LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Malte Wissmann & Helge Toutenburg & Shalabh

Role of Categorical Variables in Multicollinearity in the Linear Regression Model

Technical Report Number 008, 2007
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Role of Categorical Variables in Multicollinearity in Linear Regression Model

M. Wissmann[1], H. Toutenburg[2] and Shalabh[3]

Abstract

The present article discusses the role of categorical variable in the problem of multicollinearity in linear regression model. It exposes the diagnostic tool condition number to linear regression models with categorical explanatory variables and analyzes how the dummy variables and choice of reference category can affect the degree of multicollinearity. Such an effect is analyzed analytically as well as numerically through simulation and real data application.

Key Words: Linear regression model, multicollinearity, dummy variable, condition number

1 Introduction

The problem of multicollinearity has remain the center of attraction in the literature of linear regression analysis for a long time, see Silvey (1969). It arises when the explanatory variables in the linear regression model are correlated and thus one or more columns of the design matrix form a 'near' linear combination with other columns. The presence of multicollinearity in the data is a numerical issue as well

[1]WWZ Statistics and Econometrics, University of Basel, Petersgraben 51, 4003 Basel (Switzerland) E-Mail: malte.wissmann@unibas.ch

[2]Department of Statistics, University of Munich, Akademiestrasse 1, 80799 Munich (Germany) Email: Helge.Toutenburg@stat.uni-muenchen.de

[3]Department of Mathematics & Statistics, Indian Institute of Technology, Kanpur-208 016 (India) E-mail: shalab@iitk.ac.in, shalabh1@yahoo.com

as a statistical issue, see Silvey (1969), Belsley, Kuh and Welsch (2004) or Steward (1987) for more details. It is a statistical issue because it inflates the variance of ordinary least squares estimator and a numerical issue in the sense that the small errors in input may cause large errors in the output. The problem of multicollinearity has been attempted in the literature from different perspectives like as diagnostic tools, removal tools, estimation and testing of hypothesis of parameters. Various diagnostic tools like as condition number, singular value decomposition method, Belsleys condition indices, variance decomposition method, variance inflation factors, Belsleys perturbation analysis etc. have been suggested in the literature for the detection of multicollinearity and identification of variables causing the linear relationships, see Belsley (1991) and Rao, Toutenburg, Shalabh and Heumann (2008) for more details. The complete bibliography on multicollinearity is out of the objectives of this paper.

The condition indices are popular diagnostic tools for multicollinearity to detect the 'near' linear dependencies in data matrix. The condition indices are supplemented by the variance decomposition method. This has an advantage that it can detect the variables causing the 'near' linear dependency. On the other hand, the variance decomposition method assumes that the disturbances in linear regression model are homoscedastic.

The variance inflation factors are also used for diagnosing the multicollinearity, see e.g. Fox (1992). These measures are based on the fact that a centered and scaled design matrix is the correlation matrix of explanatory variables. The intercept term is then excluded while using this diagnostic. The homoscedastic variance of the estimate of j^{th} regression coefficient is then a function of multiple correlation from the regression of j^{th} column on all other columns of design matrix. The term around the multiple correlation is termed as variance inflation factor of the j^{th} regression coefficient. This diagnostic assumes homoscedastic errors and cannot find the variables involved in the 'near' linear dependency.

Use of categorical variable as explanatory variable is a popular strategy in regression analysis in many applications when the data is qualitative in nature. The tools of regression analysis are applied by indicating the categories of qualitative categorical variable through dummy variables. Use of dummy variables in regression analysis has its own advantages but the outcome and interpretation may not be exactly same as in the case of quantitative continuous explanatory variable. In particular, there are several issues related to the diagnostic measures for multicollinearity which are still unexplored or only partially explored in the literature when explanatory variables are dummy variables. For example, what happens to the diagnostic measures for multicollinearity when

- the explanatory variables are qualitative in nature and are represented by dummy variables; and
- the observations in design matrix are centered around their mean.

The problem whether the observations should be centered around their mean or not before applying the diagnostic tools for multicollinearity is an issue which is still not completely resolved. The discussion about this issue may be reviewed in Belsley (1984) which argues that the centering of observations around their mean is of no use when dealing with multicollinearity. It eliminates the intercept term from the linear regression model and therefore masks the role of intercept term on multicollinearity which is caused by it as well as by other variables, see also Belsley (1991). On the other hand, Marquardt (1980) states that the centering of observations removes nonessential ill conditioning. If the uncentered data is ill conditioned, then the small errors in inputs have large impact on the estimates of parameters. Belsley (1984) demonstrates that perturbed inputs have the same influence on the estimates obtained by using the centered and uncentered observations.

