# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**UMI**

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

# MULTICOLLINEARITY DIAGNOSTICS FOR MULTIPLE REGRESSION: A MONTE CARLO STUDY

BY

Peter Flom
B.A., New York University, 1980
M.A., New York University, 1984

DISSERTATION
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN THE DEPARTMENT
OF PSYCHOLOGY AT FORDHAM UNIVERSITY

NEW YORK
April 8, 1999

# FORDHAM UNIVERSITY

*Graduate School of Arts & Sciences*

April 7 19 99

**This dissertation prepared under my direction by**

Peter Flom

**entitled** Multicollinearity Diagnostics For Multiple Regression:

A Monte Carlo Study

**has been accepted in partial fulfillment of the requirements for the Degree of**

*Doctor of Philosophy*

**in the Department of**

Psychology

(Mentor)

(Reader)

(Reader)

TABLE OF CONTENTS

# List of Tables

# List of Figures

ACKNOWLEDGEMENTS

I would like to thank my committee members, Warren Tryon, Anuparma Byravan, Thanos Patelis, and Mary Procidano for their assistance and useful editorial comments. Most especially I would like to thank my mentor, John Walsh. If every mentor were like him, fewer people would have ABD their final degree.

I would also like to thank my parents for their support throughout my life, my wife, for her support throughout the dissertation process, and my son, for being himself and bringing me joy.

# CHAPTER I

# INTRODUCTION

## Thesis

Multiple regression (MR) is one of the most widely used statistical techniques in psychology and the social sciences. The MR model is:

$$[1.1] \qquad y = XB + \varepsilon$$

Where y is an n x 1 vector of the dependent variable (DV), X is an n x p matrix of independent variables (IVs) and an intercept term, B is a p x 1 vector of regression coefficients and $\varepsilon$ is an n x 1 vector of error terms.

In the most familiar form of regression, B is estimated using ordinary least squares (OLS):

$$[1.2] \qquad b = (X'X)^{-1}X'y$$

where b is a p-vector of estimates of B. OLS assumes that $\varepsilon$ is independent of X, that $E(\varepsilon) = 0$, and that $E(\varepsilon\varepsilon') = \sigma^2 I$.

It is often the case that some of the IVs are related to each other. When this relation is strong, there may be a condition called collinearity. Collinearity can be thought of as a continuum from orthogonality to exact collinearity. Data are orthogonal when each variable adds completely new information. Exact collinearity occurs when any IV is a linear combination of any set of other IVs.

Exact collinearity rarely occurs in actual research. However, near collinearity does occur, and can cause severe problems. These include: Difficulty in interpreting the partial coefficients, instability of the coefficients, and computational inaccuracies. Given these difficulties, it is important to be able to diagnose near collinearity[1]. Areas of psychological research where collinearity may occur include personality and psychopathology research where a number of personality scales are used as IVs.

This dissertation used computer simulation to examine and compare the behavior of two widely recommended MR collinearity diagnostics when the number of independent variables (IVs), the presence of an interaction between two of the IVs, and the degree of collinearity were varied. The two collinearity diagnostics which were examined are variance inflation factors (VIFs) and condition indexes (Belsley, 1991; Belsley, Kuh, & Welsch, 1980). In addition, Belsley's results were extended to new situations, and distributions for Belsley's diagnostics were computed. These findings will provide users of

---

[1] For brevity, the term "collinearity" will be used to mean "nearly exact collinearity", except where this would cause confusion.

multiple regression with further information about diagnosing collinearity.

## Literature Review

### Assessment of Collinearity

Various methods have been proposed for assessing collinearity. These methods are known as collinearity diagnostics. This review is organized as follows: First, six commonly used methods are discussed, and reasons why they are inadequate are detailed. Second, methods involving VIFs are discussed. Finally, the method proposed by Belsley, an econometrician who has studied collinearity extensively, is described.

#### Commonly Used Methods

One commonly recommended collinearity diagnostic is to examine the correlation matrix of the independent variables for coefficients that are "high". There are two problems with using the correlation matrix to diagnose collinearity. First, while a high correlation coefficient between two variables is sufficient for collinearity, it is not necessary to it. This is so for two reasons: a) It is possible for two uncorrelated variables to be highly collinear (see Appendix A); b) Each correlation coefficient describes the relationship between two IVs; collinearity may involve one IV being a linear combination of several

other IVs. In fact, it can be shown (see Appendix A) that there can be an exact collinear relation among p variables with no correlation coefficient exceeding 1/(p-1). Thus, if there are 10 IVs, the set could be exactly collinear with no correlation above 0.111.

Second, there is no way to determine how high a correlation coefficient must be to cause problems. Choosing a cut-off value is also a problem with the collinearity diagnostics discussed below, but in at least some of those cases there may be a way to resolve this issue empirically. There is no way to do this for the correlation matrix, since individual correlations can be quite low even when there is exact collinearity (Belsley, 1991).

Another diagnostic which has been proposed is that results contrary to those predicted by theory are signs of collinearity. Either signs for some estimates are incorrect, or theoretically important regressors are nonsignificant. These conditions are neither necessary nor sufficient for collinearity. These conditions could exist for at least two other reasons: a) the model could be incorrect, b) there could be other problems with the data, such as influential points, outliers, or low (Belsley,

1991). Therefore, this method is insufficient as a collinearity diagnostic.

Another diagnostic method is to examine the determinant of X'X. Two problems with this method have been identified. First, it is very sensitive to scaling; for a p x p matrix A and a constant k:

$$[1.3] \qquad \det(kA) = k^p \det(A).$$

For example, if we have five independent variables, each of which is measured in feet, and subsequently change the scale to inches, the determinant will increase by a factor of $12^5 = 248{,}832$, while the degree of collinearity is unchanged (Schott, 1997). The problem of scale is common to many diagnostics, and can be solved by scaling the columns of the matrix to some common length (usually unit length).

Second, even if the matrix is scaled so that all its columns are of unit length, it is possible to have data that cause the determinant of X'X to indicate severe collinearity when, in fact, collinearity is modest (see Stewart, 1987 for a proof).

A variation of this method is to examine the determinant of the correlation matrix, R. This method shares the second weakness of the determinant of X'X, described above.

All-subsets regression on X has also been proposed as a collinearity diagnostic. This involves regressing each column of X on all possible combinations of the other columns. This method has three major problems: 1) Computational costs are high, 2) Interpretive costs are enormous (in that a large number[2] of regressions are needed, and each must be examined) and 3) When there are several collinear relationships in the data, it is possible that this method will fail to diagnose existing collinearity. This occurs because a particular subset of the data may be collinear itself; this can cause the variances for the parameters output by the all-subsets regression output to be statistically nonsignificant (Belsley, 1991). The first two problems might not be fatal, especially for models with relatively small numbers of IVs, but the third problem makes all-subsets regression unacceptable as a collinearity diagnostic.

Variance Inflation Factors

---

[2] The number of regressions which must be computed and interpreted is $\Sigma^{p-1}_{i=1}((p-1)!/i!(p-1-i)!)$. For p = 5, this is 15 regressions. For p = 7 it is 67 regressions.

Variance inflation factors (VIFs) are based on the inverse of the correlation matrix (R). If X is centered[3] and scaled to have unit length, then the VIFs are the diagonal elements of $R^{-1}$. These can be shown to be

$$[1.4] \qquad VIF_i = 1/(1-R_i^2)$$

where $R_i^2$ is the multiple correlation coefficient of $X_i$ regressed on the remaining IVs (Belsley, 1991). VIFs indicate the degree to which the variance of the parameter estimates is inflated due to collinearity. They have been used extensively by both statisticians and numerical analysts to indicate collinearity. However, Belsley (1991) indicates some reasons that they are not ideal for this purpose.

These reasons stem from the fact that, since the VIFs are based on the inverse of the correlation matrix, the drawbacks associated with using the correlation matrix largely apply to the VIF as well. One major problem with VIFs is determining how high they have to be to indicate problematic collinearity. However, it may be possible to determine this empirically (e.g., Montgomery & Peck, 1982, suggest that practical experience indicates that VIFs

---

[3] A variable is centered by subtracting the mean of all observations from each observation.

higher than 5 or 10 are signs of problematic collinearity; Snee, 1973, recommends 4 or 5; Marquadt, 1980, recommends 10). A second problem is that VIFs are unable to determine the number of collinear relationships. A third problem is that they are sufficient but not necessary conditions of collinearity (Belsley, 1991).

Nevertheless, VIFs (under various aliases including tolerances, which are the inverse of the VIFs) are widely recommended as a collinearity diagnostic (Montgomery & Peck, 1982; Snee, 1973; Snee & Marquadt, 1984). There are some advantages to VIFs: They are relatively simple to compute, and, more importantly, to interpret, since they provide a single number. In addition, the largest VIF is an upper bound on the condition number (defined below) which is, in turn, a fairly useful collinearity diagnostic, and which forms a part of Belsley's methods.

The same advantages and disadvantages apply to variations of the VIF which have been suggested, such as the tolerance, or the square root of the VIF (Belsley, 1991). The problems with VIFs are not due to their distribution, but to what they measure, and transformations of them cannot change this.

Belsley's Method

Belsley, Kuh and Welsch (1980) introduced a new method for assessing collinearity. This method was refined in Belsley (1991). It is briefly summarized here; more extensive discussion is in the Method section, and technical details are provided in Appendix B.

First, the condition indexes ($\eta_i$) of the data matrix are determined. These are the ratios of each of the singular values of the matrix to the smallest singular value. The condition indexes indicate how many collinear relations there are, and how severely each one inflates the variances of the parameter estimates. Then the variance decomposition proportions ($\pi_{ij}$) are determined. These indicate how much of the variance of each particular parameter estimate is due to any particular collinear relation.

Centering and the Intercept Term

Two questions which must be answered before implementing either the VIF or Belsley's procedure are whether the columns of the data matrix should be mean centered and whether an intercept term should be included. These questions are closely related, since centering removes the collinearity between the IVs and the intercept. Belsley (1984a, 1984b, 1986, 1991; Belsley, Kuh & Welsch, 1980) is firmly against centering and in favor of including

an intercept term when it is relevant. However, his arguments and examples have not convinced all others (Gunst, 1983; Marquadt, 1980; Montgomery & Peck, 1982; Stewart, 1987; Weisberg, 1980).

In this dissertation, the IVs will be generated to have a mean of 0 and an intercept term will be included (see below) so the centering debate is not critical. Nonetheless, some discussion of the issues seems appropriate, since these issues are relevant to use of collinearity diagnostics in actual research. There are two aspects to this debate, one theoretical, the other practical.

Theoretical aspects. Centering a variable is an example of a linear transformation. Centering involves, for each column of the data matrix, subtracting the mean of the column and (usually) dividing by its standard deviation (Marquadt, 1980). No one recommends it when there is no constant term in the model; controversy only arises when there is such a term (Snee & Marquadt, 1984).

Centering removes the correlation between the constant term and all linear terms. It is clear that, in so doing, centering the data reduces collinearity, sometimes substantially; Marquadt and Snee (1975) give an example where the largest VIF is reduced by a factor of 1,000. In

addition, Belsley (1984a) provides an artificial data set which is "astronomically" (p. 73) collinear when not mean centered, but exactly orthogonal after mean centering. What is unclear is whether this reduction in collinearity is "real" or whether the problems related to collinearity are still there, but somehow "hidden".

Centering solves neither of the two key problems caused by collinearity: 1) That small changes in the data produce large changes in the estimates of the parameters. 2) That the variances of the estimates are inflated.

The second of these problems cannot be solved by centering or by any other linear transformation, since the variance is unaffected by such transformations. Belsley (1984a) shows that, for his data set, the first problem is also not solved by centering despite the fact that the centered data are (or appear to be) orthogonal. Just as a 1% change in the uncentered data produces a 40% change in the parameter estimates, so does a 1% change in the centered data produce a 40% change in the parameter estimates for the centered data. Thus, Belsley (1984a) argues that centering masks essential collinearity and should be avoided.

Snee and Marquadt (1984) argue that the key issue in centering is not what Belsley calls the "basic data" but

rather the domain of prediction, i.e., the range of "the predictor variables over which one wants to make predictions" (p. 83). Their recommendation is that mean centering should be applied when the domain of prediction includes the natural origin of the predictor variables. They further state that since the behavior of the model at the natural intercept point is frequently of little interest, the constant term is often little more than a nuisance factor. In Belsley's example, the data are all close together and far from 0; therefore, according to Snee and Marquadt (1984), the fact that the origin is poorly predicted is irrelevant.

In his reply to these comments Belsley (1984b) points out that while the implied domain is relevant for estimation of the model, it is usually irrelevant for assessing conditioning. The key factor, for Belsley, is structural interpretability. One of the benefits of determining $\kappa$ (the condition number, or largest condition index) is that it can be used to determine how sensitive the parameters are to changes in the data; changing the data by x% can change the parameters by, at most, $\kappa$x%. But, if the data are centered, then this applies to a change in the centered data, which is rarely substantively meaningful. Further, if the underlying model includes an

intercept term, then collinearity involving that term is relevant.

Stewart (1987) also argues for centering when the data include a constant term:

> In order for centering to have a gross effect on the diagnostic, some variable $x_i$ must have a large constant part, and in fact, the larger the constant part the more 'important' the variable becomes. Now a large constant part is usually an artifice of the way the data are collected, especially in the sciences where it is not uncommon to make very precise measurements over a narrow range. In these cases it is appropriate to regard the 'importance' of such a variable as equally artificial. Otherwise put, the real variable is masked by the large constant part. Centering simply shows the variable for what it is (p. 75).

Practical aspect. The practical aspect of the centering debate is its effect on the numerical accuracy of the results. Simon and Lesage (1988a) found that tolerances[4] and condition numbers calculated based on a

[4] The tolerance is, as noted above, the inverse of the VIF.

centered data set were unable to diagnose collinearity between an IV and the intercept. Belsley (1986) suggests that "better conditioned centered data imply a poorly conditioned centering transform, and computationally this could be out of the frying pan into the fire" (p. 152). Some support for this view, at least in certain circumstances, is given by Simon and Lesage (1988a), who found that the centering transform can involve substantial cancellation error, which is a prime source of numerical problems in computer programs (Press, Teukolsky, Vetterling, & Rannery, 1992). Cancellation error occurs because of the numerical inaccuracy of computers. When very nearly equal numbers are compared, with interest centering on the difference between them, a result of 0 may occur if the difference is sufficiently small.

On a practical basis, even those who support centering might be troubled by the fact that diagnostics based on a centered matrix cannot reveal potential numerical inaccuracies which result from either the collinearity with the intercept term or the transformation which eliminates it.

Interactions

Models which are not strictly linear pose special problems for collinearity diagnostics. Belsley (1991)

points out that, when assessing nonlinear models (such as those which include interactions) "both the data and the nature of the model must be considered" (p. 354). While there are cases where orthogonal data lead to very poorly conditioned estimators, the data which Belsley (1991) presents to demonstrate this are highly contrived.

Given the fact that many psychological models involve interaction terms, it is important to determine whether collinearity diagnostics work well in typical situations. If they do not work well, then further investigation is needed into methods which will work. If, on the other hand, they do work in most situations, then the warnings raised by Belsley (1991) may be only theoretically interesting.

## Purpose and Rationale

The purpose of this dissertation was to extend our knowledge of collinearity diagnostics in several directions. First, the diagnostic procedures developed by Belsley (1991; Belsley, Kuh, & Welsch, 1980) were used on data where the IVs are normally distributed, rather than uniformly distributed. Psychological research frequently involves data which are assumed to be normally distributed. Belsley's examples, however, involve uniformly distributed data. Second, Belsley's procedure was applied to models

which include an interaction between two of the IVs. Third, Belsley's methods were compared to VIF measures. Belsley (1991) presents reasons that his methods are theoretically superior; it is not yet known whether this superiority manifests itself under conditions typical of psychological research.

Fourth, distributions for Belsley's diagnostics and VIFs were established for certain conditions. Belsley (1991) notes that he makes "no attempt ...to infer any distributional properties through repeated sampling" (p. 81). Establishing these distributions allows researchers to make better use of the diagnostics. Belsley (1991) suggests that cutoff values to diagnose collinearity be determined empirically, and establishing the distributions is part of this process.

An additional purpose of this dissertation was to inform researchers of some of the issues regarding collinearity and its ill effects, and of the inadequacy of some commonly used methods for diagnosing it.

## Conceptual Hypotheses

1. Belsley's method would distinguish among weak, moderate, and strong collinearity and will correctly identify which variables are involved in the collinear relation when used with normally distributed variables.

2. For both Belsley's methods and VIFs, errors in either determining the presence of collinearities or the variables involved in them will be more common in cases involving an interaction term. This is so for the reasons discussed in the section titled "Interactions", above.

3. Belsley's method would prove equal or superior to VIFs in all conditions.

# CHAPTER II

## METHOD

This dissertation evaluated two collinearity diagnostics (VIFs and Belsley's method) when three parameters were varied. These parameters were number of IVs in the regression (3, 5, or 7), presence or absence of an interaction term, and degree of collinearity. Each model included an intercept term. There were several reasons for choosing these parameters.

First, a literature review of articles using regression that were published in 1996 in any of four major psychological journals (Developmental Psychology, Journal of Applied Psychology, Journal of Consulting and Clinical Psychology, and Journal of Personality and Social Psychology) indicated that this range of number of IVs is fairly typical of psychological research, and that much psychological research involves models with an interaction. Of 55 articles, 31 included at least one regression with between 3 and 7 IVs, and 19 involved at least one interaction.

Second, evaluation of models with an interaction term extends Belsley's results. Third, varying the degree of collinearity allows us to assess the performance of the

different diagnostics. Fourth, the majority of articles reviewed used an intercept term in the model.

## Belsley's Method

Belsley's proposed diagnostic can be outlined as follows: 1) Obtain scaled condition indexes and variance-decomposition proportions

2) Determine number of near dependencies

3) Determine variate involvement

The condition indexes $\eta_k$ are computed as follows. First, the p singular values $\mu_k$ of the p x p data matrix X are computed.[5] Then

[2.1] $$\eta_k = \mu_{max} / \mu_k, \quad k = 1,2,\ldots p$$

where $\mu_k$ is the $k^{th}$ largest singular value. Each large condition index is an indication of a near dependency. Thus far, the technique is not original with Belsley. The condition number has been widely recommended as a collinearity diagnostic (e.g., Lesage & Simon, 1985).

The variance-decomposition proportions ($\pi_{jk}$) are indicators of which variables are involved in each of the

---

[5] The singular values of X are the diagonal elements of D where X = UDV' is the singular value decomposition (see Appendix B).

near dependencies identified by the condition indexes. They are[6]:

$$[2.2] \qquad \pi_{jk} = \phi_{kj}/\phi_k, \; k,j = 1, 2, \ldots p$$

where

$$[2.3] \qquad \phi_{kj} = v_{kj}^2 / \mu_j^2 \text{ and } \phi_k = \Sigma \phi_{kj}, \; k = 1,2,\ldots p$$

and $v_{kj}$ are the elements of V in the singular value decomposition of X.

The number of near dependencies is equal to the number of "high" condition indexes. No precise definition of high exists, although Belsley (1991) suggests that any which exceed 30 may indicate problems. One of the goals of this dissertation is to better define high in this context.

In step 3, a variate is considered degraded by collinearity if 1) it is involved in at least one near dependency and 2) the total proportion of its variance associated with the dependencies it is involved in exceeds some threshold. Again, this threshold is, at present, somewhat arbitrary; here Belsley (1991) suggests .5. Another goal of this dissertation is to better define a proportion which indicates variate involvement.

---

[6] The switch in the order of the subscripts between the left and right sides of equation 8 is adopted to render the output of the diagnostic more readable.

Belsley's method generates considerable output. This output is typically formatted as in Table 1 (assuming there are 3 IVs plus an intercept term, $b_0$).

Table 1

Sample Output of Belsley's Diagnostics

| Condition index | Var($b_0$) | Var($b_1$) | Var($b_2$) | Var($b_3$) |
|---|---|---|---|---|
| $\eta_1$ | $\pi_{10}$ | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ |
| $\eta_2$ | $\pi_{20}$ | $\pi_{21}$ | $\pi_{22}$ | $\pi_{23}$ |
| $\eta_3$ | $\pi_{30}$ | $\pi_{31}$ | $\pi_{32}$ | $\pi_{33}$ |
| $\eta_4$ | $\pi_{40}$ | $\pi_{41}$ | $\pi_{42}$ | $\pi_{43}$ |

## Data

In this 3 (number of IVs) x 2 (presence of interaction) x 3 (degree of collinearity) factorial ANOVA design, data matrices were generated which have different numbers of IVs (3, 5, or 7), presence or absence of an interaction, and weak, moderate, or strong collinearity. These matrices were used to test the various diagnostic measures. The two diagnostic procedures that were evaluated are VIFs and Belsley's Condition Indexes ($\eta$). Both of these were evaluated using SAS (see Appendix D). The degree of collinearity was adjusted using the method used by Belsley, as follows: An exactly collinear term was

computed, and varying degrees of random error, or noise, was added to it. When more noise is added to the collinear term, the collinearity is less strong.

Data were generated using SAS call routines for random number generators for uniform (RANUNI) and normal (RANNOR) distributions (SAS Institute, 1991).

## Models

Three parameters were varied: Number of IV's (3, 5, or 7), presence or absence of an interaction term, and degree of collinearity (weak, moderate, or strong). This 3 x 2 x 3 design yields 18 models (cells), which are summarized in Tables 2 and 3.

In each model, each of the IVs was generated to be normally distributed with mean 0 and standard deviation 1. A mean of 0 was chosen because it automatically centers the variables, thus avoiding the centering controversy discussed previously. Both VIFs and Belsley's procedure scale the data matrix to have a standard deviation of 1. In view of this, a standard deviation of 1 was chosen for the input data. Since changes in the sample size do not affect the diagnostics (Belsley, 1991) each model used an $\underline{N}$ of 100, which was seen as being fairly typical of psychological research. In each model, one collinear term was added to the model.

Table 2

Models without an Interaction Term

| | Weak collinearity | Moderate collinearity | Strong collinearity |
|---|---|---|---|
| 3 IVs | Model 1 | Model 2 | Model 3 |
| 5 IVs | Model 4 | Model 5 | Model 6 |
| 7 IVs | Model 7 | Model 8 | Model 9 |

Table 3

Models with an Interaction Term

| | Weak collinearity | Moderate collinearity | Strong collinearity |
|---|---|---|---|
| 3 IVs | Model 10 | Model 11 | Model 12 |
| 5 IVs | Model 13 | Model 14 | Model 15 |
| 7 IVs | Model 16 | Model 17 | Model 18 |

Degree of collinearity was adjusted using Belsley's method. An exactly collinear term was generated, and then different amounts of random error, or noise, were added to it. When little noise is added, collinearity is strong; as more noise is added, the collinearity becomes weaker. Numerical details of this procedure as follows.

Models 1, 2 and 3 were generated by formula 2.4 (below); models 4, 5, and 6 by formula 2.5; models 7, 8, and 9 by formula 2.6; models 10, 11 and 12 by formula 2.8;

models 13, 14, and 15 by formula 2.9; and models 16, 17 and 18 by formula 2.10.

[2.4] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_cX_c + e$$

[2.5] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_cX_c + e$$

[2.6] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_cX_c + e$$

[2.7] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_5X_1X_2 + b_cX_c + e$$

[2.8] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_1X_2 + b_cX_c + e$$

[2.9] $$Y = b_0X_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_1X_2 + b_cX_c + e$$

where $X_c$ is the collinear term:

[2.10] $$X_c = X_1 + X_2 + e_i$$

and

[2.11] $$e_i \sim N(0, 10^{-i}s^2_uI)$$

[2.12] $s^2_u = var(X_1 + X_2)$ and i = 1, 2, or 3, for weak, moderate, and strong collinearity, respectively[7].

## Procedure

For each of the specified models 1,000 data matrices were generated. Each matrix was analyzed for collinearity using two methods: Belsley's collinearity diagnostics and VIFs. Matrix generation and analyses were performed using SAS (see Appendix C).

While neither VIFs nor, especially, Belsley's procedure are intended to be used in a rigid way, some cutoff must be established. The most commonly recommended levels for VIFs are 5 and 10 (tolerances of 0.2 - 0.1) (Marquadt, 1980; Montgomery & Peck, 1982; Snee, 1973) and these were adopted here as cutoffs for moderate and strong collinearity, respectively. Belsley (1991) suggests that any $\eta$ over 10 indicates moderate collinearity, and any over

---

[7] Belsley (1991) typically allows i to vary from 0 to 4. However, the extreme cases are the least taxing to any diagnostic. A method which detects strong collinearity will likely detect even stronger collinearity, and very weak collinearity does not pose severe problems to regression.

30 indicates fairly strong collinearity, these standards were also be adopted here.

As Belsley (1991) points out, for there to be collinearity at least two variables must be involved in the collinear relation. His method establishes this through the variance-decomposition proportions $\pi_{ij}$, and he suggests that if two or more variables have $\pi$s over 0.5 associated with the same high $\eta$, collinearity may exist.

Operational Hypotheses

1a. Within each set of models (i.e., models 1, 2, and 3; 4, 5, and 6; 7, 8, and 9; 10, 11, and 12; 13, 14, and 15; and 16, 17 and 18) the condition number $\kappa$ would be significantly higher[8] for moderate than for weak collinearity, and higher for strong collinearity than for moderate collinearity.

1b. The mean $\kappa$ would be over 30 for strong collinearity.

1c. The variance-decomposition proportions $\pi_{ij}$ associated with the highest $\eta_p$ would be highest for the two variables involved in the collinear relation.

---

[8] While there is no formal significance test for $\kappa$, the data generated can be used to estimate the standard errors, and approximate significance tests.

1d. The variance-decomposition proportions $\pi_{ij}$ associated with the variables involved in the collinearity would have means over 0.5.

1e. The variance-decomposition proportions $\pi_{ij}$ associated with the all other variables would have means lower than 0.5.

2. Any cases where hypotheses 1a-1e are not supported would be those where the model includes an interaction term (i.e. models 10-18).

3. In models 2, 3, 5, 6, 8, 9, 11, 12, 14, 15, 17 and 18, if the highest VIF is over 10, then $\eta_p$ would be over 30.

# CHAPTER III

## RESULTS

This dissertation measured the effects of variation in three parameters on two collinearity diagnostics for multiple regression. The three parameters were number of variables (3, 5, or 7), number of interactions (0 or 1), and degree of collinearity (weak, moderate, or strong). This 3-way design yielded 18 models; each model was replicated 1000 times. The two collinearity diagnostics were Variance Inflation Factors (VIFs) and Condition Indexes ($\eta$).

VIFs can be shown to be

[4, repeated] $VIF_i = 1/(1-R_i^2)$

where $R_i^2$ is the multiple correlation of $X_i$ regressed on the remaining IVs. Condition indexes are the ratios of each of the singular values of the data matrix to the smallest singular value.

Belsley (1991) suggested reasons why condition indexes are superior to VIFs, but offered no supporting experimental evidence. The principal purposes of this dissertation were 1) to compare the ability of condition indexes and VIFs to diagnose collinearity under the conditions listed above and 2) to further evaluate the performance of condition indexes under the same conditions.

This chapter is organized as follows: First, the distribution of the largest VIF and the largest condition index are compared for each of the 18 models. The largest values are used because they are the key to diagnosing collinearity; if the largest value does not indicate a problem, no smaller value can do so. Next, the ability of the condition index to correctly identify the degree of collinearity and variate involvement is determined. Third, the effect of having an interaction among the IVs is evaluated. Fourth, the performance of condition indexes in determining the degree of collinearity is compared to that of VIFs. Fifth, the degree of dispersion of the two statistics is compared. Finally, the results are summarized.

