C$. A small singular value implies a near exact linear dependence among the columns of X, just as does a small eigenvalue. We will ignore the computational issues and treat these two approaches as equivalent.

1.2 Collinearity and the Least Squares Estimator

Using equation (1) and the properties of the matrix of eigenvectors C, we can write $X^TX = X\Lambda C^T$, and therefore

$$(X^TX)^{-1} = C\Lambda^{-1}C^T = \sum_{i=1}^{k} \lambda_i^{-1} c_i c_i^T \tag{2}$$

defining $C = \{c_1, c_2, \ldots, c_k\}$ to be the matrix of characteristic vectors. The covariance matrix of the least squares estimator b is $cov(b) = \sigma^2(X^TX)^{-1}$, and using equation (2) the variance of b_j is

$$var(b_j) = \sigma^2 \left(\frac{c_{j1}^2}{\lambda_1} + \frac{c_{j2}^2}{\lambda_2} + \ldots + \frac{c_{jk}^2}{\lambda_k} \right) \tag{3}$$

The orthonormality of C implies that $\sum_{\ell=1}^{k} c_{j\ell}^2 = 1$, so variance of b_j depends upon three distinct factors. First, the magnitude of the error variance, σ^2; second, the magnitudes of the constants c_{jk}; and third, the magnitude of the eigenvalues, λ_ℓ. A small eigenvalue may cause a large variance for b_j if it is paired with a constant $c_{j\ell}$ that is not close to zero. The constants $c_{j\ell} = 0$ when x_j and x_ℓ, are orthogonal. This fact is an important one for it will allow one to determine which variables are "not" involved in collinear relationships.

Suppose β_j is a critical parameter in your model, and there is one small eigenvalue, $\lambda_k \approx 0$.

If x_j is not involved in the corresponding linear dependency , then c_{jk} will be small, and the fact that will not adversely affect the precision of estimation of β_j. The presence of collinearity in the data does not automatically mean that "all is lost." If X^TX has one or more small eigenvalues, then you must think clearly about the objectives of your research, and determine if the collinearity reduces the precision of estimation of your key parameters by an unacceptable amount. This leads us to the next question, "What is a small eigenvalue?"

1.3 Variance Decomposition of Belsley et al. (1980)

A useful property of eigenvalues is that $tr(X^TX) = \sum_{i=1}^{k} \lambda_i$. This implies that the size of the eigenvalues is determined in part by the scaling of the data. Data matrices consisting of large numbers will have larger eigenvalues, in total, than data matrices with small numbers. To remove the effect of scaling Belsley et al. (1980)), whose collinearity diagnostic procedure is proposed here, suggest scaling the columns of X to unit length. This scaling is only for the purpose of diagnosing collinearity, not for model estimation or interpretation.

To diagnose collinearity, examine the proportion of the variance of each least squares coefficient contributed by each individual eigenvalue. Define $\phi_{jk} = c_{jk}^2/\lambda_k$, and let ϕ_j be the variance of b_j, apart from the error variance, σ^2.

$$\phi_j = \left(\frac{c_{j1}^2}{\lambda_1} + \frac{c_{j2}^2}{\lambda_2} + \ldots + \frac{c_{jk}^2}{\lambda_k} \right)$$

Then, the proportion of the variance of b_j associated with the k^{th} eigenvalue λ_k is $\frac{\phi_{jk}}{\phi_j}$. Note the reversal of the subscripts. This is convenient for tabling the variance proportions, which has a now standard format. The columns of the table correspond to the variances of individual least squares coefficients, and the sum of each column is one. The rows of this matrix correspond to the different eigenvalues, which have been scaled in a certain way. The "condition index" is the square root of the ratio of the largest eigenvalue, λ_1, to the ℓ^{th} largest, λ_ℓ, that is,

$$\eta_\ell = \left(\frac{\lambda_1}{\lambda_\ell} \right)^{\frac{1}{2}}.$$

The condition indices are ordered in magnitude, with $\eta_1 = 1$ and η_k being the largest, since its denominator is the smallest eigenvalue.

Table 1 summarizes much of what we can learn about collinearity in data. BKW carried out extensive simulations to determine how large condition indices affect the variances of the least squares estimators. Their diagnostic procedures, also summarized in Belsley (1991, Chapter 5), are these:

Condition Index	Variance Proportions of OLS $var(b_1)$	$var(b_2)$	$\cdots$	$var(b_k)$
η_1	ϕ_{11}	ϕ_{12}	$\cdots$	ϕ_{1k}
η_1	ϕ_{21}	ϕ_{22}	$\cdots$	ϕ_{2k}
.				
.	$\cdots$	$\cdots$	$\cdots$	$\cdots$
.				
η_k	ϕ_{k1}	ϕ_{k2}		ϕ_{kk}

Table 1: Matrix of Variance Proportions

Step 1 Begin by identifying large condition indices. A small eigenvalue and a near exact linear dependency among the columns of X is associated with each large condition index. BKW's experiments lead them to the general guidelines that indices in the range 0-10 indicate weak near dependencies, 10-30 indicate moderately strong near dependencies, 30-100 is a large condition index, associated with a strong near dependency, and indices in excess of 100 are very strong. Thus when examining condition indexes values of 30 and higher should immediately attract attention.