The issue of having an intercept term in linear regression model from the multi-

collinearity point of view is also an unsettled issue in the literature. It is often argued that the intercept term is of no use for the interpretation of regression results. But if the linear regression model has dummy variables then the intercept term represents the mean level of study variable at the reference or baseline categories of all categorical variables, when all other variables are set to be zero. So it has an interpretable feature as a baseline level of the study variable. When explanatory variables are dummy variables, then the aspect of centering of observation is not meaningful because then the centered dummy variables as well as their regression coefficients loose their interpretation.

In linear regression analysis, the dummy variables can also play an important role as a possible source for multicollinearity. The choice of reference category for a categorical variable may affect the degree of multicollinearity in the data.

Such issues have not yet been addressed in the literature to the best of our knowledge. There is one conference paper about multicollinearity and categorical data, see Hendrickx, Belzer, Grotenhuis and Lammers (2004). But this paper basically applies Belsleys perturbation analysis to a data set using a software that can deal with categorial variables, which the original Belsleys approach cannot do. There is some available literature about the variance inflation factors, see e.g. Steward (1987) for a numerical approach to variance inflation factors and Fox (1992), who considers the dummy variables in case of generalized variance inflation factors.

We have attempted in this direction and have tried to explore these issues. We have considered the condition number and variance decomposition as diagnostic tools for multicollinearity in this paper. We assume that some of the explanatory variables are categorical in nature and are represented by dummy variables. We have analytically and numerically analyzed the role of dummy variables and the choice of reference category in causing the multicollinearity. We find that the multicollinearity with dummy variables may be reduced by choosing the correct reference category.

The plan of the paper is as follows. In Section 2, we describe the model and diagnostics for multicollinearity. Several linear regression models with different combinations of qualitative and quantitative variables are considered. The tools of regression and multicollinearity analysis are exposed to these models. Their discussion and results are presented in Section 3. In the last Section 4, the effect of categorical variable is explored in two real data set and its findings are reported. Some conclusions are placed in Section 5.

2 The model and the diagnostics for multicollinearity:

Consider the following linear regression model:

$$y = X\beta + \varepsilon \tag{1}$$

where y is a $(n \times 1)$ vector of study variable, $(n \times p)$ design matrix X is of full rank and represents n observations on each of the $(p-1)$ explanatory variables and an intercept term with $(n \times 1)$ vector of elements unity, ε is a $(n \times 1)$ vector of disturbances which is a random variables with zero mean and positive definite covariance matrix Σ, and β is a $(p \times 1)$ vector of associated regression coefficients.

The ordinary least squares estimator of β from (1) is

$$b = (X'X)^{-1}X'y. \tag{2}$$

If we assume that the disturbances ε are homoscedastic, then the variance-covariance matrix of b is

$$Var(b) = \sigma_\varepsilon^2 (X'X)^{-1}, \tag{3}$$

where $\Sigma = \sigma_\varepsilon^2 I$.

In the linear regression model (1), we assume that some of the explanatory variables are categorical variables. We use the set up of dummy variables to model the categorial variables. A categorial variable with m categories is represented by $(m-1)$ dummy variables. The reference category or baseline category is denoted by r, which the analysts may choose freely ($r \; \varepsilon \; \{1, \ldots, m\}$), see Rao et al. (2008). A dummy variable D_k, $k \; \varepsilon \; m \backslash r$, is defined as

$$D_k = \begin{cases} 1 & \text{if in category } \; k \\ 0 & \text{else.} \end{cases} \tag{4}$$

Note that when dummy variables are used to represent the categorical explanatory variables, then an intercept term is needed in the model. Clearly the level of a study variable y at the reference category is where all dummy variables are zero. So the intercept term reflects this baseline level of y and is therefore necessary in the regression model.

Now we assume that the problem of multicollinearity is present in data where some of the explanatory variables are categorical in nature. We examine the role of dummy variables under the aspect of multicollinearity. We measure the multicollinearity in the design matrix with condition number following Belsley et al. (2004). A problem with the condition number is that it has its own scaling problems, see Steward (1987). When the dimensions of the data are changed, then the condition number is also changed. Belsley et al. (2004) recommends to scale each column of the design matrix using the Euclidian norm of each column before computing the condition number. The methods of Belsley et al. (2004) are implemented in the statistical software R using its package *perturb*. This package uses the root mean square of each column for scaling as its standard procedure.