## Sampling Distribution of the Statistics

The distributions of the largest condition indexes and VIFs for the various models are shown in Figures 2 - 37 (see Appendix E) and in Tables 4 and 5, below. Table 4 lists the mean, standard deviation, skewness, kurtosis, and Shapiro-Wilks (SW) test of normality for the largest condition index ($\eta$) for each model, and Table 5 does so for the largest VIF.

Skewness is a measure of the tendency of deviations from the mean to be larger in one direction than the other.

The normal distribution has skewness = 0. Kurtosis is a measure of the heaviness of the tails and shape of a distribution. The normal distribution has kurtosis = 0.

The SW test is one of the most powerful omnibus tests of the null hypothesis that the sample comes from a normal population (Royston, 1988). It was developed by Shapiro and Wilk (1965), and later extended to large samples by Royston (1982). It is bounded by 0 and 1, and smaller values indicate greater nonnormality (D'Agostino, 1982). For further details, see D'Agostino (1988), Royston (1980, 1982), and Shapiro and Wilk (1965).

Based on the SW statistic, all of the distributions of the condition indexes and the VIFs are statistically different from normal. This is due, in large part, to the large sample sizes. In every case, the SW test is closer to 1 for the condition index than for the VIF, indicating that the condition index is closer to normality (e.g., for the model with no interaction, 3 IVs, and weak collinearity, the SW test is 0.95 for the condition index and 0.88 for the VIF).

Figures 2 - 37 show the degree of skewness and kurtosis graphically. Each page shows the distribution of the condition index and the VIF for one model. For several of the models (e.g., the models with no interaction, three

IVs, and moderate collinearity) it is apparent that the VIF is more skewed than the condition index. None of the distributions are "grossly" nonnormal; i.e., they are all unimodal, they all have modes which are near the center of the distribution, and none have any obvious outliers, by visual inspection.

For both diagnostics, the distributions for the models with no interaction and three IVs were markedly more skewed and had higher kurtosis than those for the other models. Apart from this, there were no clear patterns within each diagnostic. For each of the 18 models, the VIF had a more skewed and leptokurtotic distribution than the condition index did (e.g., for the model with no interaction, 7 IVs, and moderate collinearity, the largest condition index had a skewness of 0.43, and kurtosis of 0.46; the largest VIF had skewness of 0.71 and kurtosis of 1.02).

## Performance of Condition Indexes

### Ability to Determine Existence of Collinearities

These results are summarized in Tables 4, 6, 7, and 8. Table 4 contains statistics for all models; Table 6 contains additional results for models with weak collinearity, Table 7 for models with moderate collinearity and Table 8 for models with strong collinearity.

Table 4

Mean, Standard Deviation, Skewness, Kurtosis, and Shapiro-Wilks Test of the Largest Condition Index, 18 Models, N = 1000 For Each Model.

| Model | Mean | SD | Skewness | Kurtosis | SW test |
|---|---|---|---|---|---|
| No interaction | | | | | |
| 3 IVs | | | | | |
| Weak collin. | 8.04 | 1.63 | 1.11 | 2.87 | 0.95 |
| Mod. collin. | 24.91 | 5.07 | 1.01 | 1.70 | 0.94 |
| Strong collin. | 79.14 | 16.10 | 1.00 | 1.62 | 0.94 |
| 5 IVs | | | | | |
| Weak collin. | 6.90 | 0.56 | 0.45 | 0.40 | 0.98 |
| Mod. collin. | 21.28 | 1.59 | 0.45 | 0.30 | 0.98 |
| Strong collin. | 67.11 | 5.31 | 0.63 | 1.14 | 0.98 |
| 7 IVs | | | | | |
| Weak collin. | 7.01 | 0.56 | 0.56 | 0.32 | 0.98 |
| Mod. collin. | 21.75 | 0.56 | 0.43 | 0.46 | 0.98 |
| Strong collin. | 68.25 | 5.28 | 0.39 | 0.17 | 0.98 |
| Interaction | | | | | |
| 3 IVs | | | | | |
| Weak collin. | 6.84 | 0.55 | 0.37 | 0.48 | 0.98 |
| Mod. Collin. | 21.19 | 1.63 | 0.32 | 0.07 | 0.98 |
| Strong collin. | 66.91 | 5.08 | 0.36 | 0.42 | 0.98 |
| 5 IVs | | | | | |
| Weak collin. | 6.97 | 0.55 | 0.42 | 0.37 | 0.98 |
| Mod. Collin. | 21.64 | 1.64 | 0.37 | 0.01 | 0.98 |
| Strong collin. | 68.11 | 5.30 | 0.50 | 0.64 | 0.98 |
| 7 IVs | | | | | |
| Weak collin. | 7.16 | 0.57 | 0.37 | 0.58 | 0.98 |
| Mod. Collin. | 22.09 | 1.78 | 0.33 | 0.10 | 0.98 |
| Strong collin. | 69.71 | 5.73 | 0.35 | 0.10 | 0.98 |

Table 5

Mean, Standard Deviation, Skewness, Kurtosis, and Shapiro-Wilks Test of the Largest VIF, 18 Models, N = 1000 for Each Model.

| Model | Mean | SD | Skewness | Kurtosis | SW test |
|---|---|---|---|---|---|
| No interaction | | | | | |
| 3 IVs | | | | | |
| Weak collin. | 14.64 | 5.78 | 1.91 | 7.34 | 0.88 |
| Mod. collin. | 137.87 | 56.67 | 1.94 | 6.60 | 0.86 |
| Strong collin. | 1380.50 | 557.13 | 1.57 | 3.66 | 0.88 |
| 5 IVs | | | | | |
| Weak collin. | 11.84 | 1.85 | 0.79 | 1.15 | 0.96 |
| Mod. collin. | 108.63 | 15.82 | 0.61 | 0.51 | 0.96 |
| Strong collin. | 1077.51 | 169.20 | 0.93 | 1.96 | 0.96 |
| 7 IVs | | | | | |
| Weak collin. | 7.01 | 0.56 | 0.60 | 0.55 | 0.97 |
| Mod. collin. | 111.51 | 17.05 | 0.71 | 1.02 | 0.97 |
| Strong collin. | 1091.24 | 165.53 | 0.61 | 0.48 | 0.97 |
| Interaction | | | | | |
| 3 IVs | | | | | |
| Weak collin. | 11.62 | 1.73 | 0.63 | 0.96 | 0.97 |
| Mod. collin. | 107.53 | 15.69 | 0.46 | 0.22 | 0.98 |
| Strong collin. | 1064.66 | 157.94 | 0.65 | 1.15 | 0.97 |
| 5 IVs | | | | | |
| Weak collin. | 11.84 | 1.74 | 0.64 | 0.66 | 0.97 |
| Mod. collin. | 109.62 | 16.05 | 0.57 | 0.28 | 0.97 |
| Strong collin. | 1082.08 | 163.90 | 0.67 | 0.80 | 0.97 |
| 7 IVs | | | | | |
| Weak collin. | 12.18 | 1.81 | 0.62 | 0.73 | 0.97 |
| Mod. collin. | 112.50 | 17.08 | 0.48 | 0.05 | 0.97 |
| Strong collin. | 1116.85 | 175.04 | 0.58 | 0.37 | 0.97 |

Hypothesis 1a stated that, within each set of models, (set 1 is those models with no interaction and 3 IVs, set 2 no interaction and 5 IVs, set 3 no interaction and 7 IVs, set 4 interaction and 3 IVs, set 5 interaction and 5 IVs and set 6 interaction and 7 IVs) the highest condition

index would be significantly higher for moderate collinearity than for weak, and higher for strong collinearity than moderate. This hypothesis was confirmed (see Table 4). For each set of models, the hypothesized pattern holds (e.g., for the models with interaction and 5 IVs, the mean largest condition index is 6.97 for weak collinearity, 21.64 for moderate collinearity, and 66.91 for strong collinearity). While these differences were all statistically significant, this is due to the large sample sizes. Rather than report the results of an ANOVA and corresponding p-values (the F statistics for the various effects were all over 100), I simply note that there are large differences when the degree of collinearity is varied, and small differences when either of the other conditions are varied. The mean largest condition index for models with weak collinearity was 7.15, for moderate collinearity, 22.14, and for strong collinearity, 69.87. For models with 3 IVs the mean largest condition index was 34.51, for 5 IVs, 32.00, and for 7 IVs, 32.66. Finally, for models with no interaction the mean largest condition index was 32.52, and for models with an interaction it was 32.29.

Hypothesis 1b stated that, for models with strong collinearity, the mean of the largest condition index would

be over 30 (see Table 8). All models with strong collinearity had mean condition indexes over 30 (they ranged from 66.91 - 79.14).

Ability to Determine Variable Involvement

In addition to being able to diagnose the degree of collinearity, Belsley (1991) states that the condition indexes are able to determine which variables are involved in the collinear relations. As described in Chapter 1, Belsley (1991) suggests that if two or more variables have variance decomposition proportions[9] ($\pi$s) over 0.5 associated with the same high $\eta$, collinearity may exist. In this dissertation collinearities were created with known variable involvement. Specifically, only IVs 1 and 2 were involved in the collinearity. Therefore, only these variables should have high variance decomposition proportions associated with a high condition index (see the discussion of Hypotheses 1c, 1d, and 1e, immediately below).

[9] The variance decomposition proportion is the proportion of the variance in the variable explained by that collinearity.

Table 6

Performance of Condition Indexes

Models with Weak Collinearity

| Model | Mean Condition Index | SD | Range |
|---|---|---|---|
| No interaction | | | |
| 3 IVs | 8.04 | 1.63 | 4.18 - 18.51 |
| 5 IVs | 6.90 | 0.56 | 5.37 - 9.19 |
| 7 IVs | 7.01 | 0.56 | 5.06 - 8.99 |
| Interaction | | | |
| 3 IVs | 6.84 | 0.55 | 5.44 - 9.21 |
| 5 IVs | 6.97 | 0.55 | 5.52 - 9.43 |
| 7 IVs | 7.16 | 0.57 | 5.72 - 9.83 |

Table 7

Performance of Condition Indexes

Models with Moderate Collinearity

| Model | Mean Condition Index | SD | Range |
|---|---|---|---|
| No interaction | | | |
| 3 IVs | 24.91 | 5.07 | 15.28 - 50.80 |
| 5 IVs | 21.28 | 1.59 | 17.46 - 27.78 |
| 7 IVs | 21.75 | 1.71 | 16.96 - 28.04 |
| Interaction | | | |
| 3 IVs | 21.19 | 1.63 | 16.35 - 26.55 |
| 5 IVs | 21.64 | 1.64 | 17.44 - 27.06 |
| 7 IVs | 22.09 | 1.78 | 17.93 - 28.32 |

Table 8

Performance of Condition Indexes

Models with Strong Collinearity

| Model | Mean Condition Index | SD | Range |
|---|---|---|---|
| No interaction | | | |
| 3 IVs | 79.14 | 16.10 | 48.83 - 162.88 |
| 5 IVs | 67.11 | 5.31 | 50.02 - 93.85 |
| 7 IVs | 68.25 | 5.28 | 54.19 - 89.49 |
| Interaction | | | |
| 3 IVs | 66.91 | 5.08 | 53.15 - 86.94 |
| 5 IVs | 68.11 | 5.30 | 54.50 - 95.93 |
| 7 IVs | 69.71 | 5.73 | 55.01 - 91.90 |

Hypothesis 1c stated that the $\pi_{ij}$ associated with the highest condition index would be highest for the two variables involved in the collinear relation ($b_1$ and $b_2$). This hypothesis was confirmed. Tables 9 and 10 list the mean and standard deviation for each variance decomposition proportion associated with the highest condition index for each model. Table 9 lists results for models without an interaction, and Table 10 lists results for models with an interaction. In every case, the largest $\pi_{ij}$ are associated with variable $b_1$ and $b_2$ (e.g., for the model with no interaction, 3 IVs, and moderate collinearity, the $\pi_{ij}$ associated with $b_1$ and $b_2$ are both 0.99, whereas the $\pi_{ij}$ associated with $b_3$ is 0.06).

Hypothesis 1d stated that the $\pi_{ij}$ associated with the highest condition index (and with the variables involved in the collinear relations) will have means over 0.5. This hypothesis was also confirmed. For each model, the means of $\pi_{1j}$ and $\pi_{2j}$ associated with the largest condition index are over 0.5, they range from 0.89 (for each of the models with weak collinearity) to 1.00 (for each of the models with strong collinearity) (see Tables 9 & 10).

Hypothesis 1e stated that the $\pi_{ij}$ associated with the other variables and the highest condition index will have means lower than 0.5. This hypothesis was also confirmed for every model (e.g., for the model with no interaction, 3 IVs, and strong collinearity, the variance proportions associated with the variable not involved in the collinearity ($b_3$) was 0.05.

Table 9

Variance Proportions ($\pi_{Ij}$)[10] Associated With Largest Condition Index, Models with no Interaction

| Model | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ | $b_7$ |
|---|---|---|---|---|---|---|---|---|
| 3 IVs, weak collinearity | | | | | | | | |
| Mean | 0.06 | 0.89 | 0.89 | 0.06 | | | | |
| SD | 0.08 | 0.08 | 0.07 | 0.08 | | | | |
| 3 IVs, moderate collinearity | | | | | | | | |
| Mean | 0.06 | 0.99 | 0.99 | 0.06 | | | | |
| SD | 0.08 | 0.01 | 0.01 | 0.07 | | | | |
| 3 IVs, strong collinearity | | | | | | | | |
| Mean | 0.06 | 1.00 | 1.00 | 0.05 | | | | |
| SD | 0.08 | 0.00 | 0.00 | 0.07 | | | | |
| 5 IVs, weak collinearity | | | | | | | | |
| Mean | 0.01 | 0.89 | 0.89 | 0.01 | 0.01 | 0.01 | | |
| SD | 0.02 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | | |
| 5 IVs, moderate collinearity | | | | | | | | |
| Mean | 0.01 | 0.99 | 0.99 | 0.01 | 0.01 | 0.01 | | |
| SD | 0.02 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | | |
| 5 IVs, strong collinearity | | | | | | | | |
| Mean | 0.01 | 1.00 | 1.00 | 0.01 | 0.01 | 0.01 | | |
| SD | 0.02 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | | |
| 7 IVs, weak collinearity | | | | | | | | |
| Mean | 0.01 | 0.89 | 0.89 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.02 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 7 IVs, moderate collinearity | | | | | | | | |
| Mean | 0.01 | 0.99 | 0.99 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.02 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 7 IVs, strong collinearity | | | | | | | | |
| Mean | 0.01 | 1.00 | 1.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 |

[10] Each column contains the values associated with the corresponding parameter. Thus $b_0$ contains $\pi_{0j}$.

Table 10

Variance Proportions ($\pi_{ij}$) Associated with Largest Condition Index, Models with an Interaction

| Model | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ | $b_7$ | $b_m$ |
|---|---|---|---|---|---|---|---|---|---|
| 3 IVs, weak collinearity | | | | | | | | | |
| Mean | 0.01 | 0.89 | 0.89 | 0.01 | | | | | 0.01 |
| SD | 0.02 | 0.03 | 0.03 | 0.02 | | | | | 0.02 |
| 3 IVs, moderate collinearity | | | | | | | | | |
| Mean | 0.01 | 0.99 | 0.99 | 0.01 | | | | | 0.01 |
| SD | 0.02 | 0.00 | 0.00 | 0.02 | | | | | 0.02 |
| 3 IVs, strong collinearity | | | | | | | | | |
| Mean | 0.01 | 1.00 | 0.01 | 0.01 | | | | | 0.01 |
| SD | 0.01 | 0.00 | 0.01 | 0.01 | | | | | 0.01 |
| 5 IVs, weak collinearity | | | | | | | | | |
| Mean | 0.01 | 0.89 | 0.89 | 0.01 | 0.01 | 0.01 | | | 0.01 |
| SD | 0.02 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | | | 0.02 |
| 5 IVs, moderate collinearity | | | | | | | | | |
| Mean | 0.01 | 0.99 | 0.99 | 0.01 | 0.01 | 0.01 | | | 0.01 |
| SD | 0.01 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | | | 0.01 |
| 5 IVs, strong collinearity | | | | | | | | | |
| Mean | 0.01 | 1.00 | 1.00 | 0.01 | 0.01 | 0.01 | | | 0.01 |
| SD | 0.02 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | | | 0.01 |
| 7 IVs, weak collinearity | | | | | | | | | |
| Mean | 0.01 | 0.89 | 0.89 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.02 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 7 IVs, moderate collinearity | | | | | | | | | |
| Mean | 0.01 | 0.99 | 0.99 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.02 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 7 IVs, strong collinearity | | | | | | | | | |
| Mean | 0.01 | 1.00 | 1.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.02 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |

## Effect of an Interaction Term on the Performance of the Condition Index

Hypothesis 2 was intended to test whether an interaction term affected the performance of the condition index. This hypothesis stated that any cases where Hypotheses 1a - 1e were not supported would be those where the model included an interaction term. Since there were no cases where Hypotheses 1a - 1e were not supported, this hypothesis is neither supported nor rejected. Condition indexes were not affected by the presence of an interaction term (see Tables 4 & 5 for detection of collinearity, and 9 and 10 for determination of variate involvement). That is, the results for the models with an interaction (shown in Tables 5 & 10) were not meaningfully different from those for models without an interaction (in Tables 4 & 9).

## Comparison of Condition Indexes and VIFs

Hypothesis 3 stated that in the models with moderate or strong collinearity, if the highest VIF is over 10, then the highest condition index would be over 30. This hypothesis was confirmed for strong collinearity (e.g., for the model with no interaction, 5 IVs, and strong collinearity, the mean VIF was 10787.51 and the mean condition index was 67.11). It was rejected for moderate collinearity (e.g., for the model with no interaction, 5

IVs and moderate collinearity, the mean VIF is 108.63, but the mean condition index is 21.28).

Dispersion of the Statistics

Another area where condition indexes can be compared to VIFs is in the dispersion of the statistic. In terms of the coefficient of variation[11] (CV), the condition index is less disperse than the VIFs. The CVs for the largest condition index range from 0.07 to 0.20, with a median of 0.08. For the largest VIF, the CVs range from 0.15 to 0.41 with a median of 0.16.

The proper statistical test of a difference in distribution depends on whether we are willing to assume that the underlying distributions are normal. Evidence of the normality of the statistics was discussed above; while the Shapiro-Wilks tests were significantly different from 1 for every model, these differences were small in most cases. I therefore examine tests of difference in distribution which assume normality and which do not.

---

[11] The coefficient of variation is the standard deviation divided by the mean. It should be noted that some sources, including SAS, define the CV as the above multiplied by 100.

If we are willing to assume normality, then we can test the difference in variations using the coefficient of variation. McKay (1932) proposed a method for determining approximate confidence intervals for a coefficient of variation from a normally distributed population (Fieller, 1932; Iglewicz & Myers, 1970; McKay, 1932; Pearson, 1932; Vangel, 1996). Using McKay's test, the condition index has a significantly smaller CV than VIF for all models.

If we are not willing to assume normality, there are several nonparametric tests of dispersion available (Hollander & Wolfe, 1973). All of these, however, assume (either explicitly or implicitly) that the variables are measured on the same scale (Gibbons, 1985; Sprent, 1998). If we are not willing to assume normality, and if (as appears to be the case here) the variables are not measured on the same scale, then a procedure devised by Lewontin (1966) is useful. Lewontin (1966) showed that, while the variance of a variable is affected by changes in scale, the variance of the logarithm (to any base) is not. Thus, he proposed testing the difference between two variances using the $F$ test on the ratio of the variances of the logs. Using this test, $F_{999,999}$ ranges from 2.6 - 3.9, all of which are significant at well below $p < .001$.

Thus, regardless of whether we assume that the condition indexes and VIFs are normally or not, we conclude that the condition indexes are less dispersed than the VIFs for all models.

## Summary

Both VIFs and Condition Indexes can diagnose the degree of collinearity in multiple regression (see Tables 4 & 5). Condition Indexes can, in addition, determine the number of collinear relations and the variables that are involved in each collinear relation (see Tables 9 & 1). Condition Indexes are less dispersed than VIFs, even after accounting for the differences in the means by using the coefficient of variation.

# CHAPTER IV

# DISCUSSION

This dissertation examined the performance of two collinearity diagnostics for multiple regression when three parameters were varied. These parameters were: Presence or absence of an interaction; number of IVs in the regression equation (3, 5, or 7); and degree of collinearity (weak, moderate, or strong). This discussion is organized as follows: First, the performance of the two diagnostics is discussed. Second, limitations of the present research are examined. Third, suggestions for future research are put forth. Finally, recommendations for users of these diagnostics are made.

## Performance of the Diagnostics

Collinearity diagnostics should have several properties. First, they should be able to diagnose collinearity and determine its degree. Second, they should be able to determine the number of collinear relations. Third, they should be able to determine which variables are involved in the collinearities. Fourth, they should be as precise and uniform as possible, so that specific levels of the statistics correspond with specific levels of collinearity. Finally, these properties should not vary across different conditions other than degree of

collinearity. Each of these areas is discussed in turn, with the exception of determining the number of collinearities, which was not addressed in this dissertation.

Ability to Detect Collinearity and Determine its Degree

It is important to be able to detect collinearity and determine its degree because this indicates whether the collinearity may be affecting the parameter estimates of the regression. Condition indexes were able to detect and determine the degree of collinearity. For each set of models, the mean of the largest condition index was significantly larger for moderate collinearity than for weak collinearity, and significantly higher for strong collinearity than for moderate. For 17 of the 18 models considered in this dissertation, Belsley's recommendations of using 10 as a cutoff for moderate collinearity and 30 for strong were 100% accurate. The only model where they were not 100% accurate was the model with no interaction, weak collinearity, and 3 IVs. In 7.5% of the replications of this model, the guidelines indicated moderate collinearity. In the remaining 92.5%, they indicated the correct degree of collinearity.

It is anomalous that the condition index made incorrect decisions only for the simplest model that was

evaluated (i.e., no interaction, 3 IVs) for weak collinearity; in this case, the largest condition index was over 10 in 75 of 1000 cases. In an attempt to determine the cause of this anomaly, I ran the same procedure with a different starting seed, and got essentially similar results. The cause of this anomaly remains to be determined.

While there were clear differences in the levels of VIFs for the different degrees of collinearity, these did not coincide with the recommendations found in the research. Most writers recommend using 5 or 10 as indicators of moderate and severe collinearity (Marquadt, 1970; Montgomery & Peck, 1982; Snee, 1973) the results of this dissertation indicate that the appropriate values are substantially higher; based on the results shown in Table 7, a tentative recommendation is that 100 indicates moderate collinearity, and 1,000 indicates strong collinearity. This is important because using the guidelines from the literature will lead us to diagnose collinearity where none exists, or to overestimate the degree of collinearity where it does exist.

That the VIFs are higher than the corresponding condition indexes is not surprising, since the VIF is an upper bound on the largest condition index (see Chapter 1).

What is surprising is that the recommendations for diagnosing collinearity with VIFs are generally lower than the recommendations for doing so with condition indexes, since the VIFs are necessarily as high or higher than the VIFs. In deciding which set of recommendations to adjust, it should be noted that the recommendations for VIFs are based entirely on empirical results, while Belsley (1991) offers theoretical reasons why a data matrix with a condition index over 30 is often harmful to the regression estimates.

Despite the lack of agreement between the generally recommended levels and the results of this research, the fact that VIFs do, indeed, vary significantly depending on the degree of collinearity is evidence VIFs can determine the degree of collinearity for the models considered here. Therefore, the ability to determine degree of collinearity does not help us in choosing between methods.

Ability to Determine Variable Involvement

It is important to be able to determine which variables are involved in the collinear relations, because this helps focus any remedial measures that are necessary. For example, if it is the case that one or more of the variables involved in the collinearity can be dropped without serious substantive difficulties, then this may be

an ideal solution. This is impossible without knowing which variables are involved in the collinearity.

The condition indexes and variance proportions were able to determine the variables involved in all cases for all models (see Tables 11 & 12 in Chapter 3). VIFs are unable to do this, since they provide no information about variable involvement. This argues strongly for using Belsley's methods to diagnose collinearity.

## Precision of the Diagnostics

The precision of a statistic is important because a more precise statistic gives more information than a less precise one. With regard to collinearity diagnostics, this implies that a more precise statistic allows a better estimate of the degree of collinearity.

The results of this dissertation show that, for the models considered, condition indexes are more precise estimates of the degree of collinearity than VIFs are. This is further support for using Belsley's methods. This implies that, as empirical evidence of the relationship between the level of the diagnostics and the degree of problem caused to the regression equation accumulates, condition indexes will allow better estimates of the degree to which any particular regression equation is harmed by collinearity.

## Stability Across Models

Collinearity diagnostics should not vary across conditions other than degree of collinearity. If they did vary, then it would be impossible to establish general guidelines for their use. Belsley's recommendations of 10 for moderate collinearity and 30 for strong collinearity worked well across all the models considered here. For weak collinearity, the mean of the largest condition index ranged from 6.84 - 8.04, for moderate collinearity, from 21.19 - 24.91, and for strong from 66.91 - 79.14. This means that guidelines for the use of condition indexes do not depend on the model.

VIFs varied more across models than Condition Indexes did. For weak collinearity the mean of the largest VIF ranged from 7.01 - 14.64, for moderate collinearity it ranged from 107.53 - 137.87, and for strong collinearity from 1064.66 - 1380.50. The greater variation across models argues against using VIFs. The degree of variation is not very high even for VIFs i.e., the ranges 7.01 - 14.64, 107.53 - 137.87, and 1064.66 - 1380.50 are not large. It should be kept in mind, however, that this dissertation examined three quite distinct levels of collinearity, so the effect of changes on degree of collinearity was clear. In actual research, collinearity

varies along a continuum, and the less the diagnostics vary across conditions, the more precise recommendations for using them can be.

## Limitations of the Present Research

There are three principal limitations to this dissertation. First, it deals only with models with one collinear relation. Second, it deals only with models with 3, 5, or 7 IVs. Third, it does not deal with models with multiple interactions. Each of these is discussed in turn.

### Number of Collinear Relations

When there are multiple collinear relations, diagnosis can be considerably more complex. Belsley (1991) analyzes some individual cases with two collinearities, but, as with the single collinearity case, does not provide any sample distributions. His results show that condition indexes do not have any difficulty analyzing such cases, whereas VIFs will sometimes have difficulty. Sampling statistics for models with multiple collinearities need to be determined, and research is needed on cases with more than two collinearities.

### Number of IVs

While the range of 3 - 7 IVs is fairly typical of psychological research, models with more IVs may pose greater difficulties for collinearity diagnostics. Little

research has been done in this area. Among the models considered here, however, there was no evidence that more complex models pose any difficulty for VIFs or Condition Indexes than simpler ones. Both diagnostics were very similar as the number of IVs was varied (see Table 6 for condition indexes and Table 7 for VIFs). Indeed, the only anomalous results were for the simplest model, with 3 IVs and no interaction.

## Multiple Interactions

While the presence of an interaction did not affect the performance of either VIFs or Condition Indexes, this dissertation only examined an interaction between the two variables involved in the collinearity. It is possible that interactions involving variables not involved in the collinearity could affect the performance of the diagnostics. It is also possible that a single interaction involving more than two variables may affect the performance of the diagnostics. Since Belsley (1991) showed that even a single interaction can cause problems for collinearity diagnostics, more research is needed in this area to determine if there are other models where this occurs. In addition, research is needed into cases with multiple interactions because models with several interactions are frequently proposed in psychology.