Step 2 This depends on the number of large condition numbers identified in Step 1

A single large condition number: Examine the variance-decomposition proportions. If there is a single large condition number, indicating a single near dependency associated with one small eigenvalue, collinearity adversely affects estimation when two or more coefficients have 50% or more of their variance associated with the large condition index, in the last row of Table 1. The variables involved in the near dependency have coefficients with large variance proportions.

Two or more large condition numbers of relatively equal magnitude: If there are $J \geq 2$ large and roughly equal condition numbers, then X^TX has J eigenvalues that are near zero and J and there are J near exact linear dependencies among the columns of X. Since the J corresponding eigenvectors span the space containing the coefficients of the true linear dependence, the "50% rule" for identifying the variables involved in the near dependencies must be modified. In this case, sum the variance proportions for the coefficients across the J large condition number rows in Table 1. The variables involved in the (set of) near linear dependencies are identified by summed coefficient variance proportions of greater than 50%. The variance proportions in a single row do not identify specific linear dependencies, as they did when there was but one large condition number.

Two or more large condition numbers with one extremely large: An extremely large condition index, arising from a very small eigenvalue, can "mask" the variables involved in other near exact linear dependencies. For example, if one condition index is 500 and another is 50, then there are two near exact linear dependencies among the columns of X. However, the variance decompositions associated

with the condition number of 50 may not indicate that there are two or more variables involved in a relationship. Identify the variables involved in the set of near linear dependencies by summing the coefficient variance proportions in the last J rows of Table 1, and locating the sums greater than 50%.

Step 3 Perhaps the most important step in the diagnostic process is determining which coefficients are **not** affected by collinearity. If there is a single large condition number, coefficients with variance proportions less than 50% in the last row of Table 1 are not adversely affected by the collinear relationship in the data. If there are $J \geq 2$ large condition numbers, then sum the last J rows of variance proportions. Coefficients with summed variance proportions of less than 50% are not adversely affected by the collinear relationships. If the parameters of interest have coefficients unaffected by collinearity, then small eigenvalues and large condition numbers are not a problem.

Step 4 If key parameter estimates are adversely affected by collinearity, further diagnostic steps may be taken. If there is a single large condition index the variance proportions identify the variables involved in the near dependency. If there are multiple large condition numbers, auxiliary regressions may be used to further study the nature of the relationships between the columns of X. In these regressions one variable in a near dependency is regressed upon the other variables in the identified set. The usual t-statistics may be used as diagnostic tools to determine which variables are involved in specific linear dependencies. See Belsley (1991, p. 144) for suggestions. Unfortunately, these auxiliary regressions may also be confounded by collinearity, and thus they may not be informative.

2 Identification in Nonlinear Models

Assessing the severity and consequences of nearly singular Hessians in nonlinear models is more complicated than in linear models, since the invertibility of the Hessian can be due to things other than collinearity of the variables of the model. The difficulties caused by the extension to nonlinear models can be illustrated using nonlinear least squares model and then extended to the context of maximum likelihood estimation, generalized linear models and other models that require nonlinear estimators. The basic BKW variance decomposition analysis extends easily to these situation.

2.1 Nonlinear Least Squares

Consider the nonlinear model

$$y = f(X, \beta) + e \tag{4}$$

where $e \sim (0, \sigma^2 I_T)$ and $f(X, \beta)$ is some nonlinear function that relates the independent variables and parameters to form the systematic portion of the model. The nonlinear least squares estimator chooses $\hat{\beta}$ to minimize $S(\beta) = e^T e$. The least squares solution is

$$Z(\beta)^T [y - f(X, \beta)] = 0 \tag{5}$$

where $Z(\beta) = \partial f(X, \beta)/\partial \beta$. The matrix of second derivatives is referred to as the Hessian and is $H(\beta) = \partial^2 f(X, \beta)/\partial \beta \beta^T$. If there is more than one value of β that minimizes S, then the parameters of the model are *unidentified* and cannot be estimated. This occurs when the Hessian is singular and corresponds to perfect collinearity in the linear model. When the Hessian is nearly singular, then the model is poorly identified and reliable estimates may be difficult to obtain.