The condition number of a matrix X is defined as

$$\kappa(X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{5}$$

where $\lambda_j, (j = 1, 2, \ldots, p)$ are the eigenvalues of $X'X$. If we use the singular value decompensation of X, then we can express the condition number in terms of singular values (μ) as

$$\kappa(X) = \frac{\mu_{max}}{\mu_{min}}, \tag{6}$$

where the singular values of X are the positive square root of eigenvalues of $X'X$, see Belsley et al. (2004) or Rao and Rao (1998) for further details on condition number, eigenvalue system and singular value decompensation.

Belsley et al. (2004) derived the threshold values for $\kappa(X)$ through simulation studies. These values are 10 and 30 which indicates a medium and serious degrees of multicollinearity, respectively. Based on $\kappa(X)$, Belsley et al. (2004) considered condition indices as

$$\eta_j \quad = \quad \frac{\mu_{max}}{\mu_j} \quad (j = 1, \ldots, p).$$

These indices provide a more detailed insight into the multicollinearity issues of a given matrix X. The number $\kappa(X)$ indicates whether a matrix is 'ill' conditioned or not. On the other hand, the numbers η_i provide information about the degree of involved 'near' linear dependencies. If, for example, two η_i's are greater then 30 then this indicates that there are two 'near' linear dependencies which may cause problems. A regression coefficient variance decompensation technique can then identify the involved variables in linear dependencies. By applying the singular value decomposition, $X = UDV'$ on the homoscedastic covariance matrix of (2), the variance of ordinary least squares estimator in (3) can be rewritten as

$$Var(b) = \sigma^2 V D^{-2} V'$$

where U and V are $(n \times p)$ and $(p \times p)$ orthogonal matrices, respectively and D is $(p \times p)$ diagonal matrix with nonnegative diagonal elements $\mu_1, \ldots, \mu_p$. This relates

the variance of j^{th} regression coefficient with the singular values of X as

$$Var(b_j) = \sigma^2 \sum_{k=1}^{p} \frac{\upsilon_{jk}^2}{\mu_k^2}. \tag{7}$$

When a 'near' linear dependency is present then the corresponding value of μ_k is very small, in comparison to other singular values. If a high proportion of the variance of two or more regression coefficients constitute the components of (7) aligned with a small singular value, then the corresponding variables are considered to be involved in causing the 'near' linear dependency. This is the idea behind the variance-decompositions proportions. Each proportion π_{kj} is the share of one component of (7) related with one singular value μ_k relative to the total variance (7). If a proportion is high, (say, > 0.5), then more than 50% of the variance of j^{th} coefficient is related to a small singular value μ_k. If another coefficient also has a high proportion related with μ_k and the corresponding condition index μ_{max}/μ_k is high, then a 'near' linear dependency is diagnosed. Let

$$\pi_{kj} = \frac{\phi_{jk}}{\phi_j}, \quad (j, k = 1, \ldots, p), \tag{8}$$

$$\phi_{jk} = \frac{\upsilon_{jk}^2}{\mu_k^2}, \tag{9}$$

$$\phi_j = \sum_{k=1}^{p} \phi_{jk}. \tag{10}$$

These proportions may then be summarized in a matrix Π with the condition indices on the first column and remaining p columns for the proportions. A 'near' linear dependency is present, when a condition index exceeds the threshold value. The variables involved are identified by proportions greater then 0.5 in Π.

3 Choice of reference category and multicollinearity

It may be noticed that the choice of reference category affects the values of explanatory variable. One of the objective of our study is to explore the influence of the choice of reference category of a categorical variable on the degree of multicollinearity and on the numerical stability of the estimates. We use the condition number of scaled design matrix to diagnose the multicollinearity.

To motivate and understand the issue, we first consider a simple situation of linear regression model

$$y = \beta_0 + \beta_1 D + \varepsilon \tag{11}$$

with an intercept term (β_0), slope parameter (β_1) and a dummy variable (D) representing a categorial variable with only two categories. Note that β_1 is interpreted as the difference in the expected values of y in different categories. Out of n number of observations, we observe h times the value '1' in the sample. We use the Euclidian norm to scale the design matrix following Belsley et al. (2004). The resulting scaling factors are ($\sqrt{n}$, $\sqrt{h}$). This provides a matrix where all the columns have same length. The cross-product of the scaled matrix $X_s'X_s$ is

$$X_s'X_s = \begin{pmatrix} 1 & \sqrt{f} \\ \sqrt{f} & 1 \end{pmatrix} \tag{12}$$