## Additional Suggestions for Future Research

In addition to addressing the limitations discussed above, research is needed on the relationship between the condition index and the effect of collinearity on the regression estimates. Not all collinearities affect these estimates. A starting point for this research is Chapter 7 of Belsley (1991), Belsley (1982), Belsley and Oldford (1986), Gunst (1983), and Simon and Lesage (1988).

Belsley (1991) notes that, if the error variance is small enough, collinearity is not harmful. The higher the degree of collinearity, the less error variance can be tolerated, so that "small" is necessarily relative. He recommends the use of the ratio of the parameter to its error variance as a measure of "signal to noise" $\tau = B_i/\sigma_{bi}$ (note that this recommendation involves the use of parameters, rather than their estimators). The test is then whether this statistic is sufficiently different from zero to be harmful. This can be done using the noncentral t distribution. Additional details can be found in chapter 7 of Belsley (1991).

## Recommendations for Usage

Since standard statistical packages such as SPSS and SAS can provide both Belsley's diagnostics and VIFs, all users of multiple regression should be trained in their

use. While it may seem obvious that all data sets which will be subject to multiple regression should be diagnosed for collinearity, this author's experience is that very few actually are.

This author's first choice for diagnosing collinearity is Belsley's method. While this method is clearly superior to VIFs, it does generate considerably more output. One possible compromise is to use VIFs for the initial diagnosis of collinearity, and, if collinearity exists, do further analysis with condition indexes.

Belsley's recommendations for determining the degree of collinearity using his methods (i.e., condition indexes over 10 indicate moderate collinearity, and those over 30 severe collinearity) appear to be at least a good starting point. For VIFs, however, the generally recommended levels of 5 or 10 appear to be too low. Based on the results shown in Table 7, I recommend 100 for moderate collinearity and 1,000 for severe collinearity.

Moreover, if the largest condition index or VIF is substantially larger then the second largest, or if there are large gaps in the sequence of condition indexes or VIFs, that is additional evidence of collinearity.

If there is evidence of collinearity, then further work is necessary to determine what the cause of the

collinearity is, and what steps should be taken to remedy it. These steps will depend on the nature of the problem being considered. Possible remedies include dropping one or more variables, collecting more data, and using ridge regression.

Dropping variables may be a good choice if the correlation between two or more variables is very high, and if there is no strong substantive reason for keeping both of them. This may often be the case where there are many IVs, or if the IVs were chosen in an exploratory fashion, rather than for strong substantive reasons.

Collecting more data will not always solve the collinearity, and may not be feasible for practical reasons. It is more likely to solve the problem if there are many IVs compared to the number of cases, or if the data vary little on one or more variables.

A full discussion of ridge regression is beyond the scope of this dissertation. It provides biased estimates of the parameters, but may be less sensitive to problems such as collinearity.

# CHAPTER V

## SUMMARY

In multiple regression a single dependent variable is estimated as a linear combination of a number of independent variables (IVs). Some of the IVs may be related to each other. When this relation is strong, collinearity may exist to a problematic degree. Exact collinearity exists when one variable is a linear combination of any set of other variables. While this rarely occurs, near collinearity does occur, and can cause problems including difficulty in interpreting the partial coefficients, instability of the coefficients and computational inaccuracies.

A number of methods have been proposed for evaluating the degree of collinearity. This dissertation used computer simulation to evaluate two widely recommended methods for experiments where the number of IVs, the presence of an interaction between two of the IVs and the degree of collinearity were varied. The two methods were variance inflation factors (VIFs) as recommended by Montgomery and Peck (1982), Snee (1973) and Snee and Marquadt (1984), among others, and condition indexes, as developed and recommended by Belsley (1991, Belsley, Kuh, & Welsch, 1980). In addition, Belsley's results were

extended to new situations, and distributions for Belsley's diagnostics were computed.

VIFs are based on the inverse of the correlation matrix of the IVs. The VIFs are the diagonal elements of the inverted correlation matrix, after that matrix has been centered and scaled to have unit length. VIFs indicate the degree to which the variances of the parameter estimates are inflated due to collinearity.

Condition indexes are the basis of a collinearity diagnostic procedure introduced by Belsley, Kuh, and Welsch (1980) and refined by Belsley (1991). The condition indexes are the ratios of each of the singular values of the data matrix to the smallest singular value[12] of the matrix. The number of collinearities is determined by the number of high condition indexes. Belsley (1991) suggests that a condition index over 10 indicates moderate collinearity, and one over 30 indicates strong collinearity (note that the condition indexes are numbers, which can be high or low, but that collinearity is a condition which can be strong or weak). The degree to which particular parameters in the regression equation are affected by

---

[12] The singular values of X are the diagonal elements of D where X = UDV' is the singular value decomposition.

collinearity is determined by the variance decomposition proportions, which are given by

[5.1] $$\pi_{jk} = \phi_{kj}/\phi_k, \; k,j = 1, 2, \ldots p$$

where

[5.2] $$\phi_{kj} = v_{kj}^2 / \mu_j^2 \text{ and } \phi_k = \Sigma \phi_{kj}, \; k = 1,2,\ldots p$$

and $v_{kj}$ are the elements of V in the singular value decomposition of X.

One component of this dissertation was to extend our knowledge of collinearity diagnostics in several directions. First, the diagnostic procedures developed by Belsley (1991; Belsley, Kuh & Welsch, 1980) were used on data where the IVs are normally distributed, rather than uniformly distributed, as in Belsley's examples. Second, Belsley's procedures were applied to models which include an interaction term between two of the IVs, a situation which Belsley (1991) indicates may be particularly problematic. Third, Belsley's methods were compared to VIF measures. Fourth, distributions for condition indexes and for VIFs were established for all conditions considered.

An additional feature of this dissertation was to inform researchers of some of the issues regarding collinearity and its ill effects, and of the inadequacies of some commonly used methods for diagnosing it.

The conceptual hypotheses were:

1. Belsley's method would distinguish among weak, moderate, and strong collinearity and would correctly identify which variables were involved in the collinear relation when used with normally distributed variables.

2. For both Belsley's methods and VIFs, errors in either determining the presence of collinearities or the variables involved in them would be more common in cases involving an interaction term (cf. the section entitled "Interactions" in Chapter I).

3. Belsley's method would prove equal or superior to VIFs in all conditions.

## Method

This dissertation evaluated two collinearity diagnostics (VIFs and Belsley's method) when three parameters were varied. These parameters were number of IVs in the regression (3, 5, or 7), presence or absence of an interaction term, and degree of collinearity. Each model included an intercept term. These parameters were chosen because: 1) a literature review of recent articles using regression published in Developmental Psychology, Journal of Consulting and Clinical Psychology, Journal of Applied Psychology, or Journal of Personality and Social Psychology indicated that this number of IVs was common in psychological research; 2) the presence of an interaction

term extends Belsley's results and many psychological models include an interaction; 3) varying the degree of collinearity allows us to assess the performance of the diagnostics; and 4) most of the psychological models in the literature review included an intercept term.

This 3 (3, 5, or 7 IVs) x 2 (presence vs. absence of an interaction term) x 3 (weak, moderate, or strong collinearity) factorial ANOVA design generated 18 models. Degree of collinearity was varied using the method used by Belsley (1991). Each model: 1) had IVs generated to be normally distributed with a mean of 0 and standard deviation 1; 2) included an intercept term and a single collinear term; 3) had a sample size of 100. Each model was replicated 1,000 times using SAS call routines for uniform and normal distributions (SAS Institute, 1991).

Each replication of each model was evaluated for collinearity using VIFs and Belsley's procedure. This evaluation was performed using SAS. Based on the most common recommendations, a VIF of 5 was taken to indicate moderate collinearity, and a VIF of 10 to indicate strong collinearity. Belsley (1991) recommends using cutoffs of a condition index of 10 for moderate collinearity and 30 for strong collinearity. In addition, Belsley (1991) requires that, for collinearity to be diagnosed, two or more IVs

must have variance-decomposition proportions associated over 0.5 associated with the same high condition index.

## Results

This section is organized as follows: First, the distribution of the largest VIF and the largest condition index are compared for the various models. Second, the ability of Belsley's method to correctly identify the degree of collinearity and variable involvement is determined. Third, the effect of an interaction among the IVs is evaluated. Fourth the performance of condition indexes in determining the degree of collinearity is compared to that of VIFs. Fifth, the degree of dispersion of the two statistics is compared. Finally, the results are summarized.

### Sampling Distribution of the Statistics

Each of the 18 models was replicated 1,000 times, yielding a distribution for each of the two statistics for each of the 18 models. The distributions of the largest condition index and the largest VIF are shown graphically in Appendix E, and statistics regarding them are in Tables 4 and 5 in Chapter 3. The largest values for each statistic were chosen because they are most relevant to collinearity diagnosis; if the largest VIF or condition index did not indicate collinearity, no other value could.

The normality of these distributions was tested using the Shapiro-Wilk's (SW) test (D'Agostino, 1982; Shapiro & Wilk, 1965; Royston, 1982). Based on this test, all of the distributions of both statistics are significantly different from normal; this is due, in part, to the large sample size. Further evidence of the degree of nonnormality is the degree of skewness and kurtosis. The normal distribution has a skew of 0 and a kurtosis of 0. All of the distributions of both statistics had positive skewness and positive kurtosis. For the condition indexes, the mean skewness was 0.52, and the mean kurtosis was 0.61 (see Table 4 in Chapter 3). For the VIFs, the mean skewness was 0.83 and the mean kurtosis was 1.58, indicating that, on average, the VIFs were more skewed and more leptokurtotic than the condition indexes.

The skewness was closer to 0 for the condition index than for the VIF for each of the 18 models, and the kurtosis was closer to 0 for the condition index in 17 of the 18 models. In addition, the SW test is closer to 1.00 (which indicates normality) for the condition index than for the VIF for each of the 18 models.

However, none of the distributions are "grossly" nonnormal. They are all unimodal, they all have modes near

the center of the distribution, and none have any obvious outliers by visual inspection.

## Performance of Belsley's Method

### Ability to Determine Existence of Collinearities

Belsley's method was able to determine the existence and the degree of collinearity. The largest condition index was larger for models with strong collinearity than for those with moderate collinearity, and larger for those with moderate collinearity than with weak collinearity. Both of these differences were highly statistically significant, and were also in line with Belsley's (1991) recommendations of 10 and 30 as indicators of moderate and strong collinearity, respectively.

### Ability to Determine Variable Involvement

In addition to being able to determine if collinearity exists and diagnose its degree, Belsley (1991) states that the condition indexes are able to determine which variables are involved in the collinear relations. This ability was confirmed by this dissertation.

In each of the models considered, IVs 1 and 2 (and only these) were involved in collinear relations. Therefore, by Belsley's method, only these should have large variance proportion associated with large condition indexes. This was the case for all 18 models considered in

this dissertation. For each model, the largest variance decomposition proportions were associated with IVs 1 and 2, and each of these had means over 0.5 (they range from 0.89 to 1.00), and, in each case, no other variance decomposition proportions were over 0.5.

Effect of an Interaction Term on the Performance of the Condition Index

Belsley (1991) noted that interactions can cause problems for collinearity diagnostics, including his method. This dissertation examined models with a single interaction term, and found that they did not adversely affect the performance of Belsley's method. The results for models with an interaction term were not meaningfully different from those without an interaction term.

Comparison of Condition Indexes and VIFs

One of the main purposes of this dissertation was to compare the performance of VIFs and Belsley's method. In general, both VIFs and Belsley's method were able to diagnose collinearity and determine its degree. The mean largest VIF (like the mean largest condition index) was significantly larger for moderate collinearity than for weak collinearity, and significantly larger for strong collinearity than for moderate.

The condition indexes are, however, less dispersed than the VIFs. The coefficients of variation (CV) for the largest condition indexes were significantly smaller than those for the largest VIF. For the condition indexes, the CVs ranged from 0.07 to 0.20 with a median of 0.08; for the VIFs, they ranged from 0.15 to 0.41 with a median of 0.16. The test for this difference assumes that both variables are normally distributed. If we are not willing to make this assumption, then a useful comparison of the dispersions is that suggested by Lewontin (1966), of comparing the logarithms of the variances; by this test, as well, the condition indexes are significantly less disperse than the VIFs.

## Summary

Both VIFs are able to diagnose the degree of collinearity in multiple regression. Condition indexes can, in addition, determine the number of collinear relations and the variables which are involved in the collinearity. In addition, condition indexes are less dispersed than VIFs, and are more normally distributed.

## Discussion

This dissertation examined the performance of two collinearity diagnostics (VIFs and Belsley's method) when three parameters were varied (degree of collinearity,

number of IVs, and presence or absence of an interaction term). Collinearity diagnostics should have several properties: First, they should be able to diagnose collinearity and determine its degree. Second, they should be able to determine the number of collinearities. Third, they should be able to determine which variables are involved in the collinearities. Fourth, they should be as precise and uniform as possible. Finally, these properties should not vary across different conditions other than the degree of collinearity. Each of these is discussed in turn, except for determining the number of collinearities, which was not addressed in this dissertation.

Ability to Detect Collinearity and Determine its Degree

It is important to be able to detect collinearity and determine its degree because this indicates whether the collinearity may be affecting the parameter estimates of the regression. Condition indexes were able to detect and determine the degree of collinearity. For each set of models, the mean of the largest condition index was significantly larger for moderate collinearity than for weak collinearity, and significantly higher for strong collinearity than for moderate. For 17 of the 18 models considered in this dissertation, Belsley's recommendations of using 10 as a cutoff for moderate collinearity and 30

for strong were 100% accurate. The only model where they were not 100% accurate was the model with no interaction, weak collinearity, and 3 IVs. In 7.5% of the replications of this model, the guidelines indicated moderate collinearity. In the remaining 92.5%, they indicated the correct degree of collinearity. One anomaly is that the condition index made incorrect decisions only for the simplest model that was evaluated (i.e., no interaction, 3 IVs) for weak collinearity. The cause of this anomaly remains to be determined.

While there were clear differences in the levels of VIFs for the different degrees of collinearity, these did not coincide with the recommendations of 5 for moderate collinearity and 10 for strong collinearity found in the literature (Marquadt, 1980; Montgomery & Peck, 1982; Snee, 1973) the results of this dissertation indicate that the appropriate values are substantially higher; a tentative recommendation is that 100 indicates moderate collinearity, and 1,000 indicates strong collinearity. This is important because using the guidelines from the literature will lead us to diagnose collinearity where none exists, or to overestimate the degree of collinearity where it does exist.

## Ability to Determine Variable Involvement

It is important to be able to determine which variables are involved in the collinear relations, because this helps focus any remedial measures that are necessary. For example, if it is the case that one or more of the variables involved in the collinearity can be dropped without serious substantive difficulties, then this may be an ideal solution. This is impossible without knowing which variables are involved in the collinearity.

Belsley's method was able to determine the variables involved in all cases for all models (see Tables 9 & 10 in Chapter 3). VIFs are unable to do this, since they provide no information about variable involvement. This argues strongly for using Belsley's methods to diagnose collinearity.

## Precision of the Diagnostics

The precision of a statistic is important because a more precise statistic gives more information than a less precise one. With regard to collinearity diagnostics, this implies that a more precise statistic allows a better estimate of the degree of collinearity.

The results of this dissertation show that, for the models considered, condition indexes are more precise estimates of the degree of collinearity than VIFs are. This is further support for using Belsley's methods. This

implies that, as empirical evidence of the relationship between the level of the diagnostics and the degree of problem caused to the regression equation accumulates, condition indexes will allow better estimates of the degree to which any particular regression equation is harmed by collinearity.

## Stability Across Models

Collinearity diagnostics should not vary across conditions other than degree of collinearity. If they did vary, then it would be impossible to establish general guidelines for their use. Belsley's recommendations of 10 for moderate collinearity and 30 for strong collinearity worked well across all the models considered here. For weak collinearity, the mean of the largest condition index ranged from 6.84 - 8.04, for moderate collinearity, from 21.19 - 24.91, and for strong from 66.91 - 79.14. This means that guidelines for the use of condition indexes do not depend on the model.

VIFs varied more across models than Condition Indexes did. For weak collinearity the mean of the largest VIF ranged from 7.01 - 14.64, for moderate collinearity it ranged from 107.53 - 137.87, and for strong collinearity from 1064.66 - 1380.50. The greater variation across models argues against using VIFs. The degree of variation

is not very high even for VIFs i.e., the ranges 7.01 - 14.64, 107.53 - 137.87, and 1064.66 - 1380.50 are not large. It should be kept in mind, however, that this dissertation examined three quite distinct levels of collinearity, so the effect of changes on degree of collinearity was clear. In actual research, collinearity varies along a continuum, and the less the diagnostics vary across conditions, the more precise recommendations for using them can be.

## Limitations of the Present Research

There are three principal limitations to this dissertation. First, it deals only with models with one collinear relation. Second, it deals only with models with 3, 5, or 7 IVs. Third, it does not deal with models with multiple interactions. Each of these is discussed in turn.

## Additional Suggestions for Future Research

In addition to addressing the limitations discussed above, research is needed on the relationship between the condition index and the effect of collinearity on the regression estimates. Not all collinearities affect these estimates. A starting point for this research is Chapter 7 of Belsley (1991), Belsley (1982), Belsley and Oldford (1986), Gunst (1983), and Simon & Lesage (1988).

Belsley (1991) notes that, if the error variance is small enough, collinearity is not harmful. The higher the degree of collinearity, the less error variance can be tolerated, so that "small" is necessarily relative. He recommends the use of the ratio of the parameter to its error variance as a measure of "signal to noise" $\tau = B_i/\sigma_{bi}$ (note that this recommendation involves the use of parameters, rather than their estimators). The test is then whether this statistic is sufficiently different from zero to be harmful. This can be done using the noncentral t distribution. Additional details can be found in chapter 7 of Belsley (1991).

## Recommendations for Usage

Since standard statistical packages such as SPSS and SAS can provide both Belsley's diagnostics and VIFs, all users of multiple regression should be trained in their use. While it may seem obvious that all data sets which will be subject to multiple regression should be diagnosed for collinearity, this author's experience is that very few actually are.

This author's first choice for diagnosing collinearity is Belsley's method. While this method is clearly superior to VIFs, it does generate considerably more output. One possible compromise is to use VIFs for the initial diagnosis of collinearity, and, if collinearity exists, do further analysis with condition indexes.

Belsley's recommendations for determining the degree of collinearity using his methods (i.e., condition indexes over 10 indicate moderate collinearity, and those over 30 severe collinearity) appear to be at least a good starting point. For VIFs, however, the generally recommended levels of 5 or 10 appear to be too low. Based on the results shown in Table 7, I recommend 100 for moderate collinearity and 1,000 for severe collinearity.

Moreover, if the largest condition index or VIF is substantially larger then the second largest, or if there

are large gaps in the sequence of condition indexes or VIFs, that is additional evidence of collinearity.

If there is evidence of collinearity, then further work is necessary to determine what the cause of the collinearity is, and what steps should be taken to remedy it. These steps will depend on the nature of the problem being considered. Possible remedies include dropping one or more variables, collecting more data, and using ridge regression.

## REFERENCES


Belsley, D. A. (1984a). Demeaning conditioning diagnostics through centering (with comments and rejoinder). American Statistician, 38, 73-93.

Belsley, D. A. (1984b). Reply. American Statistician, 38, 90-93.

Belsley, D. A. (1986). Centering, the constant, first-differencing and assessing collinearity. In D. A. Belsley & E. Kuh (Eds.) Model Reliability. Cambridge, MA: MIT Press.

Belsley, D. A. (1987). Comment: Well-conditioned collinearity indices. Statistical Science, 2, 86-91.

Belsley, D. A. (1991). Conditioning diagnostics: Collinearity and weak data in regression. New York: Wiley.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley.

D'Agostino, R. B. (1988). Tests for departure from normality. In S. Kotz & N. L. Johnson (Eds.) Encyclopedia of statistical sciences, volume 4, p. 315-324, New York: Wiley.

Eckart, G. & Young, G. (1936). The approximation of one matrix by another of lesser rank. Psychometrika, 1, 211-218.

Fieller, E. (1932). A numerical test of the accuracy of A. T. McKay's approximation. Journal of the Royal Statistical Society, 95, 699-702.

Gibbons, J. D. (1985). Nonparametric statistical inference, (2nd edition). New York: Marcel Dekker, Inc.

Golub, G. H. & Van Loan, C. F. (1996). Matrix computations (3rd edition). Baltimore: Johns Hopkins University Press.

Gunst, R. F. (1983). Regression analysis with multicollinear predictor variables. Communications in Statistics, A12, 2217-2260.

Hanson, R. J. & Lawson, C. L. (1969). Extensions and applications of the Householder algorithm for solving linear least squares problems. Mathematics of Computation, 23, 787-812.

Iglewicz, B. & Myers, R. H. (1970). On the percentage points of the sample coefficient of variation. Technometrics, 12, 166-169.

Kuh, E. & Belsley, D. A. (Eds.) (1986). Model reliability. Cambridge, MA: MIT Press.

Lesage, J. P. & Simon, S. D. (1985). Numerical accuracy of statistical algorithms for microcomputers. Computational Statistics and Data Analysis, 3, 47-57.

Lewontin, R. C. (1966). On the measurement of variability. Systematic Zoology, 15, 141-142

Marquadt, D. W. (1980). You should standardize the predictor variables in your regression models. Journal of the American Statistical Association, 75, 87-91.

Marquadt, D. W. & Snee, R. D. (1975). Ridge regression in practice. The American Statistician, 29, 3-20.

McKay, A. T. (1932). Distribution of the coefficient of variation and the extended 't' distribution. Journal of the Royal Statistical Society, 95, 695-698.

Montgomery, D. C. & Peck, E. A. (1982). Introduction to linear regression analysis. New York: Wiley

Pearson, E. S. (1932). Comparison of A. T. McKay's approximation with experimental sampling results. Journal of the Royal Statistical Society, 95, 703.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Rannery, B. P. (1992). Numerical recipes in C (2nd edition). New York: Cambridge University Press.

Royston, J. P. (1988). Shapiro-Wilk W Statistics. In S. Kotz & N. L. Johnson (Eds.) Encyclopedia of statistical sceinces, volume 4, p. 315-324, New York: Wiley.

SAS Institute Inc. (1991). SAS system for regression (2nd edition). Cary, NC: SAS Institute.

SAS Institute, Inc. (1990). SAS language: Reference, Version 6, 1st Edition, Cary, NC: SAS Institute.

SAS Institute Inc. (1989). SAS/STAT users' guide, version 6 (4th edition). Cary, NC: SAS Institute.

Schott, J. R. (1997). Matrix analysis for statistics. New York: Wiley.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality. Biometrika, 52, 591-611.

Simon, S. D. & Lesage, J. P. (1988a). The impact of collinearity involving the intercept term on the numerical accuracy of regression. Computer Science in Economics and Management, 1, 137-152.

Simon, S. D. & Lesage, J. P. (1988b). Benchmarking numerical accuracy of statistical algorithms, Computational Statistics and Data Analysis, 7, 197-209.

Smith, G. & Campbell, F. (1975). A critique of some ridge regression methods. Journal of the American Statistical Association, 75, 74-103.

Snee, R. D. (1973). Some aspects of nonorthogonal data analysis. Journal of Quality Technology, 5, 67-79.

Snee, R. D. & Marquadt, D. W. (1984). Collinearity diagnostics depend on the domain of prediction, the model, and the data. The American Statistician, 38, 83-87.

Sprent, P. (1998). Data driven statistical methods, London: Chapman and Hall.

Stewart, G. W. (1987). Collinearity and least squares regression. Statistical Science, 2, 68-100.

Thisted, R. A. (1975). A critique of some ridge regression methods: Comment. Journal of the American Statistical Association, 75, 80-86.

Vangel, M. G. (1996). Confidence intervals for a normal coefficient of variation. American Statistician, 15, 21 - 26.

Wampler, R. H. (1980). Test procedures and problems for least-squares algorithms. Journal of Econometrics, 12, 3-22.

Weisberg, S. (1980). Applied linear regression. New York: Wiley.

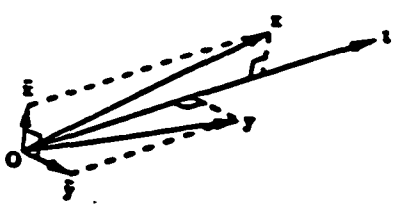
## Appendix A

## Collinearity and Correlation

Geometrically, the correlation is the cosine of the angle between the mean-centered variates, whereas collinearity is a measure of the angle between the variates themselves. Thus, the concepts are different. There are two ways of demonstrating that collinearity and correlation can differ by large amounts. First, it can be shown that two variables can be arbitrarily collinear while having a correlation of 0 (see Figure 1). Second, it can be shown that there can be an exactly collinear relation among a set of variables with no pair having a particularly high correlation.

Figure 1

Highly collinear but uncorrelated variates




From Belsley, D. A. (1991). Conditioning Diagnostics: Collinearity and Weak Data in Regression, by permission of John Wiley & Sons.

Near Collinearity With No Correlation

Let u and v be orthonormal vectors in A, where A is the orthogonal complement of (ι) in $R^n$. Then u'v = 0 and u'u = v'v = 1. Next, define $x(\alpha) = (\iota) + \alpha u$, and $y(\alpha) = (\iota) + \alpha v$. Then, as $\alpha$ approaches 0, $x(\alpha)$ approaches $y(\alpha)$, the angle $\phi$ between them approaches 0, and they become arbitrarily collinear since, $\cos(\phi)$ approaches 1. At the same time, however, the correlation between $x(\alpha)$ and $y(\alpha)$ is simply u'v, which is 0, for all nonzero $\alpha$ (Belsley, 1991; Schott, 1997).

Exact Collinearity With Low Correlation

In addition, it is possible for there to be an exactly collinear relation among p variables with no pairwise correlation exceeding 1/(p-1). If all the correlations are equal to $\rho$, then the correlation matrix is singular if $\rho = 1$ or if $\rho = -1/(p - 1)$.

## Appendix B

## Details of Belsley's Diagnostics

What follows is a brief summary of Belsley (1991) Chapter 3. It is intended only to trace the development of his diagnostic method, and some concepts from matrix algebra are left undefined. First, some results are established for exact collinearities; these are then generalized to near collinearities. Finally, methods of establishing the number of collinearities and variate involvement are established.

Exact Collinearity

Suppose an n x p matrix X has p-r exactly collinear relations among its columns. Then, rank (X) = r, and r < p. In the SVD X = UDV', both U and V are of full rank, so rank(X) must equal rank(D). However, D must also be of order p x p. Therefore, D can be partitioned as

$$\text{(B1)} \qquad \begin{matrix} D_{11} \\ 0 \end{matrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where $D_{11}$ is r x r and full-rank. Then

$$\text{(B2)} \qquad X = U \begin{matrix} D_{11} \\ 0 \end{matrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} V'$$

post-multiply by V and further partition to get:

$$\text{(B3)} \qquad X[V_1\ V_2] = [U_1\ U_2] \begin{matrix} D_{11} \\ 0 \end{matrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where $V_1$ is p x r, $V_2$ is p x (p-r), $U_1$ is n x r, and $U_2$ is n x (p-r). Multiplying out gives

(B4) $$XV_1 = U_1D_{11}$$

(B5) $$XV_2 = 0$$

So, when X has p - r exact collinearities, D has p - r 0s on the diagonal.

Near Collinearity

If 0 values on the diagonal of D indicate exact collinearities, it is intuitive to suppose that small values on the diagonal indicate near collinearities. There are two points which need further elaboration: One is determining if this intuition is correct, the other is determining what "small" means.

Taking the second point first, one sensible (and, it turns out, highly useful) scaling method is to scale each of the diagonal values by dividing it by the largest of them. The condition number $\kappa(X)$ is defined as the largest singular value divided by the smallest. That is

(B6) $$\kappa(X) = \mu_{max}/\mu_{min}$$

where $\mu_i$ refers to the $i^{th}$ singular value of X.