A useful algorithm for finding the minimum of $S(\beta)$ is the Gauss-Newton. The Gauss-Newton algorithm is based on a first order Taylor's series expansion of $f(X, \beta)$ around an initial guess, β_1, for the parameters, β. From that a pseudo-linear model is constructed

$$\bar{y}(\beta_1) = Z(\beta_1)\beta + e \tag{6}$$

where $\bar{y}(\beta_1) = y - f(x, \beta_1) + Z(\beta_1)\beta_1$. Notice that the dependent variable, $\bar{y}(\beta)$ and the regressors, $Z(\beta_1)$ are completely determined given β_1. The next round estimate, β_2 is obtained by using ordinary least squares on the pseudo-linear model, $\beta_2 = [Z(\beta_1)^T Z(\beta_1)]^{-1} Z(\beta_1)^T \bar{y}(\beta_1)$, on equation (6). The iterations continue until $\beta_{n=1} \approx \beta_n$.

It can be shown that asymptotically

$$Z(\beta)^T Z(\beta)/2T \doteq H(\beta)/T. \tag{7}$$

Therefore, if H is nearly singular, then $Z(\beta)^T Z(\beta)$ will be as well. This implies that the columns of $Z(\beta)$ can be treated as regressors and analyzed using the diagnostic procedures discussed in the preceding sections.

The Gauss-Newton algorithm is affected by collinearity among the columns of $Z(\beta)$ since $[Z(\beta_n)^T Z(\beta_n)]$ may become singular for any of its iterations. In fact, the model could be well conditioned at the final solution, but be nearly singular at one of the many intermediate points visited by the Gauss-Newton algorithm. Unfortunately, when a near singularity is encountered the algorithm becomes numerically unstable and it often fails to converge. A solution here is to pick better starting values that avoid regions of the parameter space for which the function is ill-conditioned.

A more common scenario is that the function itself is badly behaved for many points in the parameter space, including the actual minimum. In this instance, the collinearity problem is very similar to that in linear models and can be examined by using the collinearity diagnostics discussed above on the matrix of pseudo-regressors, $Z(\beta_n)$.

The conditioning of the data can be influenced to some degree by rescaling the data. Many convergence problems can be solved simply by scaling your variables in the appropriate way. On the other hand, the ill-effects of collinearity may persist regarless of scaling. By this we mean that precise estimates of the parameters are just not possible with the given data no matter how they are scaled. To detect collinearity in this setup it is suggested that the the columns of $Z(\beta)$ be rescaled to have the same length before computing the collinearity diagnostics. Large condition numbers indicate collinearity that cannot be further mitigated by scaling.

Although there are other algorithms for finding the minimum of $S(\beta)$ they are all likely to suffer the same ill-effects from collinearity.[1] It is possible that some may be better behaved in the intermediate steps of the iterative solution. Nevertheless, the asymptotic result in equation (7) suggests that in the end, it is unlikely that the ill-effects of collinearity can be manipulated in a material way by using another estimator of the asymptotic covariance matrix.

2.2 Maximum Likelihood

Maximum likelihood estimation can be approached in a similar fashion. Instead of minimizing the sum-of-squared errors function the goal is to choose parameter values that maximize the log-likelhihood function, $\ell(\beta, X)$. The algorithms use either first derivatives of ℓ, the second, or both. As in the Gauss-Newton algorithm for nonlinear least squares, each of the algorithms involves inversion of the hessian (e.g., Newton-Raphson), its negative expectation (the negative information matrix used in the method-of-scoring), or a cross-products matrix of partial first derivatives (e.g. the method of Berndt, Hall, Hall, and Hausman). In any of these instances, the inverted matrix evaluated at the each round of estimates is instrumental in solving for the parameter values that maximize the likelihood function. If at any point in the process it becomes singular or nearly so, estimation fails. If convergence occurs, then the inverse of the estimated asymptotic covariance matrix can be subjected to conditioning diagnostics in the same manner as the NLLS estimator.

2.3 Generalized Linear Models

This basic approach has been used in other contexts. Weissfeld and Sereika (1991) explore the detection of collinearity in the class of generalized linear models (GLM). This broad class of models includes the linear regression model, binary choice models like logit and probit,

[1] For instance, the Newton-Raphson, which is based on the second order Taylor's series approximation, uses the hessian computed at each round.

polychotomous choice models, the Poisson regression model, the cox proportional hazard model, and others (see McCullagh and Nelder (1989) for discussion). In the generalized linear models the information matrix associated with the log-likelihood function can be expressed generally as

$$I(\beta) = X^T W X \tag{8}$$

where W is a $T \times T$ diagonal weight matrix that often is a function of the unknown parameters, β, the independent variables, and the responses, y. In this form, Segerstedt and Nyquist (1992) observe that ill-conditioning in these models can be due to collinearity of the variables, X, the influence of the weights, W, or both. They suggest a transformation of the data that, when plotted in the same diagram with the original data, can illuminate the change in conditioning that occurs due to the weights. Unfortunately, the method is manageable only in a few dimensions.

In GLM, Weissfeld and Sereika (1991) suggest applying the BKW condition number diagnostics to the the scaled information matrix ($-E[H(\beta)]$). Lee and Weissfeld (1996) do the same for the Cox regression model with time dependent regressors. Although the variance decompositions can be computed in these instances, their interpretation is not as straightforward since collinearity can also be due to the way the weights interact with the explanatory variables.