with $f = \frac{h}{n} \neq 0$ as the share of 'ones' in the sample. If $f = 0$, then we have a column with all elements 'zero' in X and the scaled design matrix does not exist in this case. Clearly a model with a 'zero' column does not makes any sense at all. If f is near to 1, then there exists a close relationship between the intercept term and dummy variable in the design matrix. The condition number, which is the ratio of square root of the maximum and minimum eigenvalues of $X_s'X_s$, should also reflect the same

fact. From the characteristic polynomial

$$(1-\lambda)^2 - f, \tag{13}$$

we obtain the eigenvalues of $X_s'X_s$ as $\lambda_{1,2} = 1 \pm \sqrt{f}$. The condition number is then

$$\kappa(X_s) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{1+\sqrt{f}}{1-\sqrt{f}}}, \tag{14}$$

which is a function of f. The graph of (14) is shown in figure 1. We observe from the

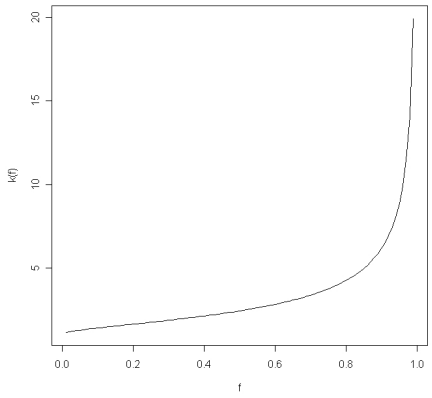

Figure 1: The condition number as a function of f

figure 1 that when f close to 0, then the condition number is close to 1 which indicates that the data is nearly orthogonal. Following Belsley et al. (2004), the condition number less then 10 indicates that there is no problem of multicollinearity at all. When $f = 0.96$, then the condition number hits the bench mark '10' and indicates the presence of a medium degree of multicollinearity. As f increases towards 1, the condition number tends towards infinity. Such an outcome is expected because then the dummy variable column is 'nearly' linearly depending on the column of intercept term in the design matrix.

From such illustration, we observe that the dummy variables can cause the multicollinearity problem. If only 4 percent of the observations are in the reference

category, then we observe a medium degree of multicollinearity in this model. It suggests further that the problem of multicollinearity can be avoided by choosing a different reference category of dummy variable. In such a case when the reference category is changed, then 96 percent of the observations lie in the reference category. The condition number in this case is around 1.2 which indicates no multicollinearity.

It is also clear that by changing the reference category, the standard errors of the corresponding regression coefficients are not changed. The homoscedastic variance of the least squares estimate of regression coefficient (b_1) of the model (1) under consideration is

$$Var(b_1) = \frac{n\sigma_\varepsilon^2}{(n-h)h} = \frac{\sigma_\varepsilon^2}{nf(1-f)} \tag{15}$$

which is symmetric due to the role of f.

However, the variance of the least squares estimate of intercept in (1) is affected. It is

$$Var(b_0) = \frac{\sigma_\varepsilon^2}{n-h} = \frac{\sigma_\varepsilon^2}{n(1-f)} \tag{16}$$

and is not symmetric like (15). If the share of 'ones' in the sample is large, then $(1-f)$ is small and the resulting variance in (16) is large. On the other hand, when $(1-f)$ is large, then the variance in (16) is small.

From such analysis in a simple case, we observe that the coding of dummy variable and the choice of reference category affects the numerical stability of design matrix as well as the variance of intercept term.

We also have obtained a closed form of the condition number in the presence of dummy variables in (14) for diagnosing the multicollinearity. This is a function of proportions of 'ones' and 'zeros'. It increases as the degree of multicollinearity increases due to the chosen reference category which affects the share of 'ones'. Such result is not reported in the literature on multicollinearity.

3.1 Interaction of one categorial variable and intercept term

In the previous subsection, we tried to have some insight on the aspect that how the choice of a reference category may affect the condition number of design matrix in a simple case with one dummy variable and an intercept term. To get more insight on this issue, we simulate some more models to see how the condition number behaves if the number of 'ones' are more than the number of 'zeros' in the system. We consider a model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon \tag{17}$$

with an intercept term (β_0), slope parameters (β_1, β_2) which contains one dummy variable (D) and a continuous quantitative explanatory variable (X_1). The observations on the dummy variable (D) are drawn from a Binomial distribution $B(1, p)$. If a smaller value of p is chosen, then we expect less number of 'ones', and as p increases the share of 'ones' increases. The continuous explanatory variable (X_1) is drawn from an exponential distribution exp(2.3). Then we compute the condition number of the system for different values of p. We simulated 10 such designs and the mean of conditional numbers is plotted against p in figure 2.