The condition number has several useful properties:

1) If X is full-rank, then X + $\delta X$ is also, provided that $||\delta X|| / ||X|| < \kappa^{-1}(X)$ where $||X||$ is the spectral norm of X (Hanson & Lawson, 1969).

2) If X is full-rank, then the smallest E which makes X + E exactly collinear has spectral norm $\mu_{min}(X)$. Thus, $\mu_{min}(X)$ provides an absolute measure of the distance of X from exact collinearity, and $\kappa^{-1}(X)$ provides a relative measure of this distance (Eckart & Young, 1936; Mirsky, 1960; Stewart, 1987).

Number of Collinearities

Since for each exact collinearity in X there is one zero singular value, and since an overall near-collinearity is indicated by a small singular value, we may hypothesize that multiple near collinearities are indicated by multiple small singular values. This turns out to be correct (Kendall, 1975; Silvey, 1969) but the issue of scaling needs to be resolved.

Belsley (1991) maintains the scaling used above. He defines the condition indexes of X as

(B7) $$\eta_k = \mu_{max} / \mu_k, \text{ for } k = 1, 2, \ldots, p$$

and it is then the case that "there are as many near dependencies among the columns of a data matrix as there are high condition indexes" (Belsley, 1991, p. 56). Exact

values of "high" which indicate problems are established empirically.

Variate Involvement

Given the usual assumptions and the OLS estimator $b = (X'X)^{-1}X'y$, the variance of b is $\sigma^2(X'X)^{-1}$, where $\sigma^2$ is the common variance of $\varepsilon$. Then, since by the SVD $X = UDV'$

$$V(b) = \sigma^2(X'X)^{-1} = \sigma^2V'D^{-2}V' \tag{B9}$$

and

$$\text{var}(b_k) = \sigma^2 \Sigma_j (v_{kj}^2/\mu_j^2) \tag{B10}$$

This decomposes the variance of $b_k$ into a sum of components, each associated with one singular value. We can then define the variance-decomposition proportions as follows. First, let

$$\phi_{kj} = v^2_{kj}/\mu^2_j \tag{B11}$$

and

$$\phi_k = \Sigma_j\phi_{kj} \quad \text{for } k = 1....p \tag{B11}$$

then the variance-decomposition proportions are:

$$\pi_{jk} = \phi_{kj}/\phi_k, \quad k = 1.....p \tag{B12}$$

## Appendix C

### Benchmarking of SAS Code Against Belsley's Results

Belsley (1991) provides a number of experiments illustrating his diagnostic procedure. Some of these use data that cannot easily be replicated. Others use random data. Below, one of the latter experiments is described, and then replicated. This is done as a check on the SAS code written by the author, and on the procedure implemented by SAS.

One reason the latter check is particularly important is that, while Belsley (1991) prefers the singular value decomposition (SVD) method of obtaining his diagnostics because it is more numerically stable, SAS uses the eigenvector method because it less computationally intensive (SAS Institute, 1989, 1991; M. Stokes, personal communication, November 18, 1997). The stability of the SVD method is due to the fact that the SVD does not require computation of a cross products matrix (Lesage & Simon, 1985[1]).

---

[1]Lesage and Simon (1985) do not actually recommend the SVD method, although they do find it superior to those based on cross-product matrices. Their first choice is the modified Gram-Schmidt computation of the Householder algorithm.

Belsley's (1991) U series involves the following basic data matrix:

(C1) $U = [D_1, D_2, D_3]$

where each D is ~U(0,1).

In series U1, which is replicated below, two collinear terms ($D_4$ and $D_5$) are added to the above:

(C2) $D_4 = .3\ D_2 - .6C_3 + e_I$

where

(C3) $e_i \sim N(0,\ 10^{-i}s^2_u I)$

and

(C4) $s^2_u = var(.3D_2 - .6D_3)$

(C5) $D_5 = .7D_1 + .4D_2 + e_j$

where

(C6) $e_j \sim N(0,\ 10^{-j}s^2_v I)$

and

(C7) $s^2_v = var(.3D_2 - .6D_3)$

where i and j = 1, 2, 3, 5.

The SAS code for replicating these results is reproduced immediately below. $Q_1$ and $q_2$ represent i and j in the formulas above.

```
/* A program to replicate data in Belsley (1991) Chapter 4, experiment U1 */
options linesize=79;
*This section assigns macro variables;
%let q1=4;  %let q2=3;
* This section creates 4 random variables ~ U(0,1)
 it also creates two dummy variables, for use below ;
Title 'A program to  replicate Belsley (1991) results, U1 series';
filename stuff 'flom.dat';
data one;
file stuff;
seed1 = int(ranuni(0)*(10**7));
seed2 = int(ranuni(0)*(10**7));
seed3 = int(ranuni(0)*(10**7));
seed4 = int(ranuni(0)*(10**7));
 do I = 1 to 24;
   call ranuni(seed1,X1);
   call ranuni(seed2,X2);
   call ranuni(seed3,X3);
   call ranuni(seed4,Y);
   dum1 = .3*X2 - .6*X3;
   dum2 = .7*X1 + .4*X2;
   put X1 X2 X3 Y Dum1 Dum2;
   output;
```

```
 end;
/*This section computes the variance of Dum1 and Dum2, for use below */
proc means data=one noprint;
 var Dum1 Dum2;
 output out=two var(Dum1 Dum2) = var_Dum1 var_Dum2;
run;
* This section combines the data from the first two sections;
title2 "U(&q1,&q2)";
data three;
 set one;
 if _n_=1 then set two;
 seed5 = int(ranuni(0)*(10**7));
 seed6 = int(ranuni(0)*(10**7));
 call rannor(seed5,rn1);
 call rannor(seed6,rn2);
 vd1 = sqrt(var_dum1/(10 ** &q1)) * rn1;
 vd2 = sqrt(var_dum2/(10**&q2)) * rn2;
 do I = 1 to 24;
   X4 = Dum1 + vd1;
   X5 = dum2 + vd2;
 end;
run;
* This section produces collinearity diagnostics;
```

```
proc reg data=three;
model y = X1 X2 X3 X4 X5/collin vif noint;
run;
proc print;
var seed1-seed5;
run;
```

Tables 13 and 14 present Belsley's(1991) U1 series, and a replication of it using the above program. In each cell, Belsley's result is reported first, and the replication second. To make the tables more easily comparable, the condition indices are rounded to the nearest integer, and the variance proportions to the 3rd decimal place (as in Belsley, 1991). In addition, the scaled condition indices are ordered from smallest to largest (as in SAS), and, in some instances, the rows for the smallest condition indices are not shown (as in Belsley). These rows are largely irrelevant to collinearity analysis. Cases where this has been done are marked with an asterisk.

It is apparent that Belsley's (1991) results have been adequately replicated. All cells are close to each other, and, in every case, the same decisions about collinearity would be reached using either set of data. The correlation between the two sets of results is over .99.

Table 13

Replication of Belsley's (1991) U1 data, i = 1, j = 1-4

| η | Proportions of | | | | |
|---|---|---|---|---|---|
| | | | U1{1,0} | | |
| | $Var(b_1)$ | $Var(b_2)$ | $Var(b_3)$ | $Var(b_4)$ | $Var(b_5)$ |
| 1, 1 | .011, .011 | .004, .003 | .001, .002 | .002, .002 | .007, .007 |
| 2, 2 | .006, .001 | .021, .016 | .003, .001 | .031, .046 | .027, .012 |
| 4, 3 | .367, .309 | .088, .063 | .007, .010 | .000, .001 | .011, .019 |
| 7, 7 | .616, .663 | .108, .061 | .003, .005 | .012, .011 | .942, .961 |
| 17, 16 | .000, .023 | .779, .856 | .986, .982 | .955, .940 | .013, .000 |
| | | | U{2,0}* | | |
| 4, 4 | .053, .027 | .045, .048 | .007, .014 | .000, .015 | .002, .003 |
| 17, 16 | .107, .003 | .110, .206 | .862, .982 | .859, .939 | .094, .003 |
| 22, 31 | .836, .972 | .831, .715 | .127, .000 | .106, .012 | .901, .992 |
| | | | U{1, 2}* | | |
| 4 | .053, .003 | .045, 010 | .007, .004 | .000, .000 | .002, .000 |
| 17, 20 | .002, .001 | .050, .094 | .982, .978 | .961, .931 | .001, .018 |
| 68, 73 | .993, .995 | .939, .893 | .007, .014 | .004, .030 | .998, .998 |
| | | | U{1,3}* | | |
| 4, 4 | .001, .001 | .001, .002 | .007, .003 | .000, .001 | .000, .000 |
| 17, 14 | .000, .000 | .006, .011 | .982, .958 | .961, .910 | .001, .000 |
| 215, 154 | .999, .999 | .993, .986 | .003, .027 | .002, .052 | 1.00, 1.00 |
| | | | U{1,4}* | | |
| 4, 4 | .000, .000 | .000, .000 | .007, .001 | .000, .003 | .000, .000 |
| 17, 18 | .000, .000 | .001, .002 | .986, .977 | .963, .950 | .000, .000 |
| 679, 564 | 1.00, 1.00 | .999, .998 | .003, .006 | .001, .004 | 1.00, 1.00 |

Table 14

Replication of Belsley's (1991) U1 data, i = 1-4, j = 3

| η | Proportions of | | | | |
|---|---|---|---|---|---|
| | Var($b_1$) | Var($b_2$) | Var($b_3$) | Var($b_4$) | Var($b_5$) |
| | | | U1{0,3} | | |
| 1, 1 | .000, .000 | .000, .000 | .008, .005 | .010, .008 | .000, .000 |
| 2, 3 | .000, .000 | .000, .000 | .009, .022 | .271, .164 | .000, .000 |
| 4, 4 | .001, .000 | .001, .001 | .065, .000 | .001, .180 | .000, .000 |
| 7, 8 | .000, .000 | .002, .000 | .913, .958 | .717, .588 | .000, .000 |
| 211, 345 | .999, 1.00 | .997, .998 | .005, .016 | .001, .061 | 1.00, 1.00 |
| | | | U1{1,3}* | | |
| 2, 2 | .000, .000 | .000, .000 | .003, .000 | .034, .016 | .000, .000 |
| 4, 5 | .001, .000 | .001, .001 | .007, .011 | .000, .006 | .000, .000 |
| 17, 24 | .000, .000 | .006, .010 | .985, .962 | .963, .956 | .000, .000 |
| 215, 288 | .999, .999 | .993, .989 | .003, .025 | .002, .021 | 1.00, 1.00 |
| | | | U1{2,3}* | | |
| 2, 2 | .000, .000 | .000, 000 | .000, .000 | .003, .004 | .000, .000 |
| 4, 5 | .001, .000 | .001, .001 | .001, .000 | .000, .000 | .000, .000 |
| 54, 60 | .001, .000 | .045, .063 | .994, .982 | .992, .980 | .001, .000 |
| 215, 324 | .998, .999 | .954, .936 | .005, .017 | .004, .015 | .999, .999 |
| | | | U1{3,3}* | | |
| 4, 3 | .001, .000 | .001, .000 | .000, .000 | .000, .000 | .000, .000 |
| 168, 174 | .069, .054 | .145, .137 | .914, .894 | .915, .895 | .067, .057 |
| 219, 254 | .930, .946 | .854, .862 | .086, .106 | .085, .104 | .933, .943 |
| | | | U1{4,3}* | | |
| 213, 241 | .993, .956 | .101, .194 | .003, .000 | .003, .000 | .992, .958 |
| 544, 461 | .007, .044 | .899, .808 | .997, 1.00 | .997, 1.00 | .008, .042 |

## Appendix D

## Programs to Generate Data and Compute Collinearity Diagnostics

All of the data were generated using SAS (SAS Institute, 1990), and all of the collinearity diagnostics were computed by SAS. There were 18 models. These models can be grouped into six sets: set 1 consists of models with no interaction and 3 IVs, set 2, no interaction and 5 IVs, set 3, no interaction and 7 IVs, set 4, interaction and 3 IVs, set 5, interaction and 5 IVs, and set 6, interaction and 7 IVs. In each set, models with weak, moderate and strong collinearity were examined.

Below are six SAS programs corresponding to the six model sets. The degree of collinearity can be adjusted by changing the value of the variable q1; weak collinearity is produced when q1 = 1, moderate collinearity when q1 = 2, and strong collinearity when q1 = 3.

Program 1: Models with No Interaction and 3 IVs

```
%macro Dissert;
options  nosource nonotes nocenter nosymbolgen nomprint nomlogic;
/* A program to generate data and test for collinearity */
/* This section assigns macro variables */
   filename seeds 'seed.dat';
   data start;
```

```
infile seeds; input seed ;
call symput('seed1',seed);
run;
%do i=1 %to 1000;
%let q1=1;
%let f = 8;
/* This section creates 5 random variables ~ U(0,1) it also creates a dummy
variable, for use below */
Title 'A program to generate data';
title2 "For formula &f";
data one;
seed=&seed1;
do I = 1 to 20;
  call rannor(seed,X0);
  call rannor(seed,X1);
  call rannor(seed,X2);
  call rannor(seed,X3);
  call rannor(seed,Y);
  XQ = X1 + X2;
  output;
end;
call symput('seed1',seed);
run;
```

```
/*This section computes the variance of X4, for use below */
   proc means data=one noprint;
    var XQ;
    output out=two var(XQ) = var_XQ;
   run;
/* This section combines the data from the first two sections */
   title3 "i = &q1";
   data three;
   seed=&seed1;
    set one;
    if _n_=1 then set two(keep = var_XQ);
    call rannor(seed,rn1);
    vd1 = sqrt(var_XQ/(10 ** &q1)) * rn1;
    do I = 1 to 20;
     XC = XQ  + vd1;
    end;
   call symput('seed1',seed);
   run;
/* This section produces collinearity diagnostics   outputs them to a file, and
computes statistics about them */
   filename newout "&f&q1";
   proc printto print =newout;
   run;
```

```
 proc reg data=three;
 model y = X1 X2 X3 XC/collin vif;
run;
proc printto;
run;
data Belsley;
infile newout lrecl=82 recfm=f;
 input word1 $ @;
 if word1 = 'Number' then
  do;
    input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35;
    keep n1--vp35;
   output;
   end;
 else if word1 = 'Variable' then
   do;
    input
#2 var1 $ df1 parest1 se1 t1 prob1 vif1
```

```
 / var2 S df2 parest2 se2 t2 prob2 vif2
 / var3 S df3 parest3 se3 t3 prob3 vif3
 / var4 S df4 parest4 se4 t4 prob4 vif4
 / var5 S df5 parest5 se5 t5 prob5 vif5;
  keep vif1-vif5;
  output;
  end;
 run;
 %end;
/* This section prints the means and sds to a file */
 filename tmp "means&f&q1";
 proc printto print=tmp new;
 run;
 proc means data=Belsley;
 run;
 options ls = 79;
/* This section creates 2 file for Excel */
 filename tmp2 "vif&f&q1";
 filename tmp3 "eta&f&q1";
 proc printto log = tmp2 new;
 run;
 data today;
 set belsley (keep = vif1-vif5);
```

```
 put (vif1-vif5) (6.2);
 where vif1 ne .;
run;
proc printto log = tmp3 new;
data today2;
 set belsley (keep = eta1-eta5);
 put (eta1-eta5) (6.2);
 where eta1 ne .;
run;
proc printto;
run;
* This section generates test of normality;
 options ls = 79;
 filename tmp4 "uni&f&q1";
 proc printto print = tmp4 new;
 run;
 proc univariate data = belsley;
  var vif2-vif5 eta5;
 run;
 data last;
 file seed;
 seed=&seed1;
 put seed 10.;
```

```
    run;
    x "delete dsk3:[flom]81.dat;*";
    %mend dissert;
%dissert;
```

Program 2: Models with No Interaction and 5 IVs

```
    %macro Dissert;
    options nosource nonotes nocenter nosymbolgen nomprint nomlogic;
    /* A program to generate data and test for collinearity */
    /* This section assigns macro variables */
      filename seeds 'seed.dat';
      data start;
      infile seeds;
      input seed ;
      call symput('seed1',seed);
      run;
      %do i=1 %to 1000;
      %let q1=1;
      %let f = 9;
    /* This section creates 5 random variables ~ U(0,1). It also creates a dummy
    variable, for use below */
    Title 'A program to generate data';
    title2 "For formula &f";
```

```
data one;
seed=&seed1;
do I = 1 to 100;
   call rannor(seed,X0);
   call rannor(seed,X1);
   call rannor(seed,X2);
   call rannor(seed,X3);
   call rannor(seed,X4);
   call rannor(seed,X5);
   call rannor(seed,Y);
   XQ = X1 + X2;
   output;
end;
call symput('seed1',seed);
run;
/*This section computes the variance of X4, for use below */
   proc means data=one noprint;
    var XQ;
    output out=two var(XQ) = var_XQ;
   run;
/* This section combines the data from the first two sections */
   title3 "i = &q1";
   data three;
```

```
seed=&seed1;

set one;

if _n_=1 then set two(keep = var_XQ);

call rannor(seed,rn1);

vd1 = sqrt(var_XQ/(10 ** &q1)) * rn1;

do I = 1 to 100;

  XC = XQ  + vd1;

end;

call symput('seed1',seed);

run;

/* This section produces collinearity diagnostics, outputs them to a file, and

computes statistics about them */

filename newout "&f&q1";

proc printto print = newout;

run;

proc reg data=three;

model y = X1 X2 X3 X4 X5 XC/collin vif;

run;

proc printto;

run;

data Belsley;

infile newout lrecl=500 recfm=f;
```

```
 input word1 $ @;
 if word1 = 'Number' then
  do;
   input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31 vp41 vp51
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32 vp42 vp52
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33 vp43 vp53
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34 vp44 vp54
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35 vp45 vp55
/ n6 eig6 eta6 vpint6 vp16 vp26 vp36 vp46 vp56
/ n7 eig7 eta7 vpint7 vp17 vp27 vp37 vp47 vp57;
   keep n1--vp57;
  output;
  end;
 else if word1 = 'Variable' then
  do;
   input
#2 var1 $ df1 parest1 se1 t1 prob1 vif1
/ var2 $ df2 parest2 se2 t2 prob2 vif2
/ var3 $ df3 parest3 se3 t3 prob3 vif3
/ var4 $ df4 parest4 se4 t4 prob4 vif4
/ var5 $ df5 parest5 se5 t5 prob5 vif5
/ var6 $ df6 parest6 se6 t6 prob6 vif6
```

```
/ var7 S df7 parest7 se7 t7 prob7 vif7:
 keep vif1-vif7;
 output;
 end;
run;
%end;
/* This section prints the means and sds to a file */
filename tmp "means&f&q1";
proc printto print=tmp new;
run;
proc means data=Belsley maxdec = 2 n mean std min max cv fw = 8;
run;
proc printto;
run;
options ls = 79;
/* This section creates 2 files for Excel */
filename tmp2 "vif&f&q1";
filename tmp3 "eta&f&q1";
proc printto log = tmp2 new;
run;
data today;
 set belsley (keep = vif1 - vif7);
 put (vif1 - vif7) (6.0);
```

```
 where vif1 ne .;
run;
proc printto log = tmp3 new;
run;
data today2;
 set belsley (keep = eta1 - eta7);
 put (eta1 - eta7) (6.0);
 where eta1 ne .;
run;
proc printto;
run;
/* This section generates tests of normality */
options ls = 79;
filename tmp4 "uni&f&q1";
 proc printto print = tmp4 new;
run;
proc univariate data = belsley normal;
 var vif1-vif7 eta7;
run;
data last;
file seed;
seed=&seed1;
put seed 10.;
```

```
run;
X "delete dsk3:[flom]91.dat;*";
%mend dissert;
%dissert;
```

Program 3: Models with no interaction and 7 IVs

```
%macro Dissert;
options nosource nonotes nocenter nosymbolgen nomprint nomlogic;
/* A program to generate data and test for collinearity */
/* This section assigns macro variables */
   filename seeds 'seed.dat';
   data start;
   infile seeds;
   input seed ;
   call symput('seed1',seed);
   run;
   %do i=1 %to 1000;
   %let q1= 1;
   %let f = 10;
/* This section creates 5 random variables ~ U(0,1). It also creates a dummy
variable, for use below */
   Title 'A program to generate data';
   title2 "For formula &f";
   data one;
```

```
seed=&seed1;
do I = 1 to 100;
   call rannor(seed,X0);
   call rannor(seed,X1);
   call rannor(seed,X2);
   call rannor(seed,X3);
   call rannor(seed,X4);
   call rannor(seed,X5);
   call rannor(seed,X6);
   call rannor(seed,X7);
   call rannor(seed,Y);
   XQ = X1 + X2;
   output;
end;
call symput('seed1',seed);
run;
/*This section computes the variance of X4, for use below */
proc means data=one noprint;
 var XQ;
 output out=two var(XQ) = var_XQ;
run;
/* This section combines the data from the first two sections */
title3 "i = &q1";
```

```
data three;
seed=&seed1;
set one;
if _n_=1 then set two(keep = var_XQ);
call rannor(seed,rn1);
vd1 = sqrt(var_XQ/(10 ** &q1)) * rn1;
do I = 1 to 100;
  XC = XQ  + vd1;
end;
call symput('seed1',seed);
run;
/* This section produces collinearity diagnostics, outputs them to a file, and
computes statistics about them */
filename newout "&f&q1";
proc printto print = newout;
run;
proc reg data=three;
model y = X1 X2 X3 X4 X5 X6 X7 XC/collin vif;
run;
proc printto;
run;
data Belsley;
infile newout lrecl=500 recfm=f;
```

```
input word1 $ @;
if word1 = 'Number' then
do;
 input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31 vp41 vp51 vp61 vp71
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32 vp42 vp52 vp62 vp72
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33 vp43 vp53 vp63 vp73
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34 vp44 vp54 vp64 vp74
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35 vp45 vp55 vp65 vp75
/ n6 eig6 eta6 vpint6 vp16 vp26 vp36 vp46 vp56 vp66 vp76
/ n7 eig7 eta7 vpint7 vp17 vp27 vp37 vp47 vp57 vp67 vp77
/ n8 eig8 eta8 vpint8 vp18 vp28 vp38 vp48 vp58 vp68 vp78
/ n9 eig9 eta9 vpint9 vp19 vp29 vp39 vp49 vp59 vp69 vp79;
 keep n1--vp79;
output;
end;
else if word1 = 'Variable' then
do;
 input
#2 var1 $ df1 parest1 se1 t1 prob1 vif1
/ var2 $ df2 parest2 se2 t2 prob2 vif2
/ var3 $ df3 parest3 se3 t3 prob3 vif3
/ var4 $ df4 parest4 se4 t4 prob4 vif4
```

```
 / var5 S df5 parest5 se5 t5 prob5 vif5
 / var6 S df6 parest6 se6 t6 prob6 vif6
 / var7 S df7 parest7 se7 t7 prob7 vif7
 / var8 S df8 parest8 se8 t8 prob8 vif8
 / var9 S df9 parest9 se9 t9 prob9 vif9;
  keep vif1-vif9;
  output;
  end;
 run;
 %end;
/* This section prints the means and sds to a file */
 filename tmp "means&f&q1";
 proc printto print=tmp new;
 run;
 proc means data=Belsley maxdec = 2 n mean std min max cv fw = 8;
 run;
 proc printto;
 run;
/* This section creates 2 files for Excel */
 options ls = 79;
 filename tmp2 "vif&f&q1";
 filename tmp3 "eta&f&q1";
 proc printto log = tmp2 new;
```

```
run;
data today;
 set belsley (keep = vif1 - vif9);
 put (vif1 - vif9) (6.0);
 where vif1 ne .;
run;
proc printto log = tmp3 new;
run;
data today2;
 set belsley (keep = eta1 - eta9);
 put (eta1 - eta9) (6.0);
 where eta1 ne .;
run;
proc printto;
run;
/* This section generates tests of normality */
options ls = 79;
filename tmp4 "uni&f&q1";
proc printto print = tmp4 new;
run;
proc univariate data = belsley normal;
 var vif2 - vif9 eta9;
run;
```

```
data last;
file seed;
seed=&seed1;
put seed 10.;
run;
X "delete dsk3:[flom]101.dat;*";
%mend dissert;
%dissert;
```

Program 4: Models with an interaction and 3 IVs

```
%macro Dissert;
options  nosource nonotes nocenter nosymbolgen nomprint nomlogic;
/* A program to generate data and test for collinearity */
/* This section assigns macro variables */
filename seeds 'seed.dat';
data start;
infile seeds;
input seed ;
   call symput('seed1',seed);
   run;
   %do i=1 %to 1000;
   %let q1=1;
   %let f = 11;
```

```
/* This section creates 5 random variables ~ U(0,1), it also creates a dummy
variable, for use below */
   Title 'A program to  generate data';
   title2 "For formula &f";
   data one;
   seed=&seed1;
   do I = 1 to 100;
      call rannor(seed,X0);
      call rannor(seed,X1);
      call rannor(seed,X2);
      call rannor(seed,X3);
      call rannor(seed,Y);
      XQ = X1 + X2;
      XM = X1 * X2;
      output;
   end;
   call symput('seed1',seed);
   run;
/*This section computes the variance of X4, for use below */
   proc means data=one noprint;
   var XQ;
   output out=two var(XQ) = var_XQ;
   run;
```

```
/* This section combines the data from the first two sections */
   title3 "i = &q1";
   data three;
   seed=&seed1;
    set one;
    if _n_=1 then set two(keep = var_XQ);
    call rannor(seed,rn1);
    vd1 = sqrt(var_XQ/(10 ** &q1)) * rn1;
    do I = 1 to 100;
     XC = XQ + vd1;
    end;
   call symput('seed1',seed);
   run;
/* This section produces collinearity diagnostics, outputs them to a file, and
computes statistics about them */
   filename newout "&f&q1";
   proc printto print =newout;
   run;
    proc reg data=three;
    model y = X1 X2 X3 XM XC/collin vif;
   run;
   proc printto;
   run;
```

```
data Belsley;
infile newout lrecl=82 recfm=f;
 input word1 $ @;
 if word1 = 'Number' then
 do;
   input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31 vpM1
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32 vpM2
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33 vpM3
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34 vpM4
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35 vpM5
/ n6 eig6 eta6 vpint6 vp16 vp26 vp36 vpM6;
   keep n1--vpM6;
  output;
  end;
 else if word1 = 'Variable' then
  do;
   input
#2 var1 $ df1 parest1 se1 t1 prob1 vif1
/ var2 $ df2 parest2 se2 t2 prob2 vif2
/ var3 $ df3 parest3 se3 t3 prob3 vif3
/ var4 $ df4 parest4 se4 t4 prob4 vif4
/ var5 $ df5 parest5 se5 t5 prob5 vif5
```

```
/ var6 S df6 parest6 se6 t6 prob6 vif6;
 keep vif1-vif6;
 output;
 end;
run;
%end;
/* This section prints the means and sds to a file */
filename tmp "means&f&q1";
proc printto print=tmp new;
run;
proc means data=Belsley maxdec = 2 n mean std min max cv fw = 8;
run;
proc printto;
run;
/* This section creates 2 files for Excel */
options ls = 79;
filename tmp2 "vif&f&q1";
filename tmp3 "eta&f&q1";
  proc printto log = tmp2 new;
  run;
  data today;
   set belsley (keep = vif1 - vif6);
   put (vif1 - vif6) (6.0);
```

```
 where vif1 ne .;
run;
proc printto log = tmp3;
run;
data today2;
 set belsley (keep = eta1 - eta6);
 put (eta1 - eta6) (6.0);
 where eta1 ne .;
run;
proc printto;
run;
/* This section generates tests of normality */
options ls = 79;
filename tmp4 "uni&f&q1";
proc printto print = tmp4 new;
run;
   proc univariate data = belsley normal;
    var vif2 - vif6 eta6;
   run;
   data last;
   file seed;
   seed=&seed1;
   put seed 10.;
```

```
run;
x "delete dsk3:[flom]111.dat;";
%mend dissert;
%dissert;
```