Lesaffre and Marx (1993) also investigate the problem of ill-conditioning in generalized linear models and take a slightly different approach. Following Mackinnon and Puterman (1989) they suggest that only the columns of X be standardized to unit length, forming X_1. Then, conditioning diagnostics are computed on $X_1 \hat{W} X_1$, where $\hat{W}$ is the estimated weight matrix based on the rescaled data.[2] The square root of the ratio of largest to smallest eigenvalue describes the worst relative precision with which linear combinations of the location parameters can be estimated. Thus, this scaling gives a structural interpretation to the conditioning diagnostic. One problem with this scaling is that $X_1 \hat{W} X_1$ could be ill-conditioned because of the effects of $\hat{W}$ which could either cause the algorithm to fail or result in very large estimated variances for the parameters of the model.

2.4 BKW diagnostics based on covariance

All of the approaches to diagnosing poorly identified models can be subjected to the condition number, variance decomposition of BKW.

Even though the BKW diagnostic can identify weaknesses of the data or model, they cannot distinguish problems in the data from problems with the parameters, since the two interact

[2]Note, $X_1 \hat{W} X_1$ is not rescaled. This is not the same as finding the condition number of the scaled estimated inverse of the information matrix.

often in a nonseparable way in the estimator's covariance. Despite these problems, we compute condition numbers and perform the BKW decomposition on the scaled estimated inverse of the variance-covariance matrix. This is convenient in gretl since the variance-covariance of on estimated model can be retrieved after estimation using the accessor `$vcv`. The inverse variance-covariance matrix is scaled so that the principal diagonal has one in each element. In recent releases of gretl, the `vif` command summons the BKW diagnostic matrix after estimation of any model. Stata offers similar functionality. Refer to Adkins et al. (2015) for computational details based on covariance.

3 Example: Ordered Probit

The ordered probit is easily estimated in modern software, gretl and Stata. It contains additional parameters that are related to the number of categories, or bins, for the dependent variable. These "cutoff" parameters determine the points at which the observation moves from one category to the adjacent one when an independent variable changes. Following (Greene, 2012, p. 787-788) the ordered probit model is treated as a latent variable, y_i^* that depends on a linear index, $x_i^T\beta$.

$$y_i^* = x_i^T\beta + e_i$$

The latent variable y^* is unobserved. Instead we observe integers, y, such that

$$\begin{array}{ll} \text{y=0} & \text{if } y^* \leq 0 \\ \text{y=1} & \text{if } 0 \leq y^* \leq \tau_1 \\ \text{y=2} & \text{if } \tau_1 \leq y^* \leq \tau_2 \\ \vdots & \vdots \\ \text{y=J} & \text{if } \tau_{J-1} \leq y^* \end{array}$$

The β and the τs are unknown parameters. If e is assumed to be normally distributed across observations. To identify β, it is conventional to let $\sigma = 1$ in order to identify β. Then,

$$\begin{array}{l} \text{Prob}(y = 0|x) = \Phi(-x^T\beta) \\ \text{Prob}(y = 1|x) = \Phi(\tau_1 - x^T\beta) - \Phi(-x^T\beta) \\ \text{Prob}(y = 2|x) = \Phi(\tau_2 - x^T\beta) - \Phi(\tau_1 - x^T\beta) \\ \vdots \\ \text{Prob}(y = J|x) = 1 - \Phi(\tau_{J-1} - x^T\beta) \end{array}$$

with $0 < \tau_1 < \tau_2 < \ldots < \tau_{J-1}$ and $\Phi(t)$ is the standard normal cdf evaluated at t.

In the following empirical example I model the probability of having children less than six years of age using the data are from Mroz. The regressors include a constant, mother's

education, mother's experience, and mother's age. The dependent variable, `kidls6`, takes integer values 0, 1, 2, and 3. Identification is about linear relationships among the regressors as well as the parameterization of the model itself, which includes three regression coefficinets, βs, and three cutoff parameters, τs.

The results:

Model 2: Ordered Probit, using observations 1–753
Dependent variable: kidsl6
Standard errors based on Hessian

	Coefficient	Std. Error	z	p-value
educ	0.0437362	0.0265277	1.6487	0.0992
exper	−0.0282272	0.0101281	−2.7870	0.0053
age	−0.102429	0.00984913	−10.3999	0.0000
cut1	−2.91206	0.516607	−5.6369	0.0000
cut2	−1.74208	0.512837	−3.3970	0.0007
cut3	−0.692334	0.547788	−1.2639	0.2063

Mean dependent var	0.237716	S.D. dependent var	0.523959
Log-likelihood	−356.2304	Akaike criterion	724.4609
Schwarz criterion	752.2053	Hannan–Quinn	735.1494

Number of cases 'correctly predicted' = 610 (81.0 percent)
Likelihood ratio test: $\chi^2(3) = 196.349$ [0.0000]

Test for normality of residual –
Null hypothesis: error is normally distributed
Test statistic: $\chi^2(2) = 2.42029$
with p-value = 0.298154

The results suggest that `educ` is not significant at 5% and neither is the right-most cutoff parameter. The collinearity diagnostics could shed light as to why.

The `vif` command in GRETL produces the BKW table as follows:

```
                        --- variance proportions ---
lambda      cond      educ      exper       age       cut1      cut2      cut3
```

3.460	1.000	0.002	0.022	0.002	0.001	0.000	0.000
1.360	1.595	0.000	0.001	0.000	0.001	0.013	0.033
0.888	1.974	0.000	0.002	0.000	0.002	0.005	0.110
0.256	3.673	0.010	0.963	0.007	0.002	0.002	0.002
0.029	10.833	0.505	0.002	0.385	0.000	0.002	0.005
0.006	23.627	0.482	0.012	0.605	0.994	0.977	0.850

The second column includes the condition numbers and the last six columns are the variance decompositions for the 3 variables and the 3 cutoff points. Overall, the conditioning is not too bad since the largest condition number is 23.627, which is below the extreme threshold of 30. The largest condition number plays a significant role in estimation of all three cutoff parameters, which all have variance proportions greater than 85%. This suggests that these parameters are relatively weakly identified by the model.

With one fairly high condition number there appears to be one moderately collinear relationship that involves education and age, and made worse by the existence of the cutoff parameters. Surprisingly, of the four variables only education falls victim to weakness of the data. Experience does not appear to be collinear with any of the other variables. The model seems to be fairly well identified despite these issues.

Simulation

There are two sources of trouble with the identification of the ordered probit model. 1) the data could be collinear and 2) the likelihood function as parameterized could be relatively flat. The simulation explores both scenarios.

Collinear regressors There are three regressors in the model. A constant is not needed since it is unidentified. However, in generation of the regressors for the simulations, one is included so that a comparison with linear regression diagnostics can be made. For the simulation the regressor matrix

$$X = \{\text{const}, \text{age}, \text{educ}, \text{exper}\}. \tag{9}$$

is decomposed using SVD. $X = UDV$, where D is a diagonal matrix containing the eigenvalues of the original data. These are replaced using the desired ones Λ as in $X = U\Lambda V$.

Four sets of eigenvalues for the SVD are considered. The first makes the regressors mutually

orthogonal. Others impart various degrees and types of collinearity. The sets considered are:

$$\begin{array}{cc} \text{Eigenvalues} & \text{Collinearity} \\ \Lambda = \{1, 1, 1, 1\} & \text{None} \\ \Lambda = \{10, 7, 4, 1\} & \text{Moderate} \\ \Lambda = \{10, 1, 1, 0.1\} & \text{Moderate} \\ \Lambda = \{10, 10, 0.1, 0.1\} & \text{Moderately Severe} \\ \Lambda = \{10, 0.1, 0.05, 0.05\} & \text{Severe} \end{array} \tag{10}$$

To keep the overall variability of the data constant, the eigenvalues are rescaled to have equal length that is similar to the original Mroz data.

Parameters Gretl estimates the model using three cutoff parameters and no constant, which is not identified. Because the identification of the model could be affected by the cut-off parameters, several sets are used. The cutoff parameters are set by taking the mean of $X\beta$ and adding c such that $P(X < c) = p$ with the parameter p chosen from

Cumulative Prob	Bin Size
$p = \{0.1, 0.5, 0.9\}$	Wide
$p = \{0.1, 0.3, 0.5\}$	Moderate
$p = \{0.1, 0.2, 0.3\}$	Narrow

(11)

The parameters from the regression are set to β is set to $\{-0.1, 0.05, -0.02\}$, which mimic the values of the parameters when the original data are used.

Table 2 shows the average condition numbers for each of the collinearity designs.

Table 2: Average condition numbers for the various collinearity designs. Large bins, $p = \{0.1, 0.5, 0.9\}$.

	Collinearity among variables				
Cond	{1,1,1,1}	{10,1,1,0.1}	{10,0,0.1,0.1}	{10,.1,.05,.05}	
η_1	1.000	1.000	1.000	1.000	1.000
η_2	1.101	1.260	1.568	1.331	1.601
η_3	1.138	1.351	1.831	1.517	1.899
η_4	1.326	1.573	5.270	1.669	17.320
η_5	1.972	2.624	8.134	15.470	62.650
η_l	14.280	18.530	24.280	75.860	171.300

Notice that even when the variables are mutually orthogonal (column 1) there is still evidence of weakening identification. This has to be due to the functional form for ordered probit. As collinearity worsens, the largest condition number increases in magnitude. By the time collinearity is very bad (rightmost column), there are several condition numbers that lie above the BKW suggested threshold of 30.