Program 5: Models with an Interaction and 5 IVs

```
%macro Dissert;
options nosource nonotes nocenter nosymbolgen nomprint nomlogic;
/* A program to generate data and test for collinearity */
/* This section assigns macro variables */
filename seeds 'seed.dat';
data start;
infile seeds;
input seed ;
call symput('seed1',seed);
run;
%do i=1 %to 1000;
%let q1=1;
%let f = 12;
/* This section creates 5 random variables ~ U(0,1), it also creates a dummy
variable, for use below */
   Title 'A program to  generate data';
   title2 "For formula &f";
   data one;
```

```
seed=&seed1;
do I = 1 to 100;
   call rannor(seed,X0);
   call rannor(seed,X1);
   call rannor(seed,X2);
   call rannor(seed,X3);
   call rannor(seed,X4);
   call rannor(seed,X5);
   call rannor(seed,Y);
   XQ = X1 + X2;
   XM = X1*X2;
   output;
end;
call symput('seed1',seed);
run;
/*This section computes the variance of X4, for use below */
proc means data=one noprint;
var XQ;
output out=two var(XQ) = var_XQ;
run;
/* This section combines the data from the first two sections */
title3 "i = &q1";
data three;
```

```
seed=&seed1;
set one;
if _n_=1 then set two(keep = var_XQ);
call rannor(seed,m1);
vd1 = sqrt(var_XQ/(10 ** &q1)) * m1;
do I = 1 to 100;
  XC = XQ + vd1;
end;
call symput('seed1',seed);
run;
/* This section produces collinearity diagnostics, outputs them to a file, and
computes statistics about them */
filename newout "&f&q1";
proc printto print = newout;
run;
proc reg data=three;
model y = X1 X2 X3 X4 X5 XM XC/collin vif;
run;
proc printto;
run;
data Belsley;
infile newout lrecl=500 recfm=f;
```

```
 input word1 $ @;
 if word1 = 'Number' then
 do;
  input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31 vp41 vp51 vpM1
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32 vp42 vp52 vpM2
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33 vp43 vp53 vpM3
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34 vp44 vp54 vpM4
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35 vp45 vp55 vpM5
/ n6 eig6 eta6 vpint6 vp16 vp26 vp36 vp46 vp56 vpM6
/ n7 eig7 eta7 vpint7 vp17 vp27 vp37 vp47 vp57 vpM7
/ n8 eig8 eta8 vpint8 vp18 vp28 vp38 vp48 vp58 vpM8;
  keep n1--vpM8;
 output;
 end;
 else if word1 = 'Variable' then
 do;
  input
#2 var1 $ df1 parest1 se1 t1 prob1 vif1
/ var2 $ df2 parest2 se2 t2 prob2 vif2
/ var3 $ df3 parest3 se3 t3 prob3 vif3
/ var4 $ df4 parest4 se4 t4 prob4 vif4
/ var5 $ df5 parest5 se5 t5 prob5 vif5
```

```
    / var6 S df6 parest6 se6 t6 prob6 vif6
    / var7 S df7 parest7 se7 t7 prob7 vif7
    / var8 S df8 parest8 se8 t8 prob8 vif8;
     keep vif1-vif8;
     output;
     end;
   run;
   %end;
/* This section prints the means and sds to a file */
   filename tmp "means&f&q1";
   proc printto print=tmp new;
   run;
   proc means data=Belsley maxdec = 2 n mean std min max cv fw = 8;
   run;
   proc printto;
   run;
/* This section creates two files for Excel */
options ls = 79;
filename tmp2 "vif&f&q1";
filename tmp3 "eta&f&q1";
proc printto log = tmp2 new;
   run;
```

```
data today;
 set belsley (keep = vif1-vif8);
 put (vif1-vif8) (6.0);
 where vif1 ne .;
run;
proc printto log = tmp3 new;
run;
data today2;
 set belsley (keep = eta1 - eta8);
 put (eta1 - eta8) (6.0);
 where eta1 ne .;
run;
/* This section generates tests of normality */
options ls = 79;
filename tmp4 "uni&f&q1";
proc printto print = tmp4 new;
run;
proc univariate data = belsley normal;
 var vif2 - vif8 eta8;
run;
data last;
file seed;
```

```
seed=&seed1;
put seed 10.;
run;
X "delete dsk3:[flom]121.dat;*";
%mend dissert;
.%dissert;
```

Program 6: Models with an Interaction and 7 IVs

```
%macro Dissert;
options nosource nonotes nocenter nosymbolgen nomprint nomlogic;
/* A program to generate data and test for collinearity */
/* This section assigns macro variables */
   filename seeds 'seed.dat';
   data start;
   infile seeds;
   input seed ;
   call symput('seed1',seed);
   run;
   %do i=1 %to 1000;
   %let q1= 1;
   %let f = 13;
/* This section creates 5 random variables ~ U(0,1), it also creates a dummy
variable, for use below */
   Title 'A program to  generate data';
```

```
title2 "For formula &f";
data one;
seed=&seed1;
do I = 1 to 100;
   call rannor(seed,X0);
   call rannor(seed,X1);
   call rannor(seed,X2);
   call rannor(seed,X3);
   call rannor(seed,X4);
   call rannor(seed,X5);
   call rannor(seed,X6);
   call rannor(seed,X7);
   call rannor(seed,Y);
   XQ = X1 + X2;
   XM = X1*X2;
   output;
end;
call symput('seed1',seed);
run;
/*This section computes the variance of X4, for use below */
proc means data=one noprint;
var XQ;
output out=two var(XQ) = var_XQ;
```

```
run;
/* This section combines the data from the first two sections */
title3 "i = &q1";
data three;
seed=&seed1;
set one;
if _n_=1 then set two(keep = var_XQ);
call rannor(seed,rn1);
vd1 = sqrt(var_XQ/(10 ** &q1)) * rn1;
do I = 1 to 100;
  XC = XQ  + vd1;
end;
call symput('seed1',seed);
run;
/* This section produces collinearity diagnostics
   outputs them to a file, and computes statistics about them */
filename newout "&f&q1";
proc printto print = newout;
run;
proc reg data=three;
model y = X1 X2 X3 X4 X5 X6 X7 XM XC/collin vif;
run;
proc printto;
```

```
run;
data Belsley;
infile newout lrecl=500 recfm=f;
 input word1 $ @;
 if word1 = 'Number' then
 do;
   input
#2 n1 eig1 eta1 vpint1 vp11 vp21 vp31 vp41 vp51 vp61 vp71 vpM1
/ n2 eig2 eta2 vpint2 vp12 vp22 vp32 vp42 vp52 vp62 vp72 vpM2
/ n3 eig3 eta3 vpint3 vp13 vp23 vp33 vp43 vp53 vp63 vp73 vpM3
/ n4 eig4 eta4 vpint4 vp14 vp24 vp34 vp44 vp54 vp64 vp74 vpM4
/ n5 eig5 eta5 vpint5 vp15 vp25 vp35 vp45 vp55 vp65 vp75 vpM5
/ n6 eig6 eta6 vpint6 vp16 vp26 vp36 vp46 vp56 vp66 vp76 vpM6
/ n7 eig7 eta7 vpint7 vp17 vp27 vp37 vp47 vp57 vp67 vp77 vpM7
/ n8 eig8 eta8 vpint8 vp18 vp28 vp38 vp48 vp58 vp68 vp78 vpM8
/ n9 eig9 eta9 vpint9 vp19 vp29 vp39 vp49 vp59 vp69 vp79 vpM9
/ n10 eig10 eta10 vpint10 vp110 vp210 vp310 vp410 vp510 vp610 vp710
vpM10;
   keep n1--vpM10;
  output;
  end;
else if word1 = 'Variable' then
  do;
```

```
    input
   #2 var1 S df1 parest1 se1 t1 prob1 vif1
   / var2 S df2 parest2 se2 t2 prob2 vif2
   / var3 S df3 parest3 se3 t3 prob3 vif3
   / var4 S df4 parest4 se4 t4 prob4 vif4
   / var5 S df5 parest5 se5 t5 prob5 vif5
   / var6 S df6 parest6 se6 t6 prob6 vif6
   / var7 S df7 parest7 se7 t7 prob7 vif7
   / var8 S df8 parest8 se8 t8 prob8 vif8
   / var9 S df9 parest9 se9 t9 prob9 vif9
   / var10 S df10 parest10 se10 t10 prob10 vif10;
    keep vif1-vif10;
    output;
    end;
   run;
   %end;
  /* This section prints the means and sds to a file */
   filename tmp "means&f&q1";
   proc printto print=tmp new;
   run;
   proc means data=Belsley maxdec = 2 n mean std min max cv fw = 8;
   run;
   proc printto;
```

```
run;
/* This section creates 2 files for Excel */
options ls = 79;
filename tmp2 "vif&f&q1";
filename tmp3 "eta&f&q1";
proc printto log = tmp2 new;
run;
data today;
 set belsley (keep = vif1 - vif10);
 put (vif1 - vif10) (6.0);
 where vif1 ne .;
run;
proc printto log = tmp3 new;
run;
data today2;
 set belsley (keep = eta1 - eta10);
 put (eta1 - eta10) (6.0);
 where eta1 ne .;
run;
proc printto;
run;
/* This section generate tests of normality */
filename tmp4 "uni&f&q1";
```

```
        proc printto print = tmp4 new;
        run;
        proc univariate data = belsley normal;
         var vif2 - vif10 eta10;
        run;
        data last;
        file seed;
        seed=&seed1;
        put seed 10.;
        run;
        X "delete dsk3:[flom]131.dat;*";
        %mend dissert;
%dissert;
```

## Appendix E

### Distribution of the Statistics

One of the goals of this dissertation was to compare the performance of VIFs and Condition Indexes for each of the 18 models generated by the 3 x 3 x 2 design. Below are histograms of 1000 replications of the largest VIF and largest condition index for each of the models.

These histograms show graphically that, as discussed in Chapter 3, the VIFs have greater dispersion than the Condition Indexes for each of the models. They also show, again as discussed in Chapter 3, that both measures distinguish among weak, moderate, and strong collinearity.