Bin size also affects the condition numbers. This is shown in Table 3. The collinearity regime for these results is $\Lambda = \{10, 7, 4, 1\}$, which is very mild. The first panel has a small bin at each extreme ($p < 0.1$ and $p > 0.9$), but wide bins in the middle of the distribution. The second panel has one very wide bin at the upper extreme and moderately narrow ones at the lowest. The bottom panel has very narrow bins at the lower end and 70% probability of landing in the last one. One can see that as the bins get narrower at the bottom end, conditioning worsens as η_l increases from 18.53 to 29.41. You also see that as the condition number gets larger, the standard errors of the cutoff parameters gets larger. This is indicated by the large and increasing variance proportions in each of the last three columns of the BKW tables.

Specific subsets of the simulation results appear in Tables 4-7 below. In Table 4 the best case scenario for ordered probit is found. Here, all of the regressors are orthogonal to one another and the bins for the ordered response are centered around 0.5 and relatively wide (at least in the center of the distribution). The summary statistics show that the MLE has low bias (perhaps unbiased) for this design. The overall variability of the estimates is very small. The t-ratio measures the relative precision of the MLE (not its unbiasedness).

The variance proportion table shows evidence of some ill-conditioning that is independent of the data, which are orthogonal. The first six condition numbers increase steadily from one to two and the last is 14.27. This falls within the moderate level for the BKW diagnostics in linear models.

Further, the largest condition number affects each of the parameters, though mostly $x1$ and the three cutoff parameters.

The standard errors of the condition numbers and the variance decompositions appear at the bottom panel of each table. Since the regressors are not changing in the simulation, all of the variance is due to the parameter estimates. The relative precision of the parameter estimates is evidenced by the small standard deviations in this panel for all of the statistics computed. It is easy to conclude that parameter variation due to their estimation is not contributing to any significant degree the relative magnitudes of the diagnostics themselves. For instance, comparing the bottom panels of tables 4 and 7 (no collinearity and high collinearity, respectively) more collinear regressors are causing more variance in the condition numbers, but not in the variance proportions. These appear to be fairly invariant to collinearity among regressors. The implication is that the diagnostics themselves change in predictable ways as collinearity changes, but the variation of the measures is fairly constant in each design.

In Table 5 the degree of collinearity in the variables is introduced. The overall level is kept moderately low. The standard errors of the estimated cutoff parameters increases relative to the baseline.

The average largest condition number is 24.31 indicating that identification is becoming more challenging. The high variance proportions for the cutoff parameters is consistent with the higher standard errors found in the upper panel. The collinearity of the variables themselves does not appear to be much of a problem, except possibly for x1 which shares a large variance proportion with the cutoffs.

Tables 6 and 7 show the effects of increasing collinearity among the regressors. In the first case, only two of the regressors are highly collinear and in the second, all three are collinear with each other. In Table 6 there is a single large condition number that affects estimation of the βs, but not the cutoffs. Relative to the βs, the cutoffs are not affected to a large extent by collinearity of the variables. So, as collinearity of the variables gets worse, identification of their parameters weakens while that of the cutoffs improves, at least relatively.

In Table 7 there are two large condition numbers, with β_1 and β_2 being weakly identified. The other parameter, β_3, appears to be fairly well identified. In all of these cases, the variation caused by parameter estimation is low.

Table 3: Various bin sizes for a mildly collinear data in Ordered Probit. $\Lambda = 10, 7, 4, 1$. There is only 1 relatively large condition number in this scenario.

	Bin Size: $p = 0.1, 0.5, 0.9$					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-3.366	-2.084	-0.803
Mean	-0.099	0.051	-0.020	-3.340	-2.062	-0.769
Std Err	0.012	0.009	0.003	0.377	0.365	0.354
η_l	ϕ_{l1}	ϕ_{l2}	ϕ_{l3}	ϕ_{l4}	ϕ_{l5}	ϕ_{l6}
18.530	0.978	0.634	0.448	0.977	0.989	0.968

	Bin Size: $p = 0.1, 0.3, 0.5$					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-3.366	-2.609	-2.084
Mean	-0.099	0.052	-0.020	-3.322	-2.558	-2.043
Std Err	0.013	0.010	0.003	0.392	0.388	0.381
η_l	ϕ_{l1}	ϕ_{l2}	ϕ_{l3}	ϕ_{l4}	ϕ_{l5}	ϕ_{l6}
23.220	0.979	0.619	0.472	0.984	0.995	0.992

	Bin Size: $p = 0.1, 0.2, 0.3$					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-3.366	-2.926	-2.609
Mean	-0.099	0.052	-0.020	-3.326	-2.880	-2.562
Std Err	0.014	0.010	0.004	0.429	0.424	0.423
η_l	ϕ_{l1}	ϕ_{l2}	ϕ_{l3}	ϕ_{l4}	ϕ_{l5}	ϕ_{l6}
29.410	0.980	0.594	0.500	0.991	0.997	0.995

	Monte Carlo Summary Statistics					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-2.353	-1.071	0.2104
Mean	-0.101	0.051	-0.020	-2.361	-1.073	0.2263
Std Err	0.018	0.008	0.004	0.289	0.281	0.2793
t-ratio	-0.040	0.120	0.014	-0.030	-0.008	0.0569

	Averages of the Variance Proportion Table					
η_l	β_1	β_2	β_3	τ_1	τ_2	τ_3
1.0000	0.0058	0.0200	0.0031	0.0027	0.0004	0.0070
1.1010	0.0020	0.0322	0.0043	0.0132	0.0009	0.0037
1.1380	0.0017	0.0014	0.0019	0.0020	0.0153	0.0061
1.3260	0.0013	0.0003	0.3843	0.0001	0.0000	0.0001
1.9690	0.0056	0.1656	0.0034	0.0191	0.0001	0.0257
14.2700	0.9836	0.7805	0.6029	0.9629	0.9832	0.9574

	Standard Errors of the Variance Proportion Table					
-	0.0003	0.0021	0.0005	0.0006	0.0003	0.0006
0.0082	0.0006	0.0037	0.0021	0.0017	0.0014	0.0018
0.0128	0.0010	0.0019	0.0022	0.0016	0.0018	0.0014
0.0107	0.0002	0.0004	0.0048	0.0001	0.0001	0.0001
0.0584	0.0006	0.0101	0.0041	0.0022	0.0001	0.0028
0.1252	0.0006	0.0098	0.0035	0.0029	0.0008	0.0036

Table 4: Orthogonal Regressors: $\Lambda = \{1, 1, 1, 1\}$, with Wide Bins: p=$\{0.1, 0.5, 0.9\}$. Results based on 100 Monte Carlo samples.

	Monte Carlo Summary Statistics					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-5.197	-3.916	-2.634
Mean	-0.101	0.049	-0.019	-5.252	-3.943	-2.667
Std Err	0.007	0.014	0.010	0.380	0.364	0.348
t-ratio	-0.119	-0.086	0.126	-0.145	-0.077	-0.096
	Averages of the Variance Proportion Table					
η_l	β_1	β_2	β_3	τ_1	τ_2	τ_3
1.0000	0.0016	0.0043	0.0111	0.0006	0.0003	0.0005
1.5680	0.0000	0.0000	0.0001	0.0035	0.0080	0.0047
1.8320	0.0000	0.0009	0.0006	0.0125	0.0001	0.0144
5.2570	0.0035	0.0885	0.7976	0.0111	0.0044	0.0008
8.1320	0.1624	0.4825	0.0942	0.0022	0.0092	0.0208
24.3100	0.8325	0.4238	0.0964	0.9701	0.9780	0.9587
	Standard Errors of the Variance Proportion Table					
-	0.0001	0.0001	0.0001	0.0000	0.0000	0.0000
0.0149	0.0000	0.0000	0.0000	0.0006	0.0003	0.0006
0.0045	0.0000	0.0001	0.0002	0.0012	0.0001	0.0015
0.0691	0.0007	0.0057	0.0225	0.0010	0.0004	0.0002
0.1256	0.0072	0.0208	0.0162	0.0003	0.0008	0.0022
0.2551	0.0073	0.0159	0.0095	0.0015	0.0012	0.0032

Table 5: Moderate Collinearity: $\Lambda = \{10, 1, 1, 0.1\}$, with Wide Bins: p=$\{0.1,\ 0.5,\ 0.9\}$. Results based on 100 Monte Carlo samples.

	Monte Carlo Summary Statistics					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-3.650	-2.368	-1.086
Mean	-0.103	0.054	-0.020	-3.706	-2.410	-1.124
Std Err	0.044	0.179	0.004	0.285	0.284	0.287
t-ratio	-0.062	0.023	-0.108	-0.200	-0.147	-0.131

	Averages of the Variance Proportion Table					
η_l	β_1	β_2	β_3	τ_1	τ_2	τ_3
1.0000	0.0002	0.0001	0.0000	0.0008	0.0007	0.0007
1.3310	0.0000	0.0000	0.0003	0.0053	0.0064	0.0057
1.5160	0.0000	0.0000	0.0612	0.0068	0.0001	0.0068
1.6720	0.0000	0.0000	0.0612	0.0083	0.0002	0.0118
15.4900	0.0185	0.0031	0.1978	0.6253	0.6094	0.5738
75.8800	0.9813	0.9968	0.6794	0.3534	0.3833	0.4012

	Standard Errors of the Variance Proportion Table					
-	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001
0.0107	0.0000	0.0000	0.0003	0.0006	0.0002	0.0007
0.0154	0.0000	0.0000	0.0032	0.0010	0.0001	0.0011
0.0235	0.0000	0.0000	0.0035	0.0013	0.0001	0.0016
0.1385	0.0004	0.0001	0.0044	0.0091	0.0064	0.0038
0.2485	0.0004	0.0001	0.0050	0.0092	0.0065	0.0032

Table 6: Severe Collinearity: $\Lambda = \{10, 10, 0.1, 0.1\}$, with Wide Bins: p=$\{0.1,\ 0.5,\ 0.9\}$. Results based on 100 Monte Carlo samples.