**Figure 2**
**Largest eta, no interaction, 3 IVs, weak collinearity**

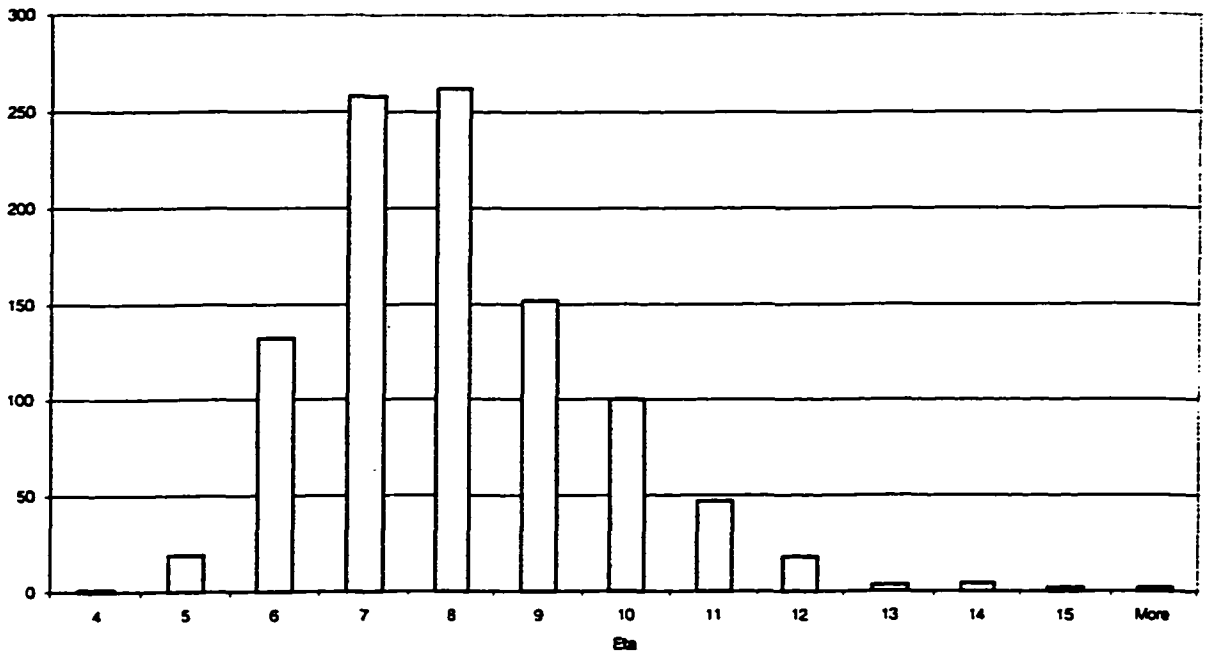



**Figure 3**
**Largest VIF, no interaction, 3 IVs, weak collinearity**

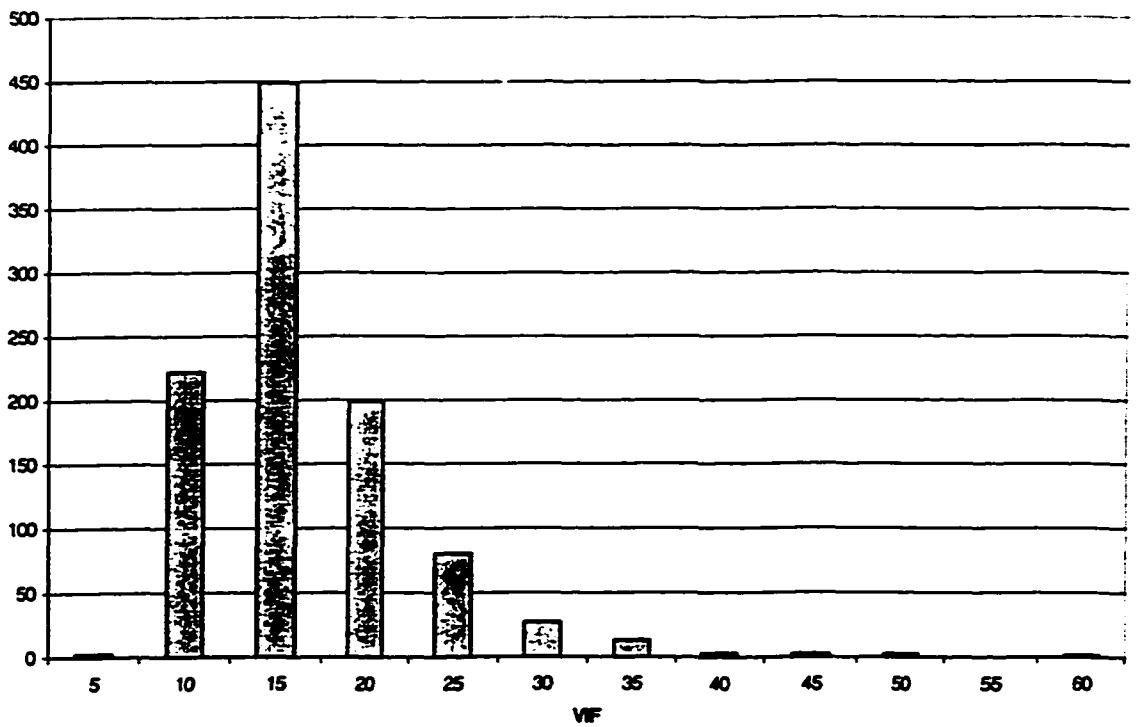


**Figure 4**
**Largest eta, no interaction, 3 IVs, moderate collinearity**

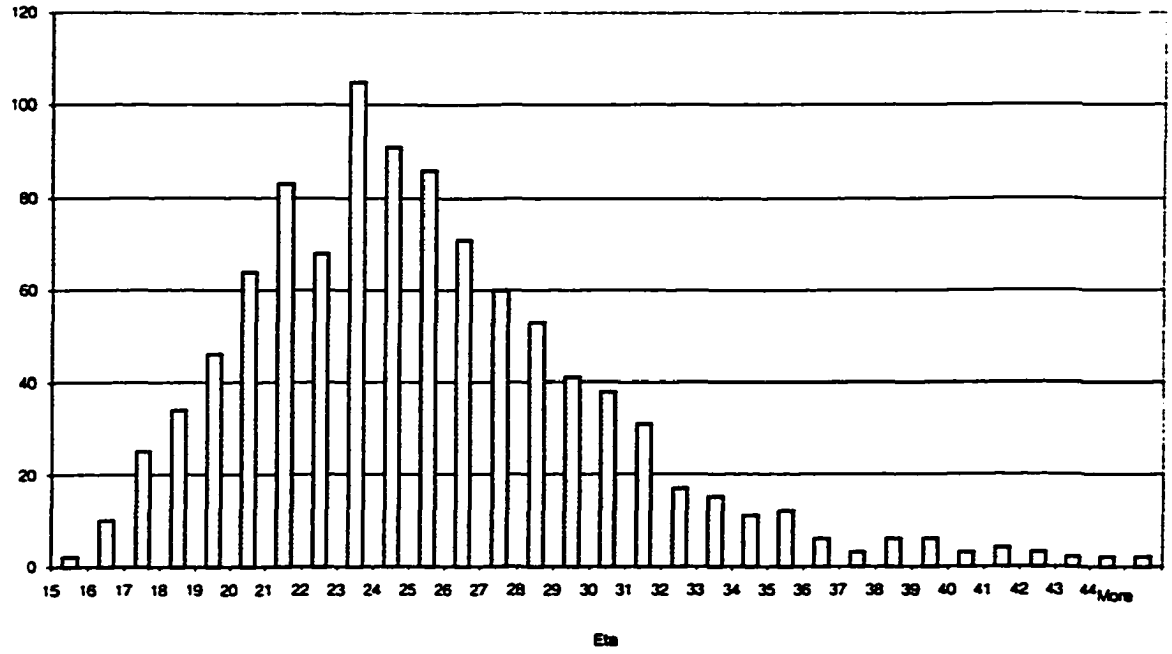



**Figure 5**
**Largest VIF, no interaction, 3 IVs, moderate collinearity**

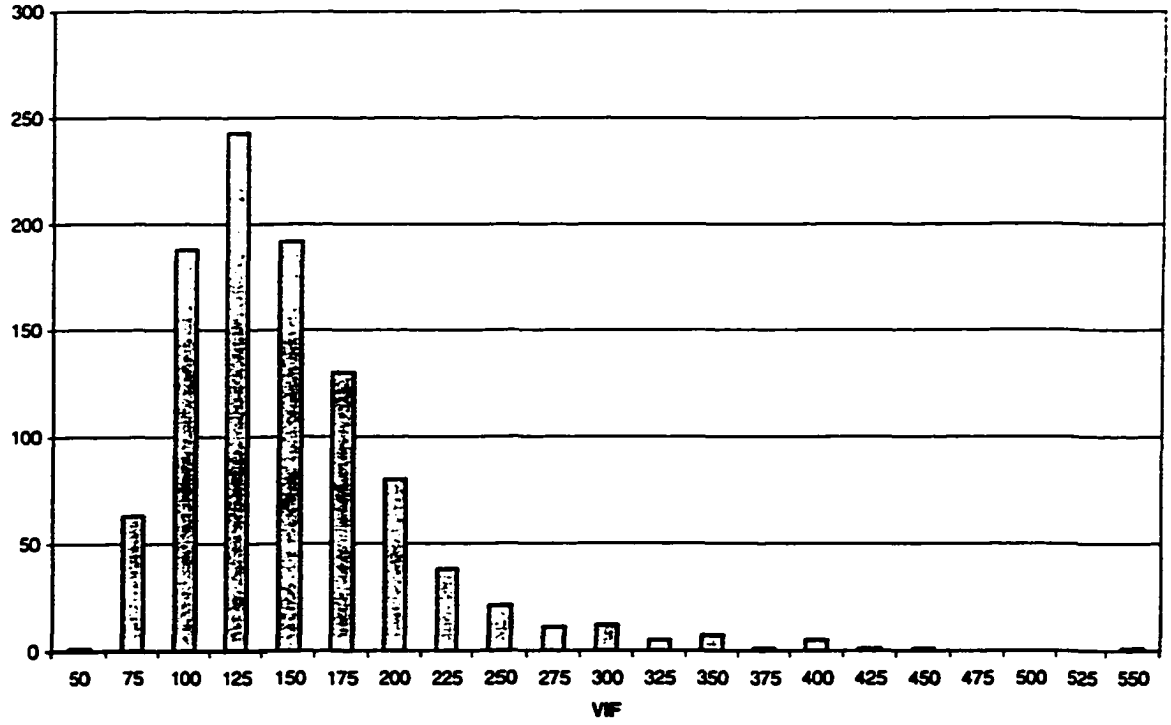


Figure 6
Largest eta, no interaction, 3 IVs, strong collinearity

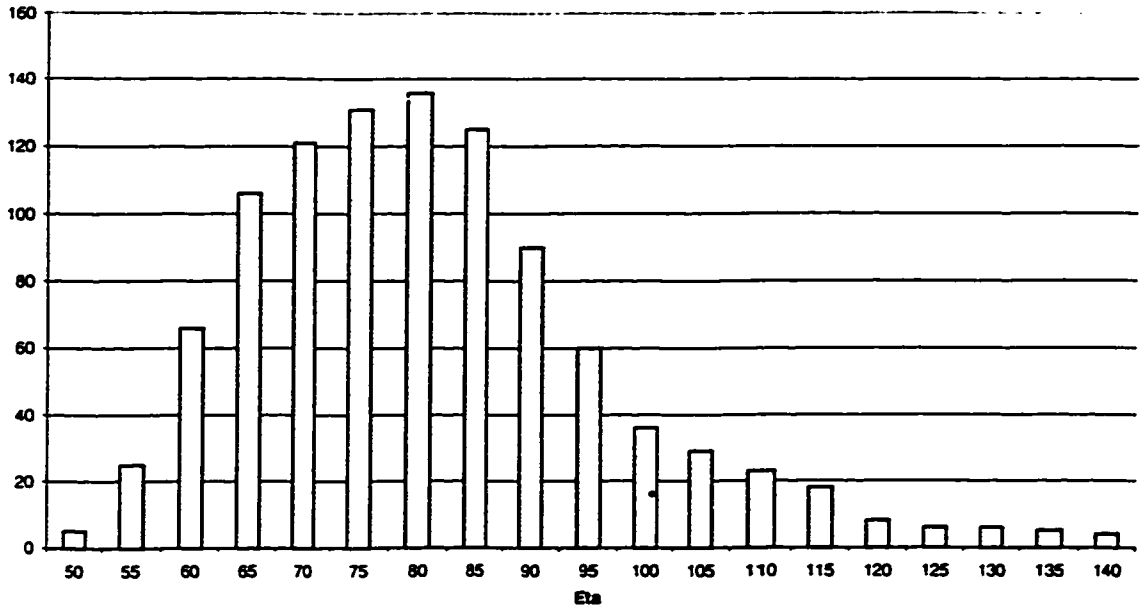



Figure 7
Largest VIF, no interaction, 3 IVs, strong collinearity

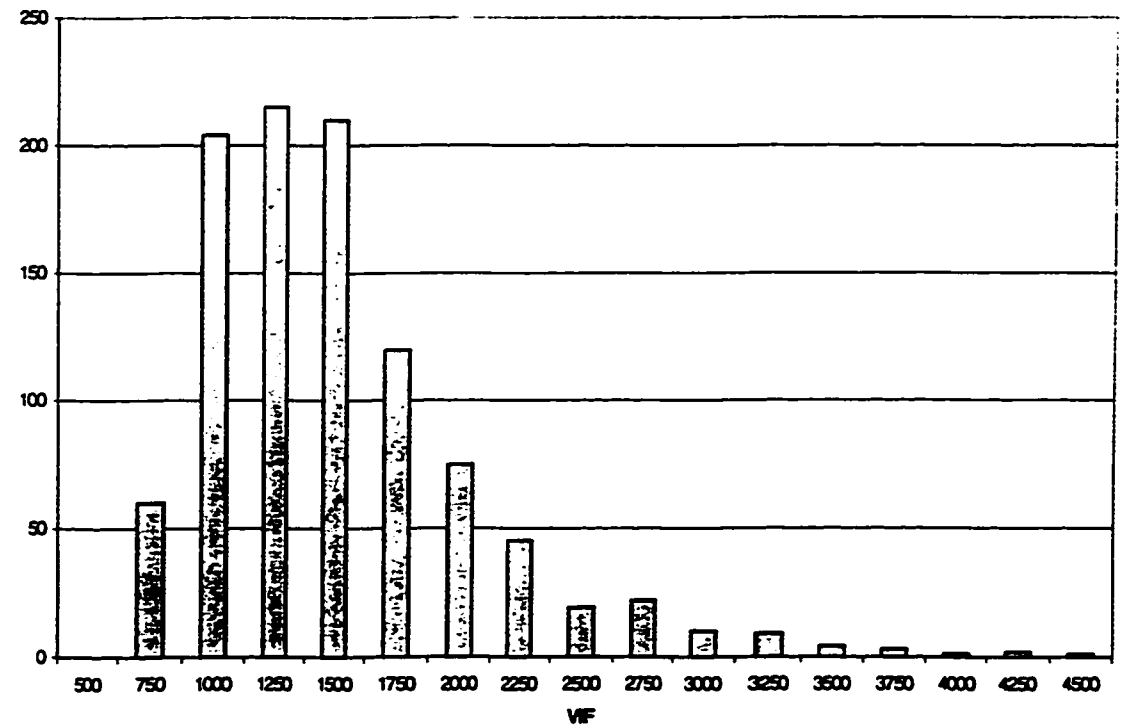


**Figure 8**
**Largest eta, no interaction, 5 IVs, weak collinearity**

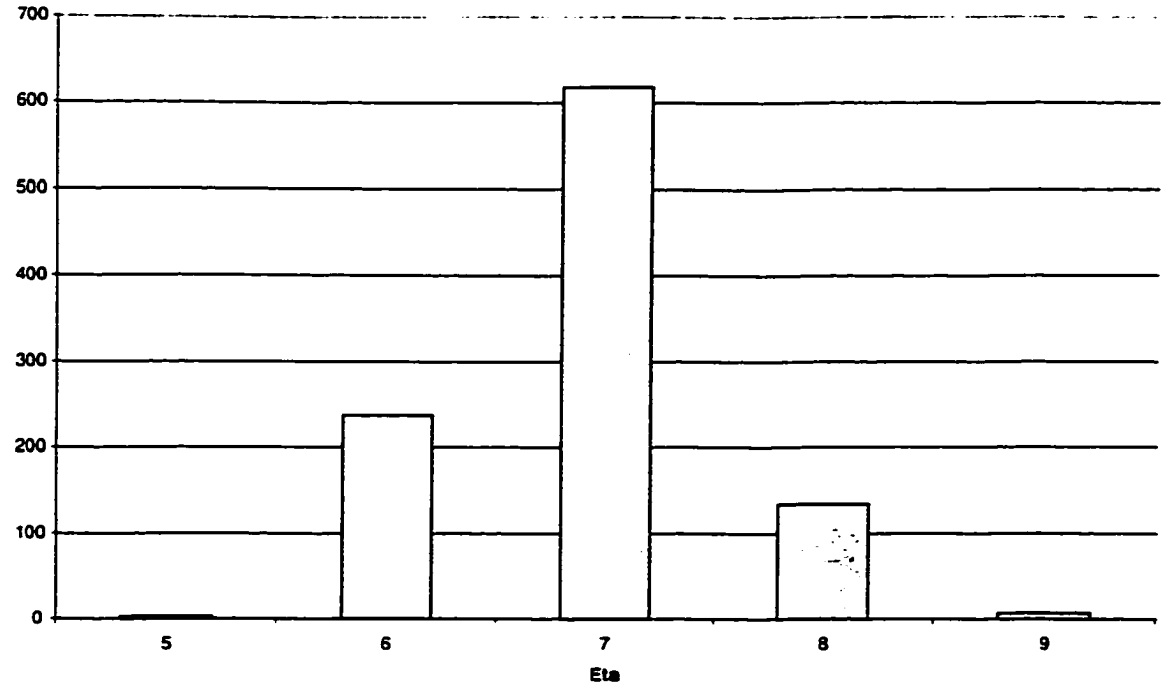



**Figure 9**
**Largest VIF, no interaction, 5 IVs, weak collinearity**

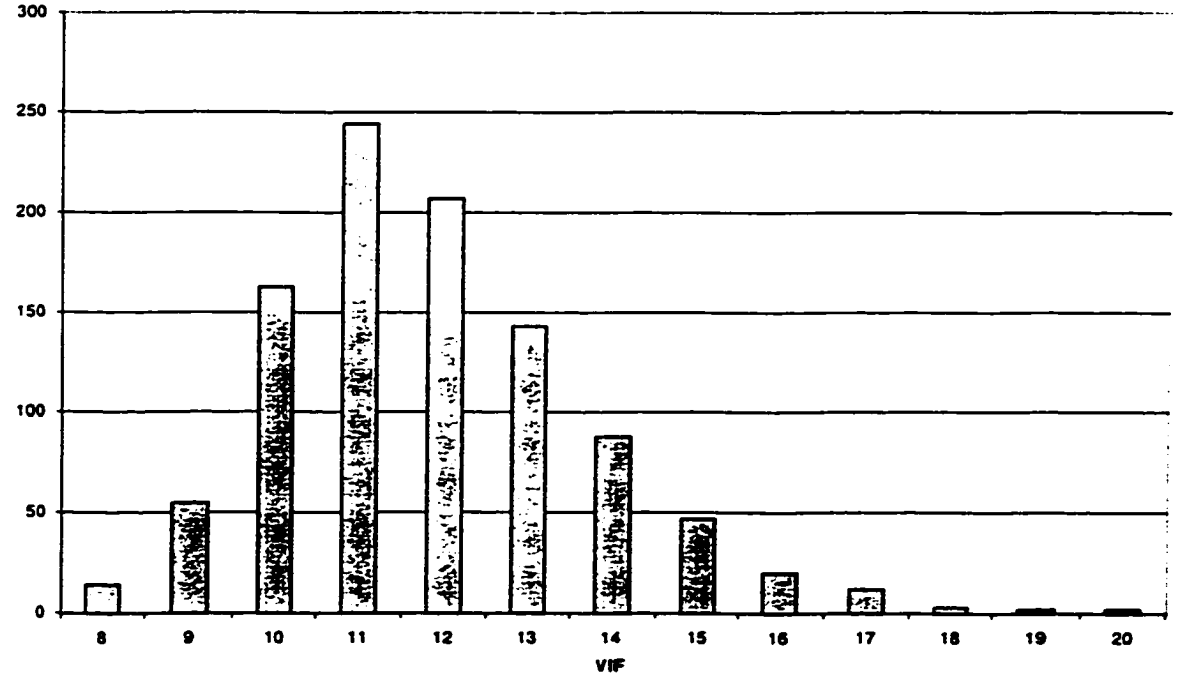


**Figure 10**
**Largest eta, no interaction, 5 IVs, moderate collinearity**

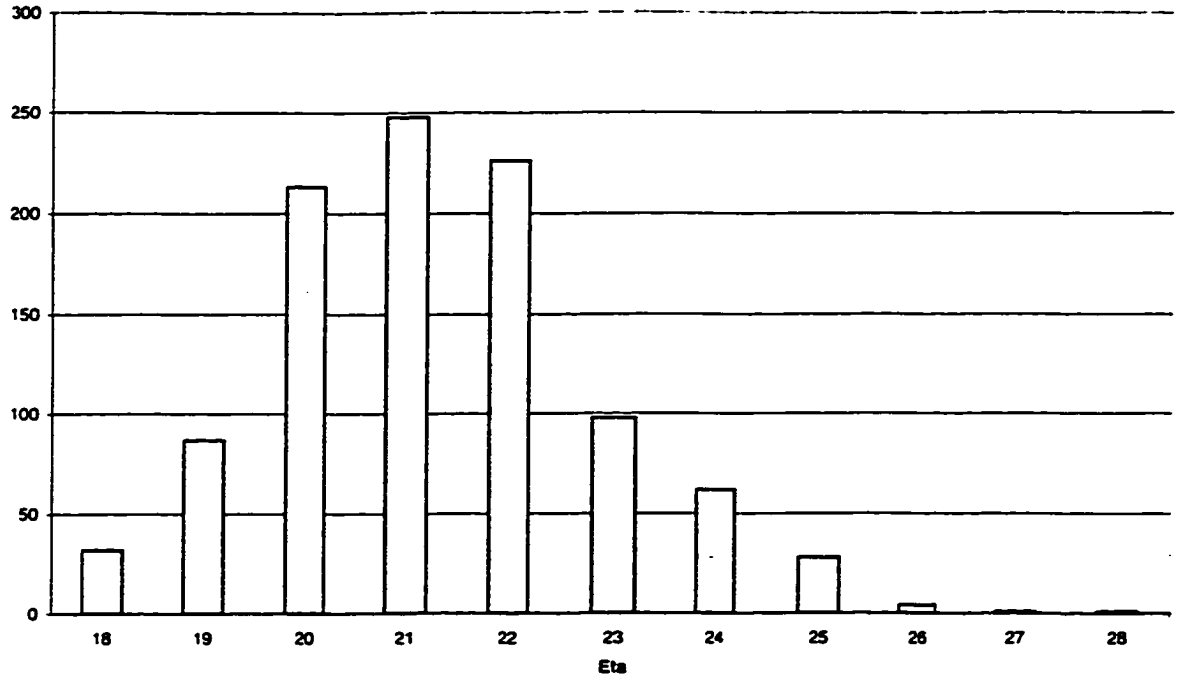



**Figure 11**
**Largest VIF, no interaction, 5 IVs, moderate collinearity**

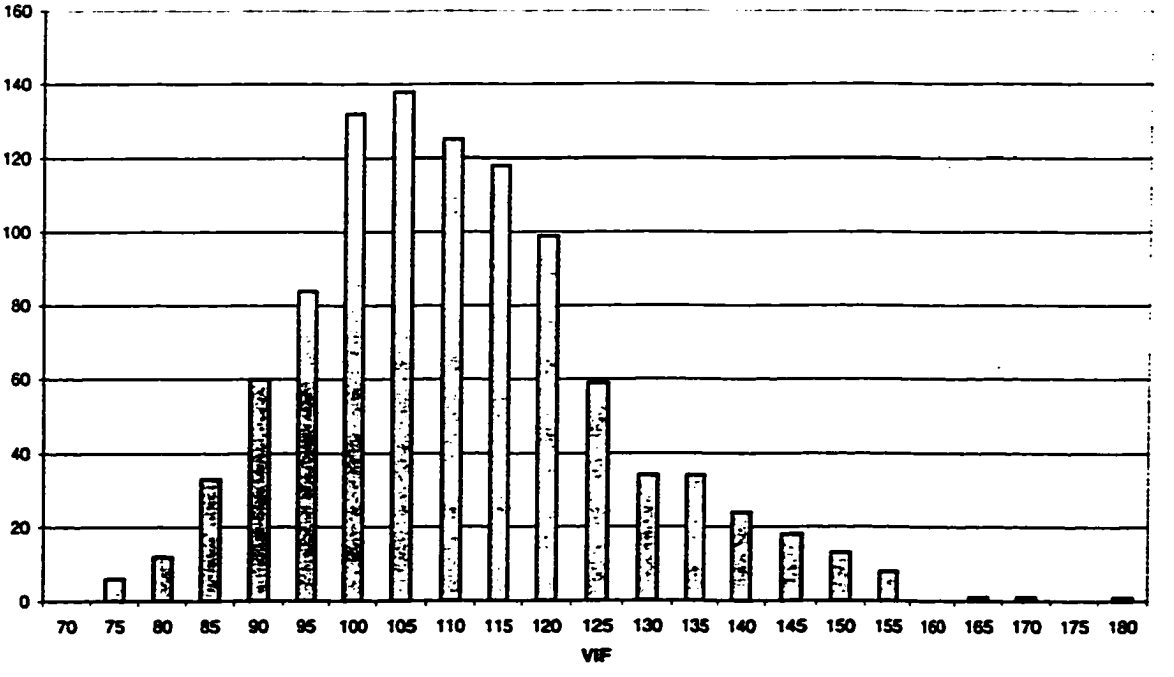


**Figure 12**
**Largest eta, 5 IVs, strong collinearity**

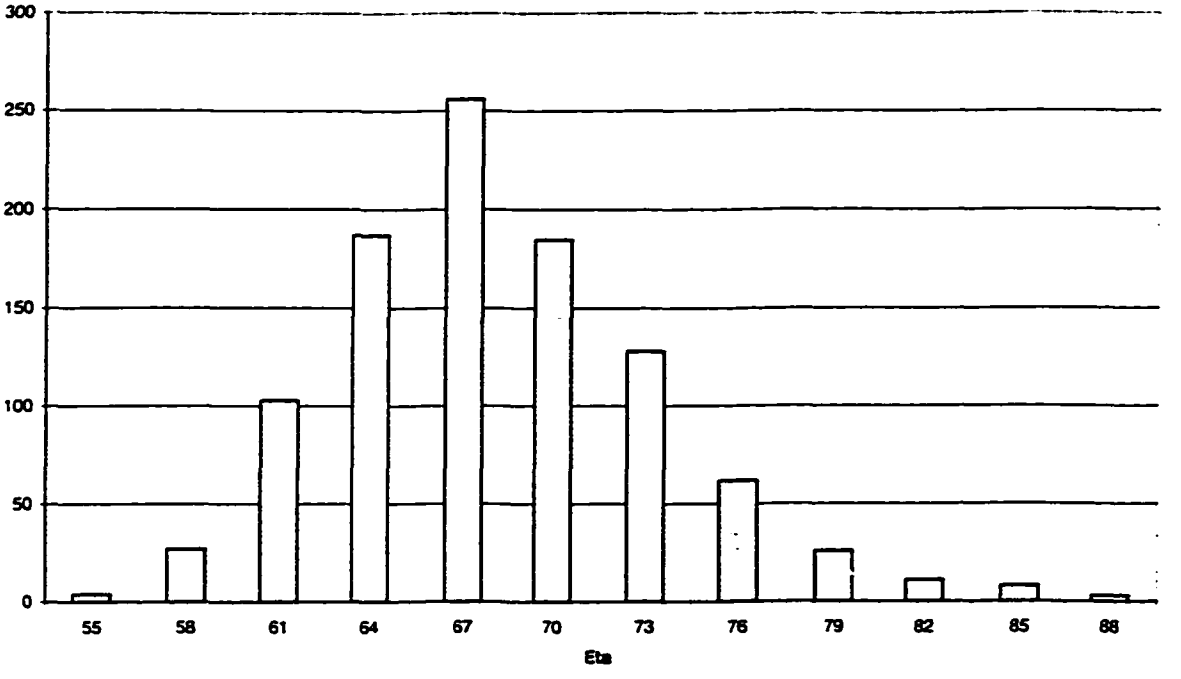



**Figure 13**
**Largest VIF, no interaction, 5 IVs, strong collinearity**

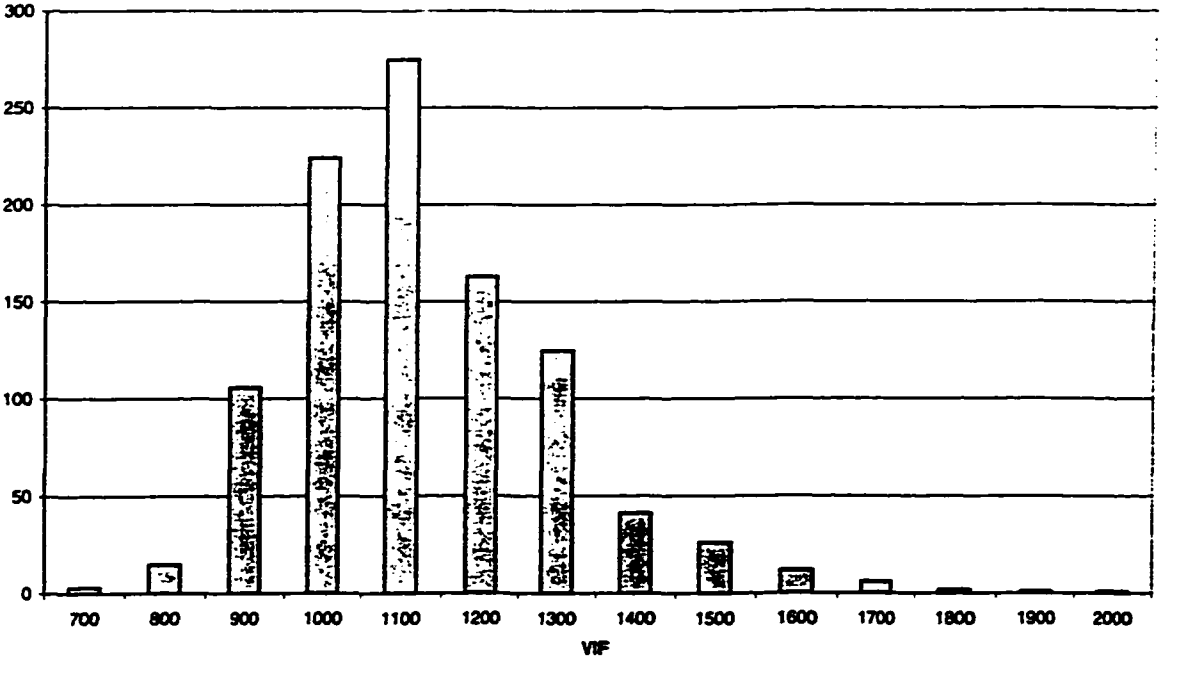


**Figure 14**
**Largest eta, no interaction, 7 IVs, weak collinearity**

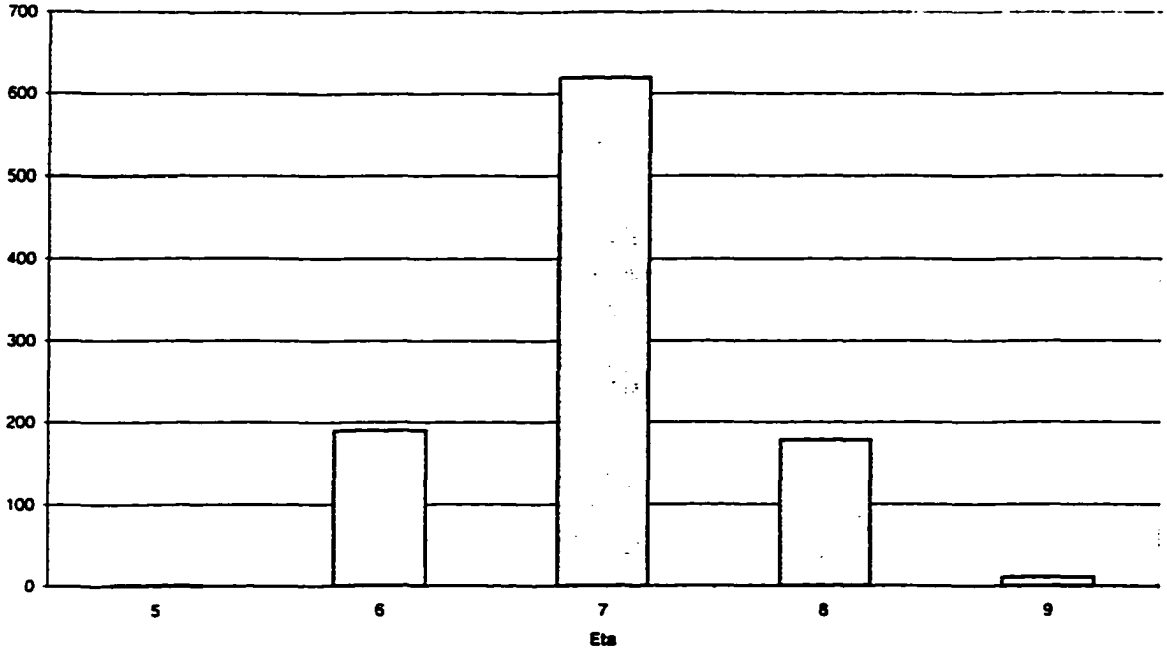



**Figure 15**
**Largest VIF, no interaction, 7 IVs, weak collinearity**

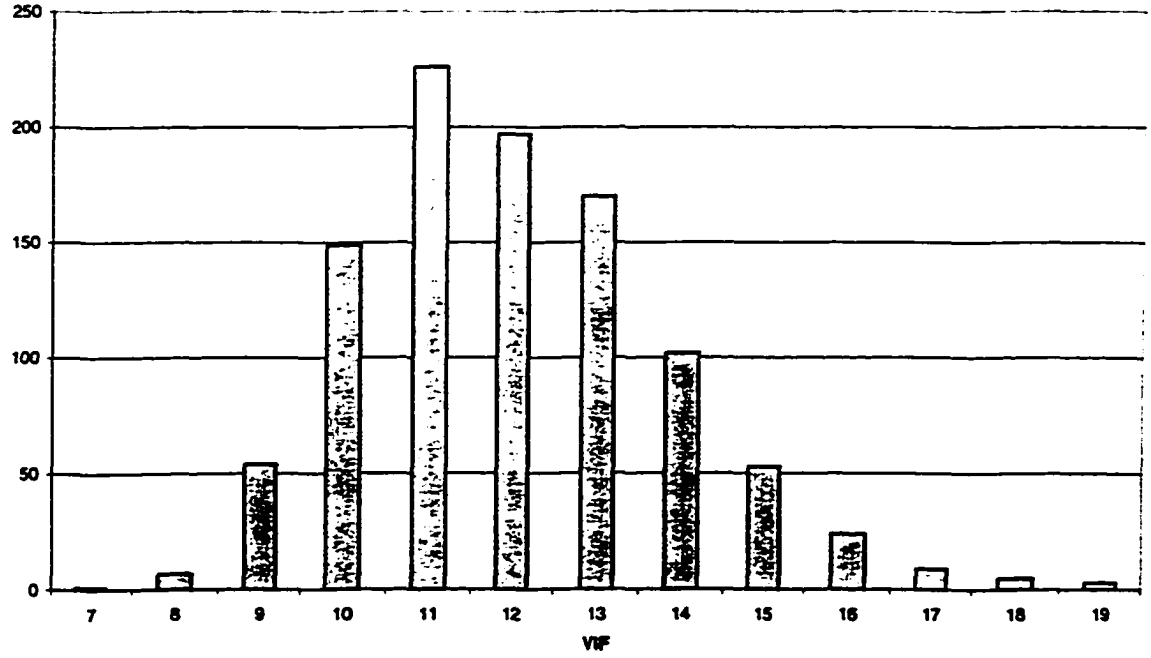


**Figure 16**
**Largest eta, no interaction, 7 IVs, moderate collinearity**

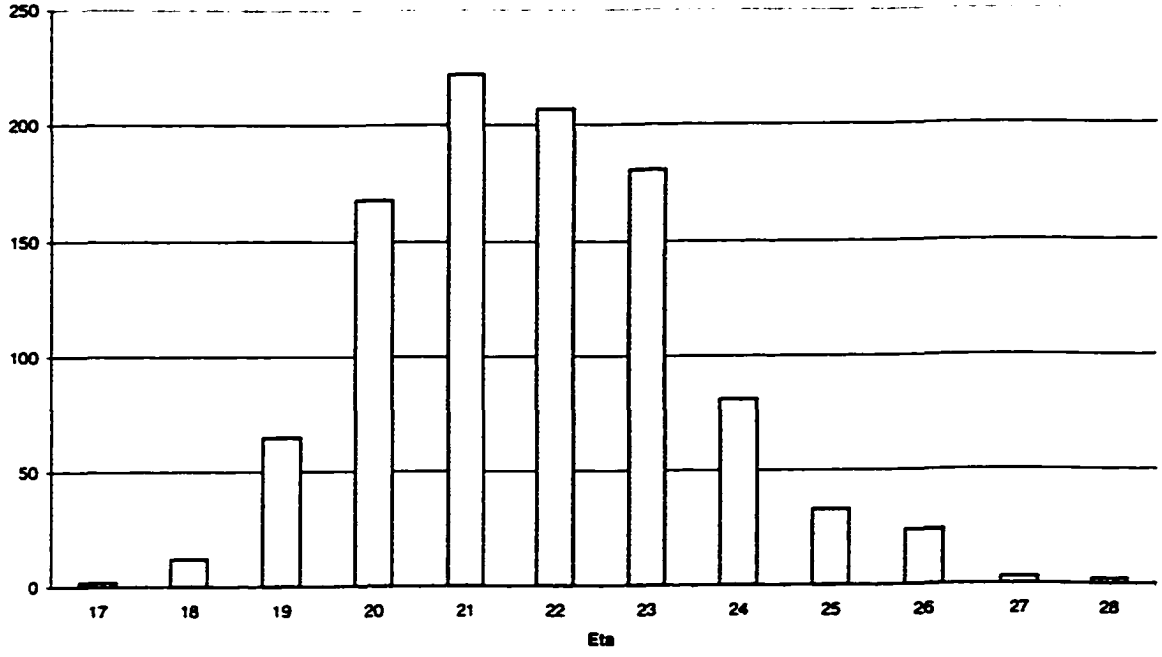



**Figure 17**
**Largest VIF, no interaction, 7 IVs, moderate collinearity**

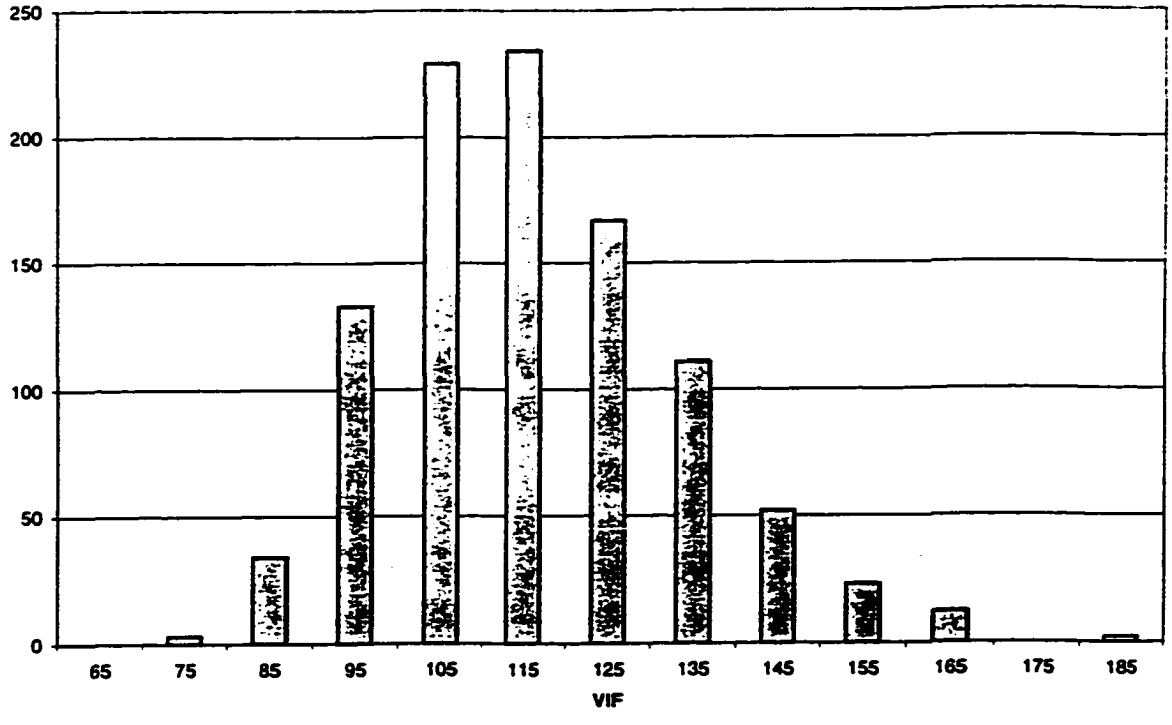


**Figure 18**
**Largest eta, no interaction, 7 IVs, stong collinearity**

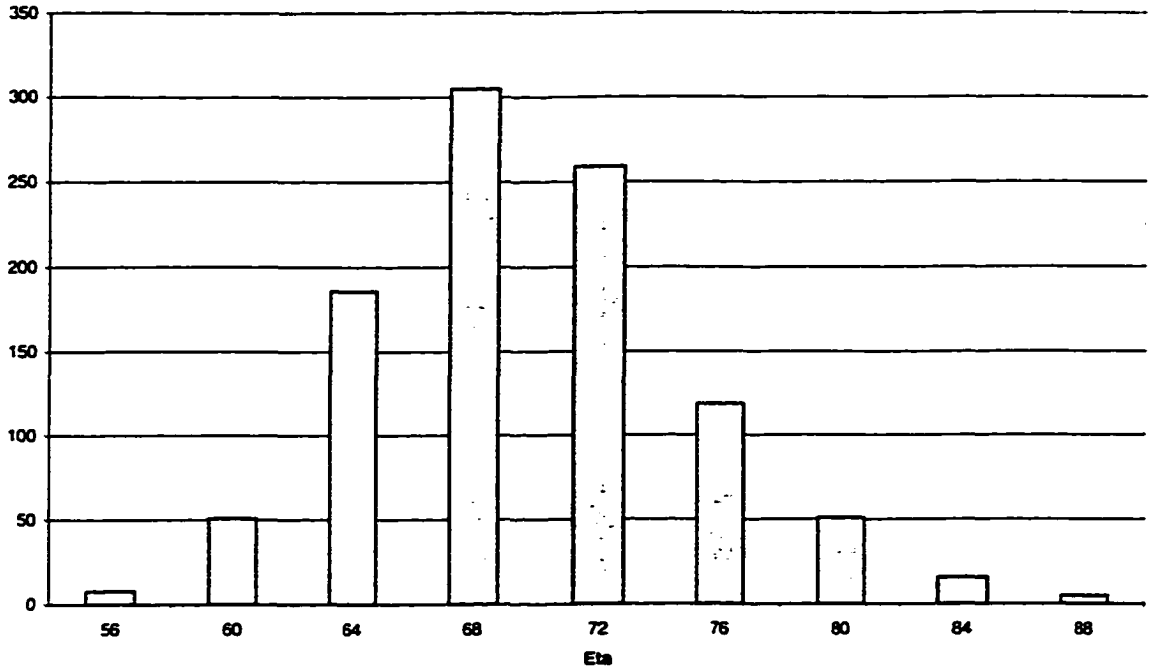



**Figure 19**
**Largest VIF, no interaction, 7 IVs, strong collinearity**

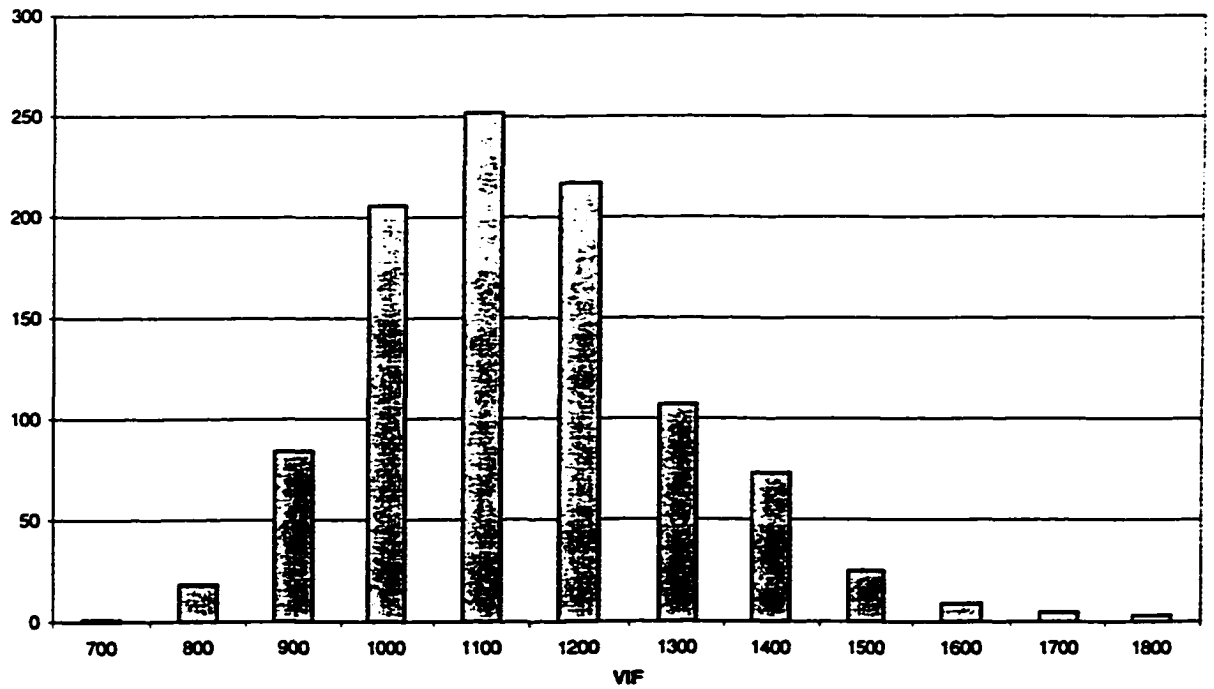


**Figure 20**
**Largest eta, interaction, 3 IVs, weak collinearity**

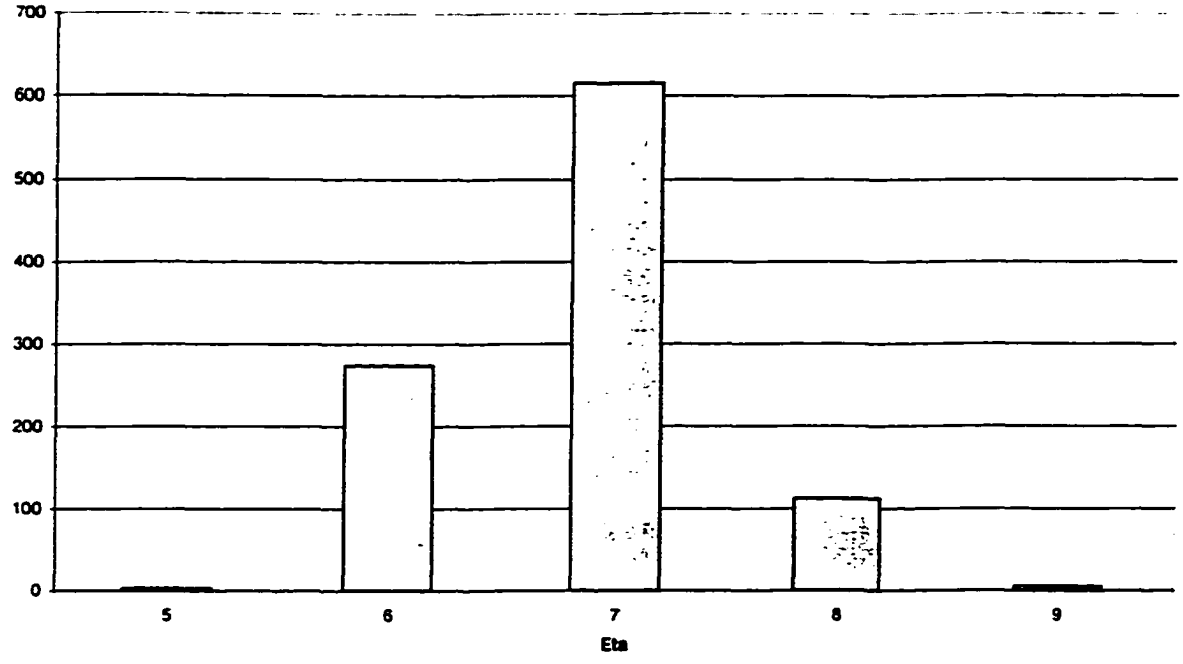



**Figure 21**
**Largest VIF, interaction, 3 IVs, weak collinearity**

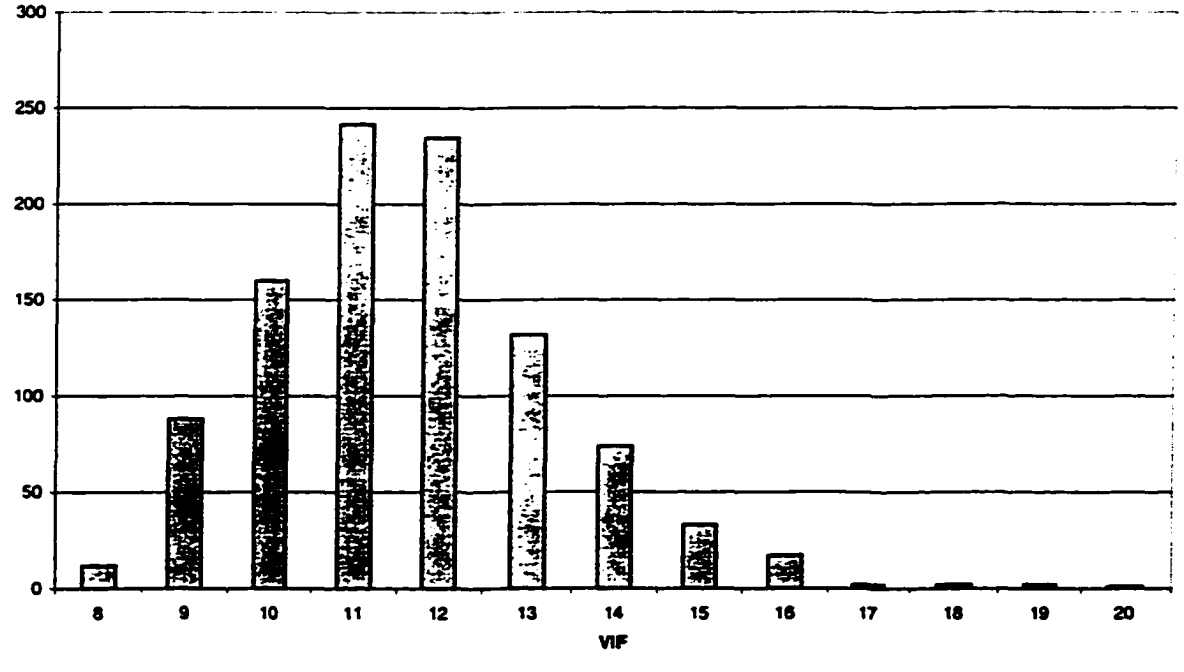


**Figure 22**
**Largest eta, interaction, 3 IVs, moderate collinearity**

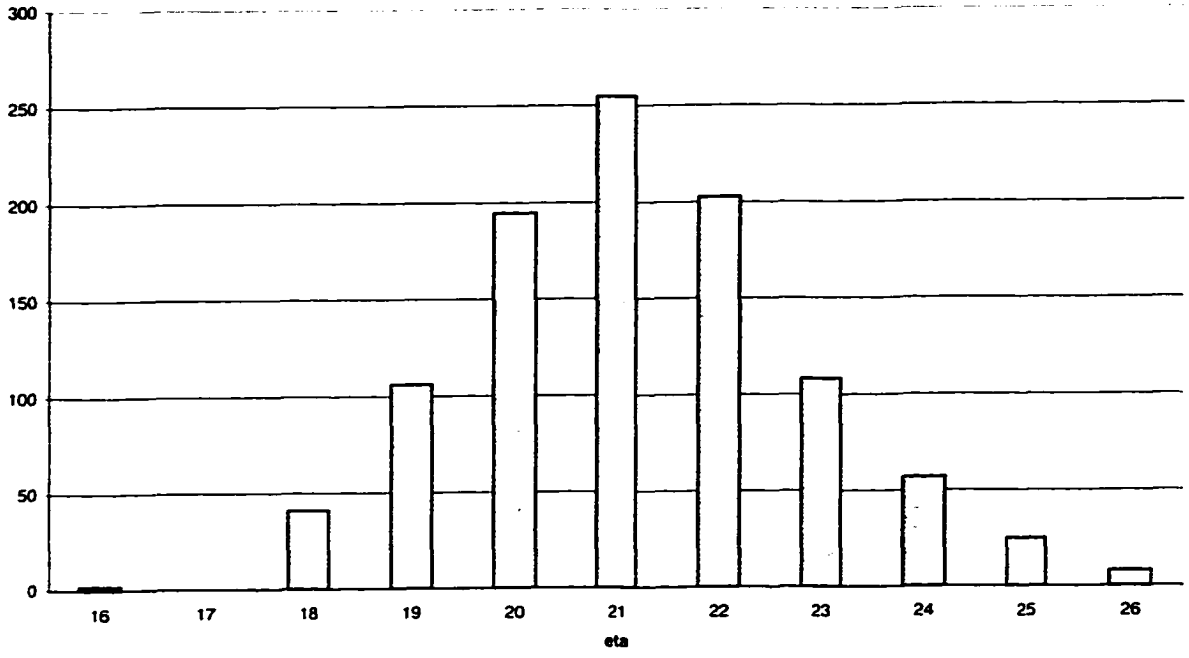



**Figure 23**
**Largest VIF, interaction, 3 IVs, moderate collinearity**

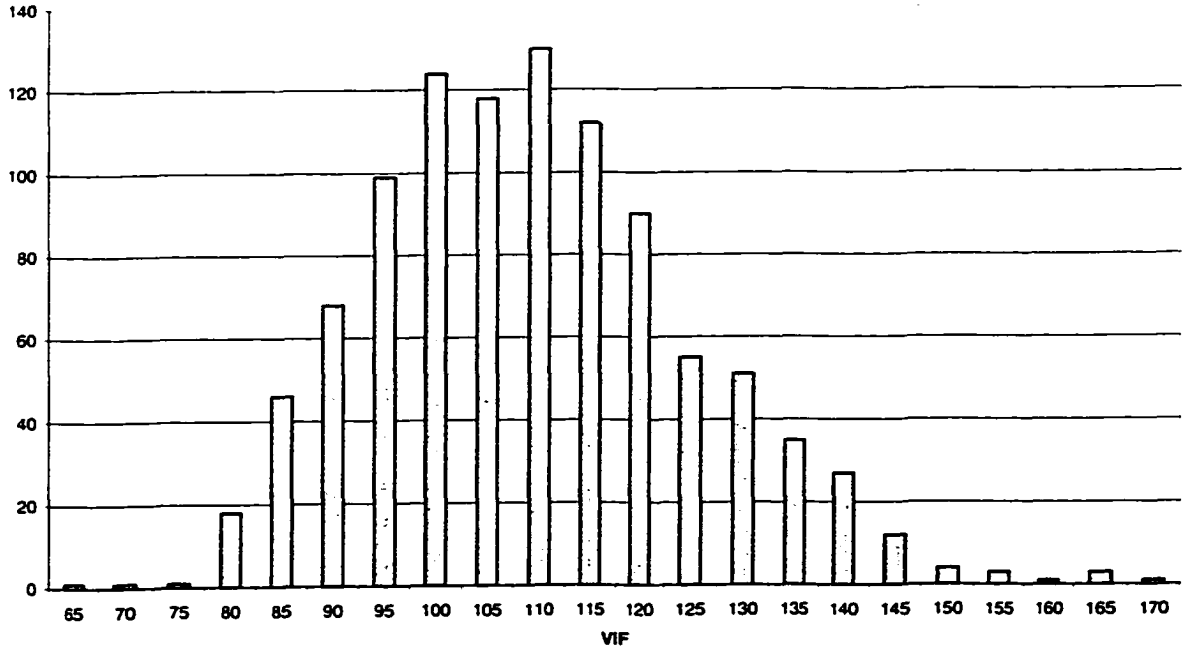


**Figure 24**
**Largest eta. interaction, 3 IVs, strong collinearity**

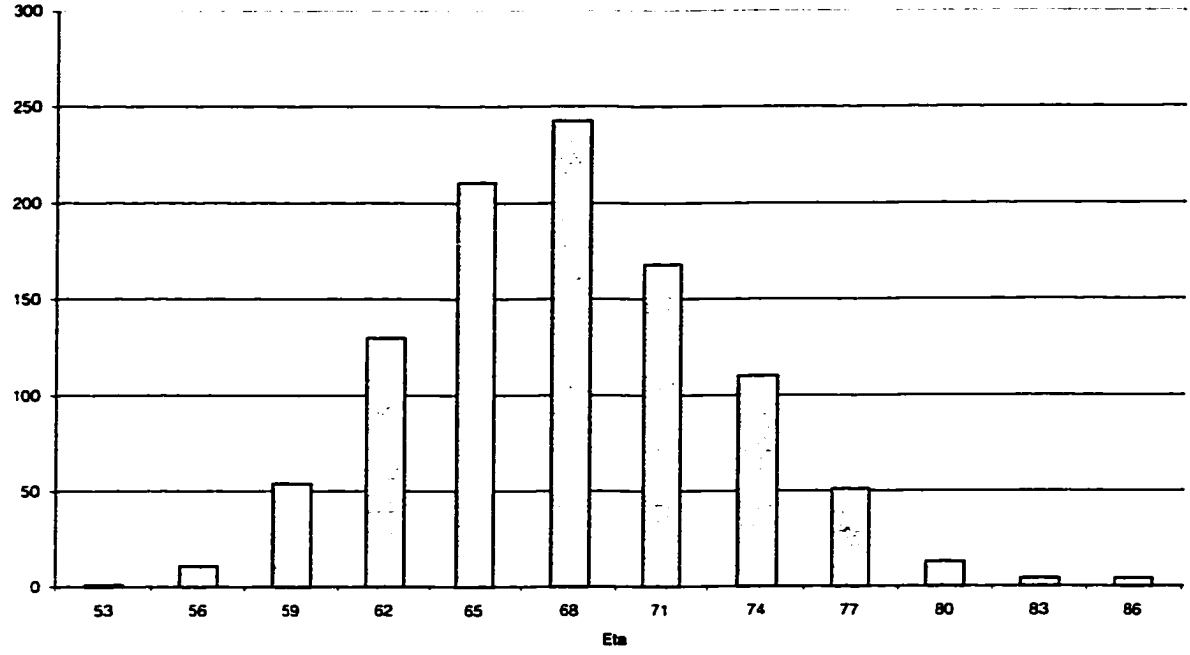



**Figure 25**
**Largest VIF, interaction, 3 IVs, strong collinearity**

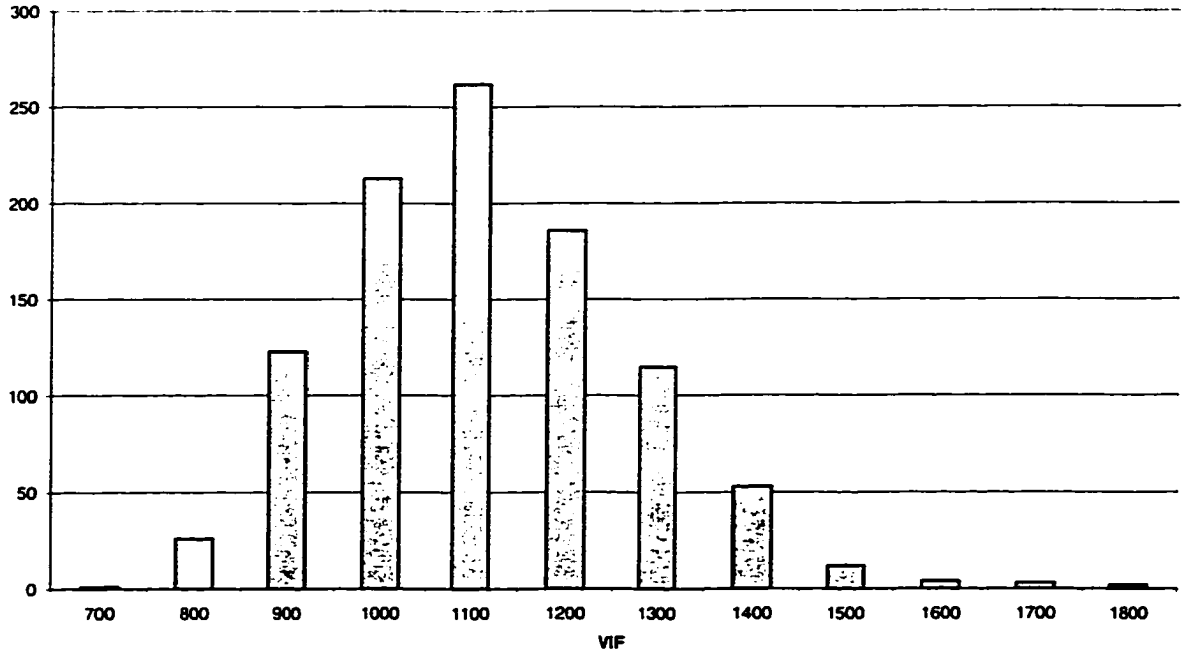


**Figure 26**
**Largest eta, interaction, 5 IVs, weak collinearity**

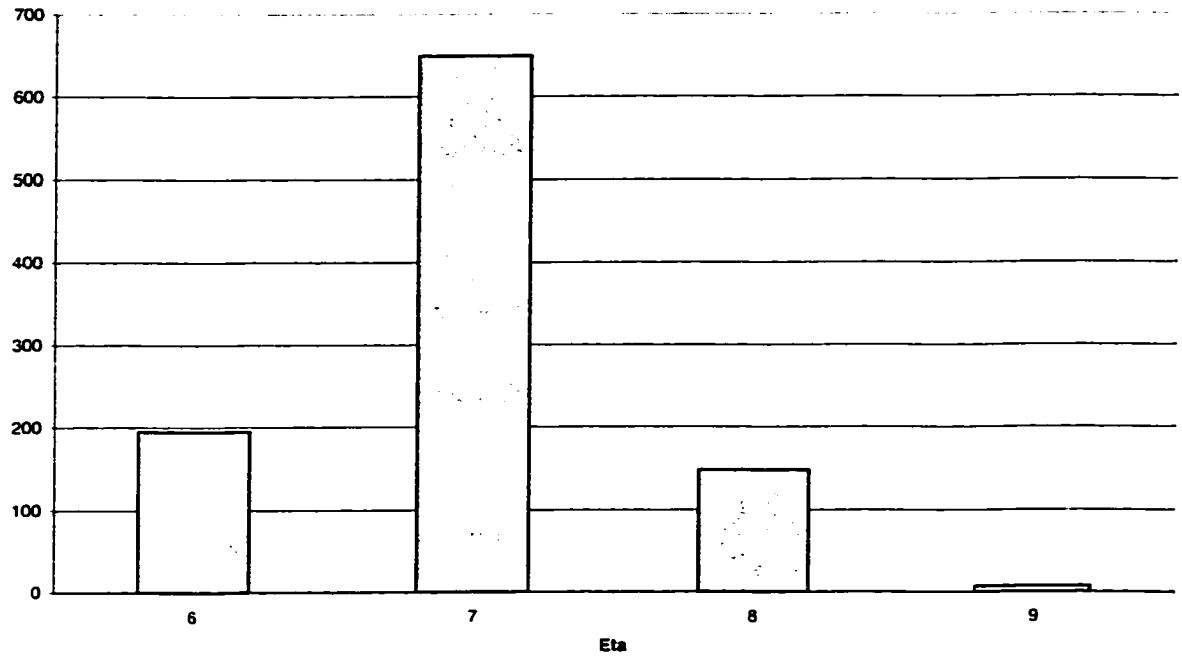



**Figure 27**
**Largest VIF, interaction, 5 IVs, weak collinearity**

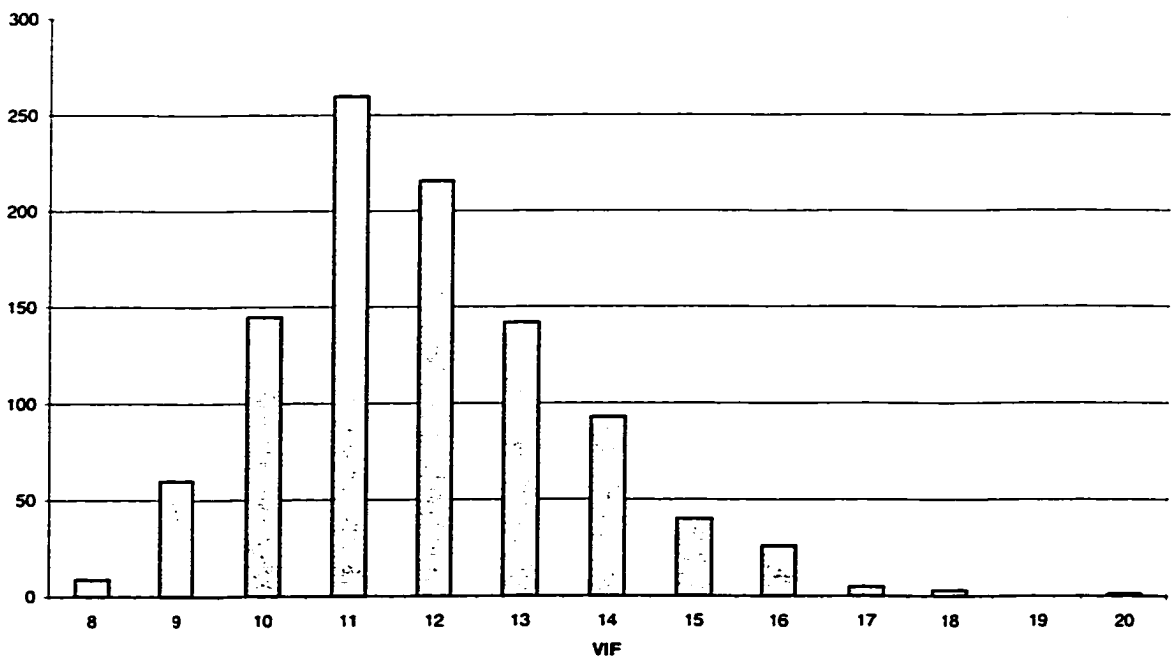


**Figure 28**
**Largest eta, interaction, 5 IVs, moderate collinearity**

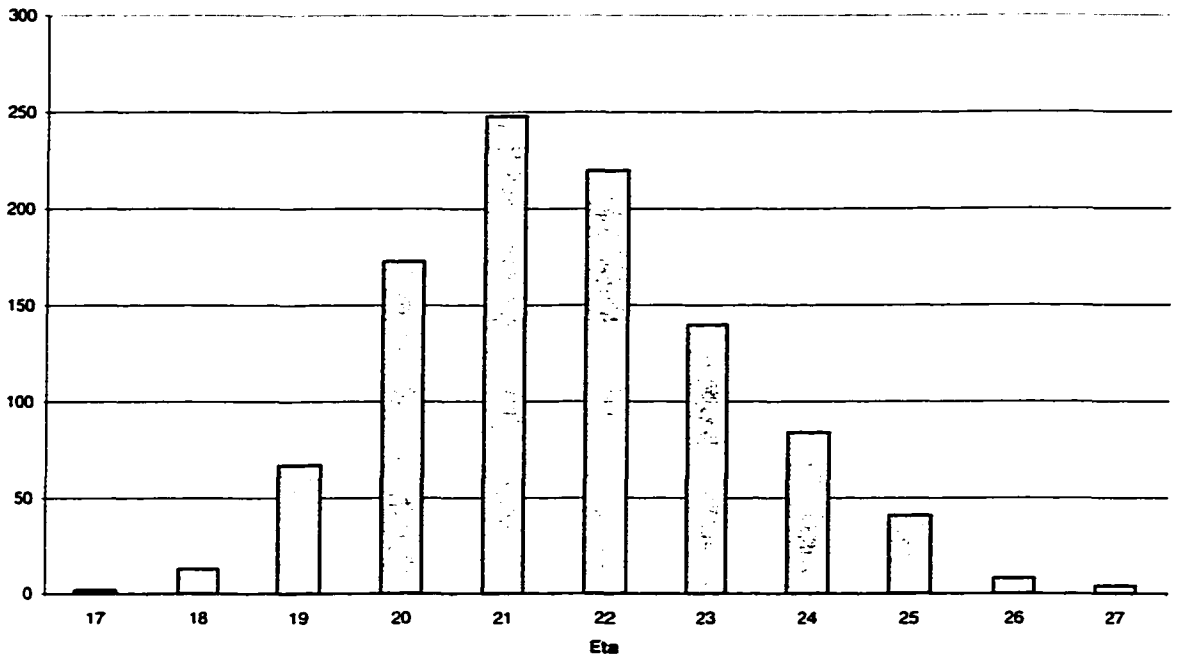



**Figure 29**
**Largest VIF, interaction, 5 IVs, moderate collinearity**

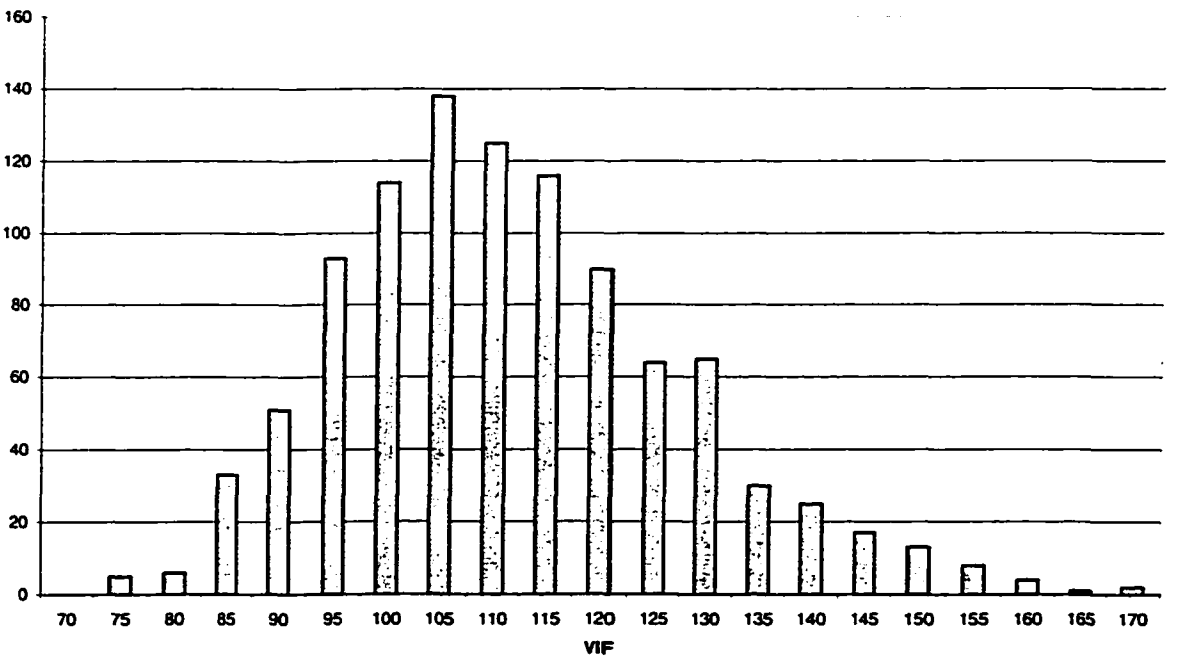


**Figure 30**
**Largest eta, interaction, 5 IVs, strong collinearity**

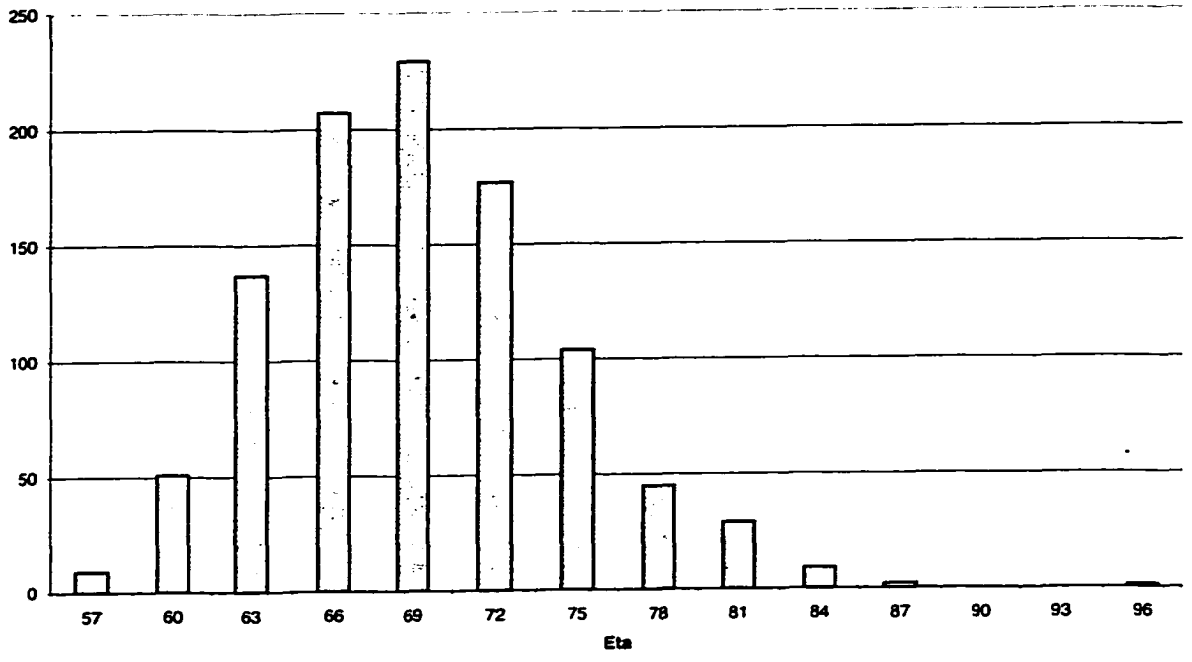



**Figure 31**
**Largest VIF, interaction, 5 IVs, strong collinearity**

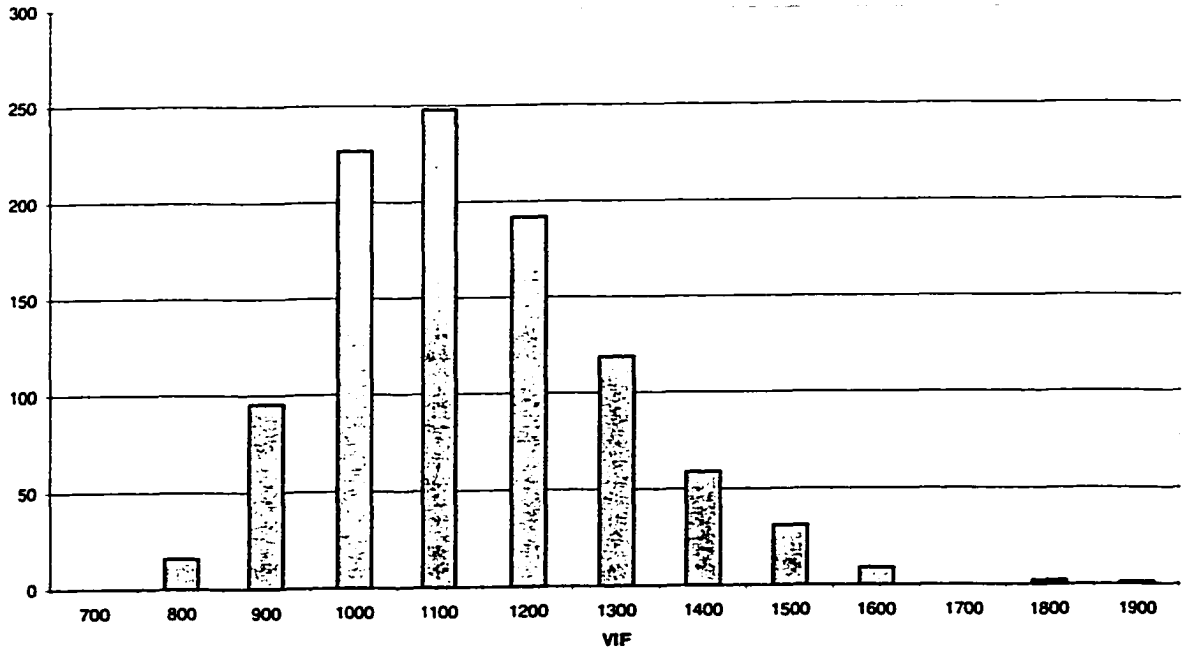


**Figure 32**
**Largest eta, interaction, 7 IVs, weak collinearity**

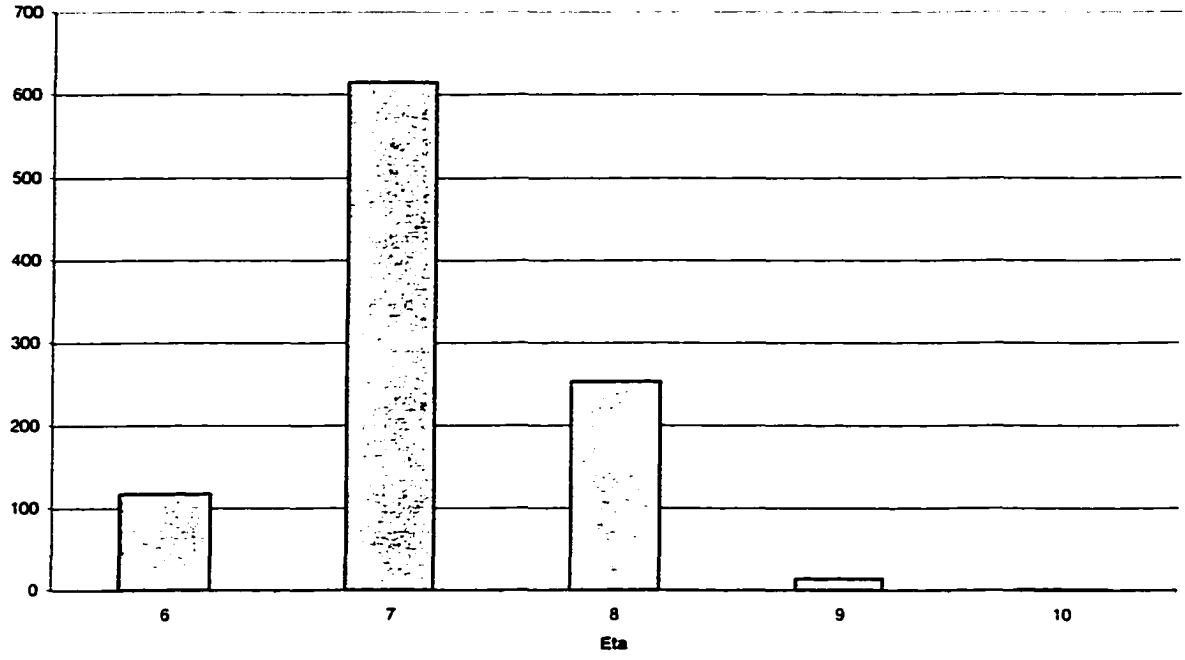



**Figure 33**
**Largest VIF, interaction, 7 IVs, weak collinearity**

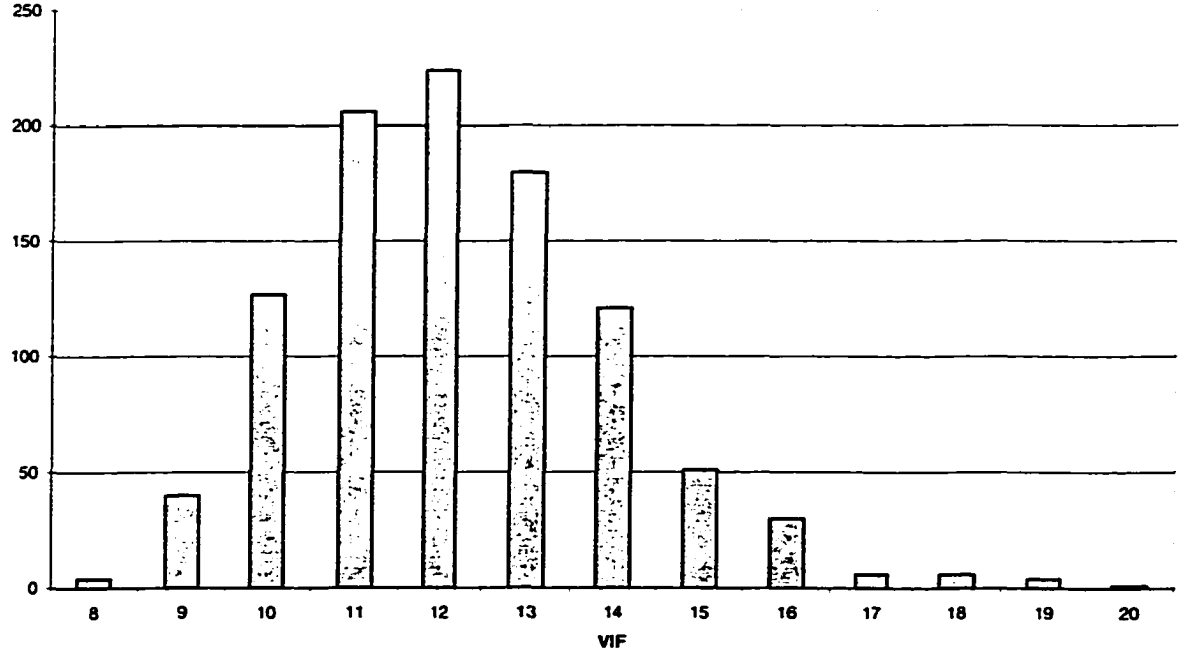


**Figure 34**
**Largest eta, interaction, 7 IVs, moderate collinearity**

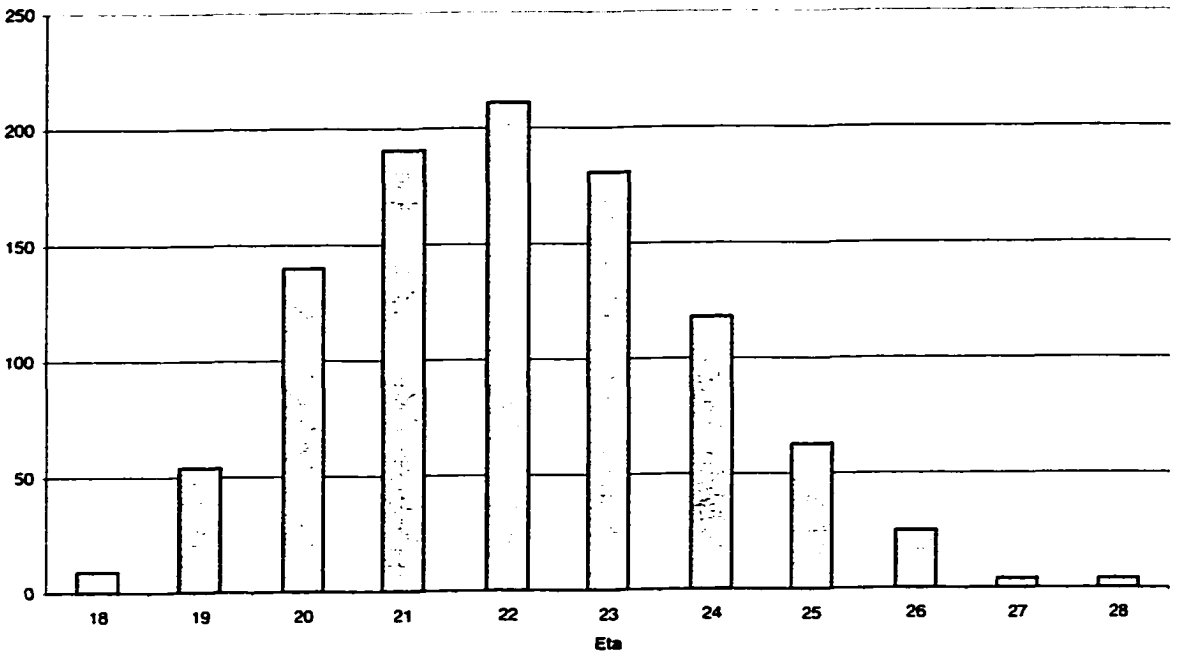



**Figure 35**
**Largest VIF, interaction, 7 IVs, moderate collinearity**

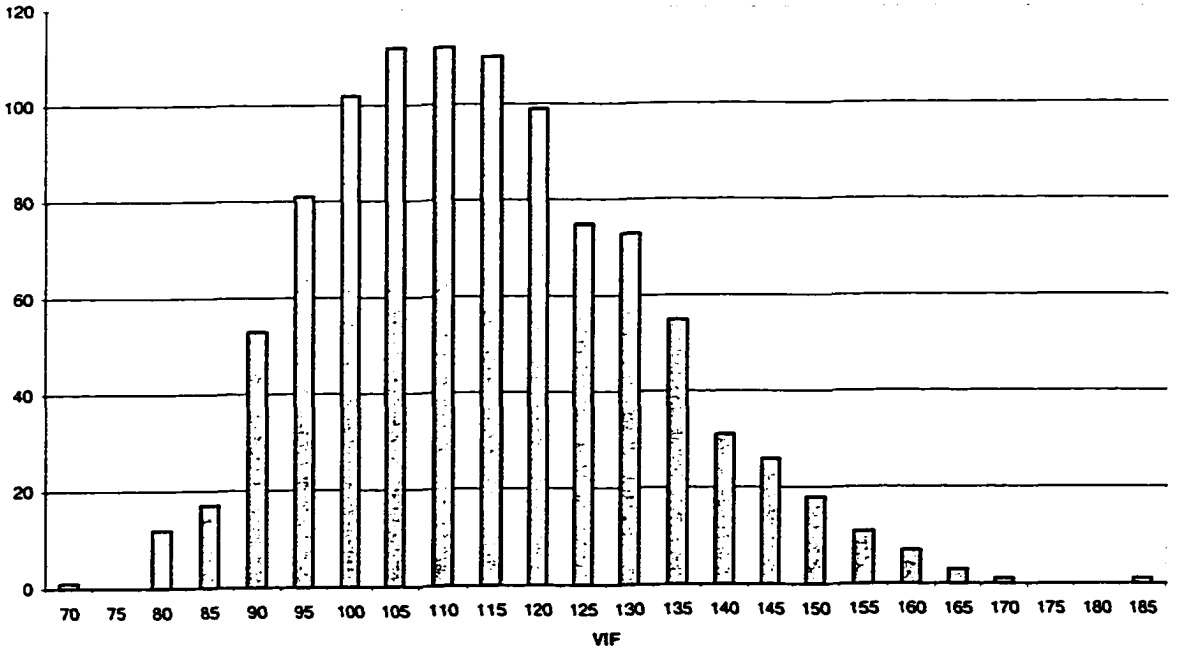


**Figure 36**
**Largest eta, interaction, 7 IVs, strong collinearity**

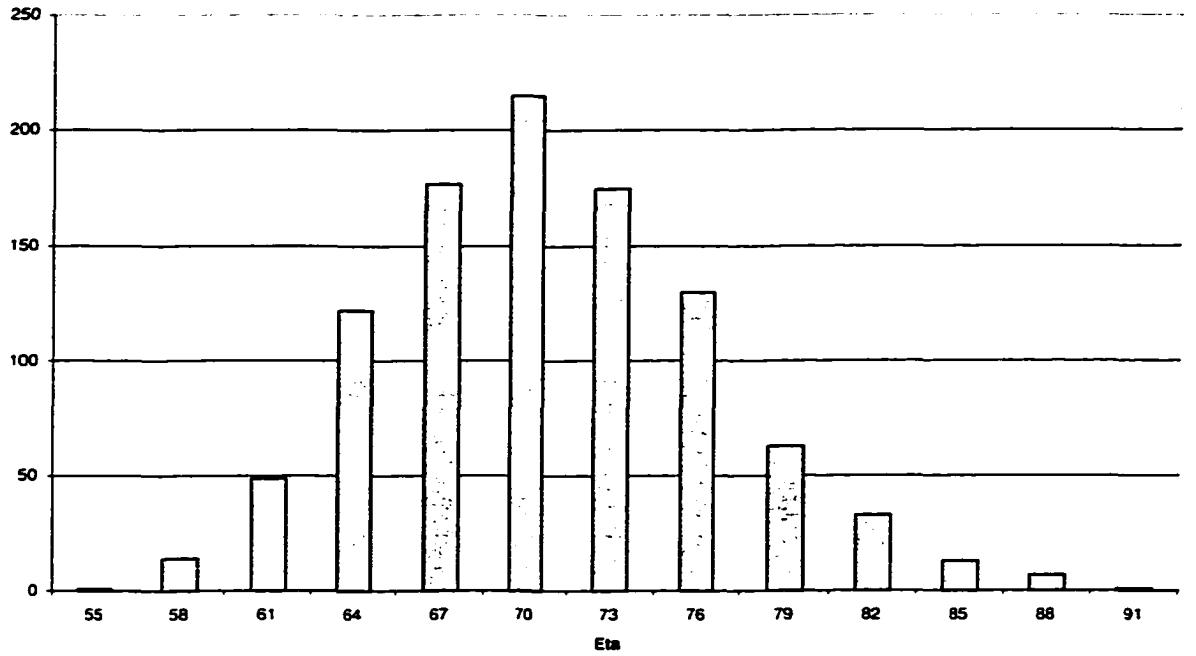
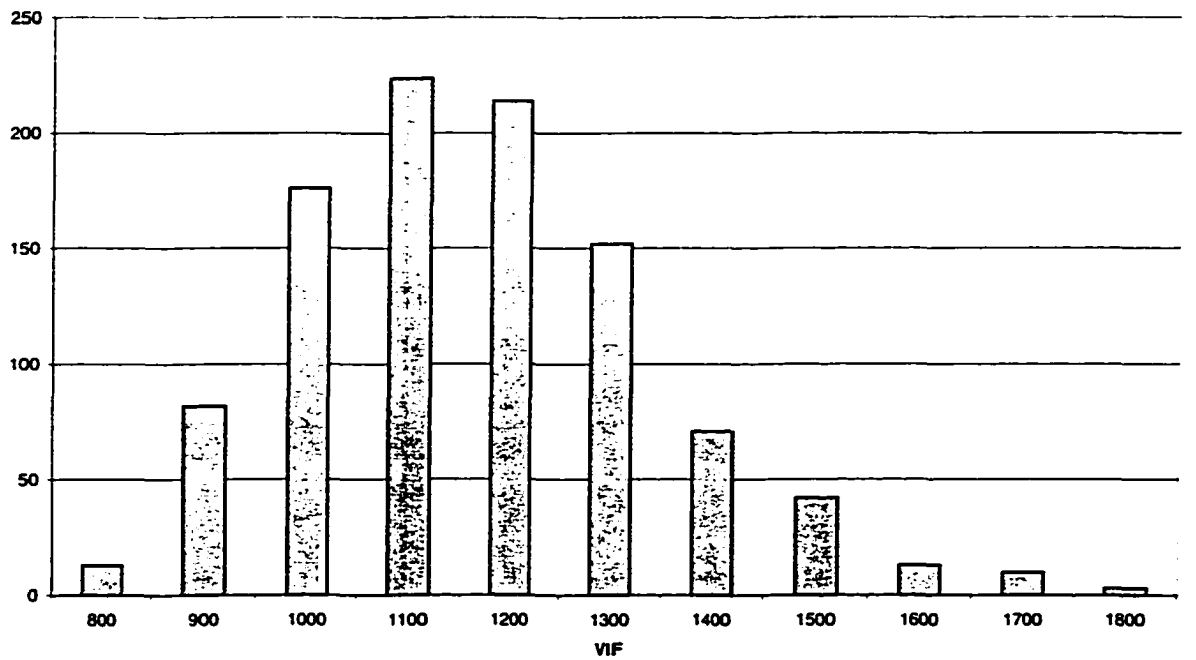



**Figure 37**
**Largest VIF, interaction, 7 IVs, strong collinearity**



Peter Leslie Flom

B. A. New York University

M. A. New York University

M. A. Fordham University

Multicollinearity Diagnostics for Multiple Regression: A Monte Carlo Study

Dissertation directed by John Walsh, Ph. D.