	Monte Carlo Summary Statistics					
	β_1	β_2	β_3	τ_1	τ_2	τ_3
True	-0.100	0.050	-0.020	-5.974	-4.692	-3.411
Mean	-0.098	0.027	-0.009	-6.069	-4.771	-3.472
Std Err	0.050	0.191	0.078	0.356	0.344	0.337
t-ratio	0.041	-0.123	0.144	-0.267	-0.228	-0.181

	Averages of the Variance Proportion Table					
η_l	β_1	β_2	β_3	τ_1	τ_2	τ_3
1.0000	0.0000	0.0000	0.0001	0.0005	0.0003	0.0004
1.6050	0.0000	0.0000	0.0000	0.0037	0.0074	0.0049
1.9000	0.0000	0.0000	0.0000	0.0122	0.0001	0.0164
17.3200	0.0010	0.0003	0.0153	0.5755	0.5384	0.4904
62.7300	0.0660	0.0355	0.9750	0.1401	0.1510	0.1537
171.5000	0.9330	0.9643	0.0096	0.2680	0.3028	0.3342

	Standard Errors of the Variance Proportion Table					
-	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0142	0.0000	0.0000	0.0000	0.0006	0.0003	0.0007
0.0069	0.0000	0.0000	0.0000	0.0014	0.0001	0.0019
0.2352	0.0001	0.0000	0.0005	0.0167	0.0132	0.0113
0.3206	0.0010	0.0006	0.0013	0.0128	0.0112	0.0099
0.6394	0.0010	0.0006	0.0012	0.0160	0.0134	0.0105

Table 7: Severe Collinearity: $\Lambda = \{10, 0.1, 0.05, 0.05\}$, with Wide Bins: p={0.1, 0.5, 0.9}. Results based on 100 Monte Carlo samples.

The variance decomposition diagnostics clearly show that the smallest eigenvalue is wreaking havoc on estimation of the βs, but not so much for the cutoffs which had very substantial average t-ratios due to smallish standard errors. I would tentatively conclude that the ordered probit functional form is fairly robust with respect to identification of parameters. Of course, other designs may reveal otherwise.

4 Final Thoughts

To sum up, the BKW diagnostics appear to be quite useful in identifying problems with nonlinear estimation. Large condition numbers that have large effects on the variances of two or more coefficients can signal issues with the data or the model. Also, higher collinearity among variables may actually be useful if interest is on other parameters in the model. This is counterintuitive and deserves additional investigation.

References

Adkins, Lee C. (2017), 'Weak identification in nonlinear econometric models', *Presented at 5th Gretl Conferenece, Athens, Greece* .
URL: *https://learneconometrics.com/pdf/GC2017/collin_gretl_170523.pdf*

Adkins, Lee C., Melissa Waters and R. C. Hill (2015), 'Collinearity diagnostics in gretl', *Presented at 4th Gretl Conferenece, Berlin Germany* .
URL: *http://learneconometrics.com/pdf/Collin/collin_gretl.pdf*

Belsley, D. A. (1991), *Collinearity Diagnostics: Collinearity and Weak Data in Regression*, Wiley, New York.

Belsley, David A. (1991), *Collinearity Diagnostics: Collinearity and Weak Data in Regression*, John Wiley & Sons, New York.

Belsley, David A., E. Kuh and R. E. Welsch (1980), *Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity*, John Wiley & Sons, New York.

Greene, W. (2012), *Econometric Analysis*, 7th edn, Prentice Hall, Upper Saddle River, NJ.

Hill, R. Carter and Lee C. Adkins (2001), Collinearity, *in* B.Baltagi, ed., 'Companion to Theoretical Econometrics', Blackwell Publishers Ltd, Malden Massachusettes, pp. 256–278.

Lee, Kyung Yul and L. A. Weissfeld (1996), 'A multicollinearity diagnostic for the cox model with time dependent covariates', *Communications in Statistics - Simulation* **25**, 41– 60.

Lesaffre, E. and B. D. Marx (1993), 'Collinearity in generalized linear regression', *Communications in Statistics - Theory and Methods* **22**, 1933–1952.

Mackinnon, M. J. and M. L. Puterman (1989), 'Collinearity in generalized linear models', *Communications in Statistics - Theory and Methods* **18**, 3463–3472.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Segerstedt, B. and H. Nyquist (1992), 'On the conditioning problem in generalized linear models', *Journal of Applied Statistics* **19**, 513–522.

Silvey, S. (1969), 'Multicollinearity and imprecise estimation', *Journal of the Royal Statistical Society, B* **31**, 539–552.

Weissfeld, L. A. and S. M. Sereika (1991), 'A multicollinearity diagnostic for generalized linear models', *Communication in Statistics, A* **20**, 1183–1198.