**Objectives:** To demonstrate the ineffectiveness of some commonly used collinearity diagnostics, and present a more effective method which was developed by David Belsley.

**Methods:** Collinearity is a problem in multiple regression when there is a relationship among the independent variables. When collinearity exists, parameter estimates may be incorrect, and their standard errors may be inflated. Some commonly used methods for diagnosing collinearity (e.g. the correlation matrix) are shown to inadequate. A Monte Carlo study of two widely recommended collinearity diagnostics (variance inflation factors and condition indexes) when three conditions were varied. These conditions were the number of IVs (3, 5 or 7), the presence (or absence) of interaction, and degree of collinearity (weak, moderate, or strong) are varied. This 3x2x3 factorial ANOVA design yielded 18 models, each of which was replicated 1000 times. Each replication was tested for collinearity using 2 diagnostics.

**Results:** Both VIFs and condition indexes were able to diagnose the presence of collinearity, but condition indexes were more precise. Condition indexes were able to determine which variables were involved in the collinearity, but VIFs were not. The

most commonly recommended values for diagnosing collinearity with VIFs appear to be substantially too low.

**Conclusions:** Data which are to be analyzed using multiple regression should be tested for collinearity using condition indexes.

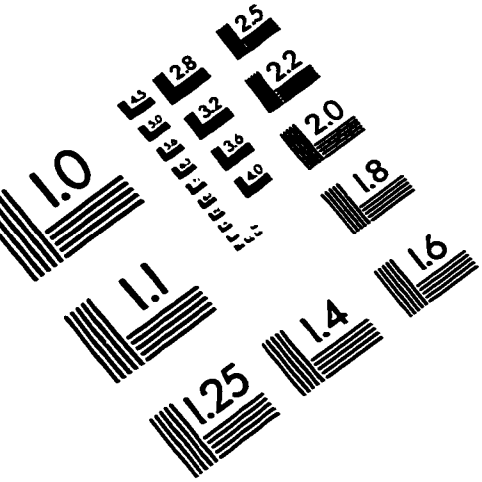
# VITA

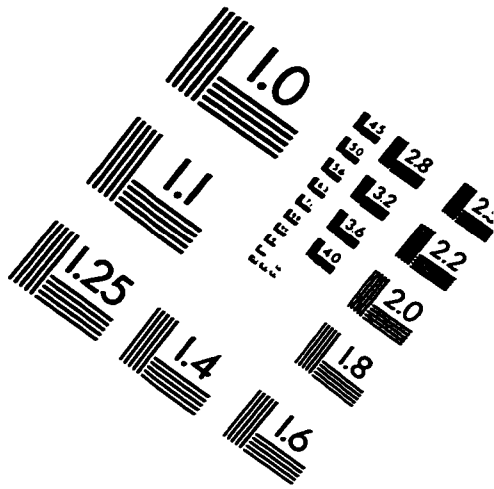
Peter L. Flom, son of Joseph Harold Flom and Claire Flom, was born on July 2, 1959, in New York City, New York. He attended York Preparatory School in New York City, and was graduated in June, 1976.

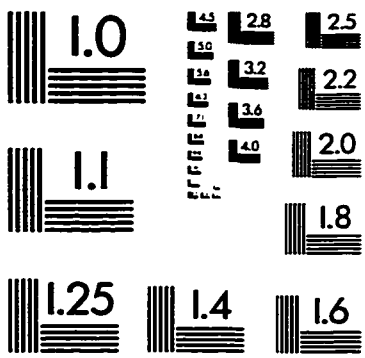
He entered New York University in September, 1977, and received the degree of Bachelor of Arts in February, 1980.

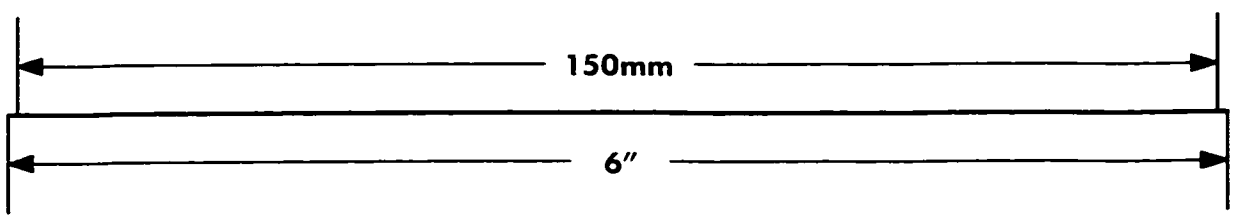
In September, 1991, he was accepted as a graduate student in the Graduate School of Arts and Sciences at Fordham University, where he majored in Psychometrics under the mentorship of Professor John Walsh. He was awarded the Richard Bennett Distinguished Fellowship and a Presidential Scholarship in the 1992, 1993, and 1994 academic years.

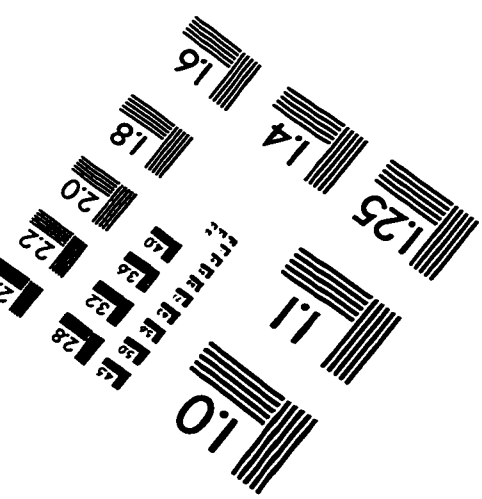
# IMAGE EVALUATION TEST TARGET (QA-3)

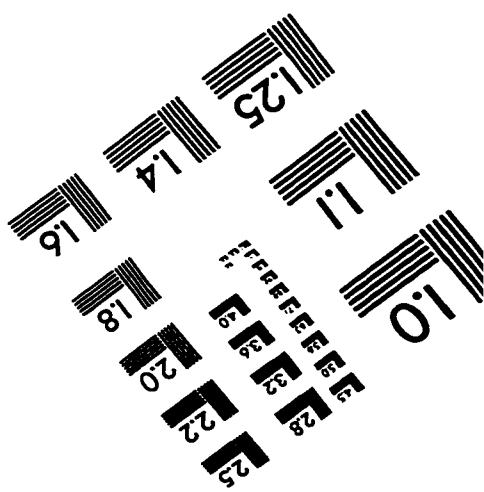



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989