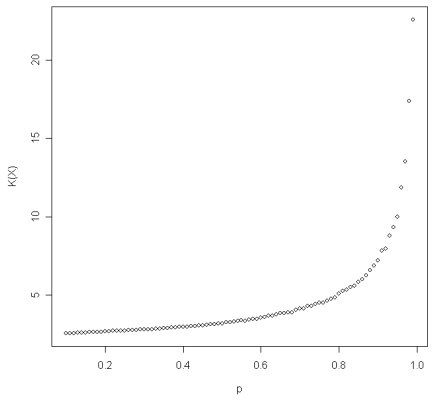

Figure 2: Conditional numbers vs. different values of p for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon$

We see an exponential relationship between the condition number of design matrix and value of p. The rate of increase of the condition number is moderate as long as the expected share of 'ones' in the sample is less then 0.8. As soon as p becomes larger than 0.8, we observe a higher rate of increment in the condition number with respect to p. We observe a medium multicollinearity problem when the value of p is around 0.95, which is quite similar to the results of the previous section.

Now we consider the simulation for the case of $p = 0.95$ in more detail. We draw 100 designs matrices for $p = 0.95$. We compute the condition number and present the results on its descriptive statistics in table 1. We observe that the mean and median

Minimum	1^{st} Quartile	Median	Mean	3^{rd} Quartile	Maximum	Standard deviation
9.048	9.738	10.15	10.22	10.69	11.69	0.643

Table 1: Descriptive statistics of condition numbers for $y = \beta_0 + \beta_1 X_1 + \beta_2 D + \varepsilon$

are close to 10. The deviation from the mean is relatively small. Next we present the box-plot of the results of table 1 in figure 3. We observe that the condition numbers

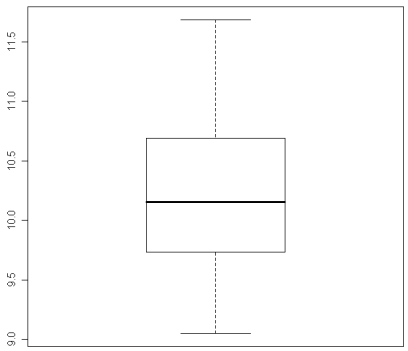

Figure 3: Boxplot of 100 condition numbers for $p = 0.95$

are relatively symmetrically distributed.

Now we explore the question that what happens if an additional category is added in the model (17)? To understand this issue, we add another dummy variable, representing the new category, in the model (17). So now we have a model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon \tag{18}$$

with an intercept term (β_0), slope parameters ($\beta_1, \beta_2, \beta_3$), one quantitative variable (X_1) and two dummy variables (D_1, D_2) representing the three categories denoted by '0', '1', and '2'. The categorical variable is drawn from the Binomial distribution $B(2, p)$. The first dummy variable D_1 takes the value '1' if the categorical variable is '1' and the second dummy variable D_2 is '1' if the categorical variable is '2'. The reference category is where the categorical variable is '0'. If we choose p similar to earlier cases, then we observe no differences with the case in (18) when only one dummy variable was considered in (17). The outcomes of the simulation are plotted in figure 4 which are very similar to the outcomes as in figure 2. This clearly shows that the nature of problem remains same even when the number of categories increase.

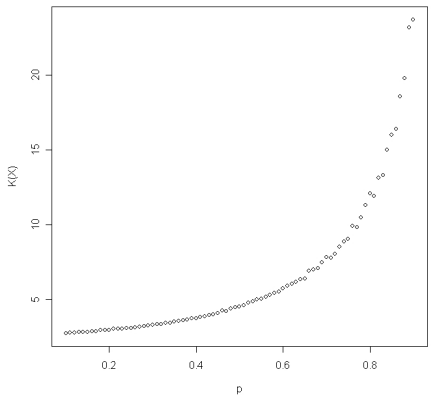

Figure 4: Conditional numbers vs. different values of p for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$

3.2 Interaction of more then one explanatory categorial variables and intercept Term

We now study a model with an intercept term and two explanatory categorical variables. The simulation model is equivalent to the model in section 3.1 but we add a second categorical variable. Both categorical variables (or factors A and B) have only two categories which are represented by two dummy variables D^A and D^B. More precisely, the model is now

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \varepsilon \tag{19}$$

with $D^A \sim B(1, p_1)$, $D^B \sim B(1, p_2 = 0.5)$ and $X_1 \sim \exp(2.3)$. Then we compare the case with only one dummy variable with this one to see the sensitivity of the condition number to this addition of dummy variable in (17). We consider three values of $p_1 = (0.7, 0.9, 0.95)$ based on the idea that $p_1 = 0.7$ corresponds to 'below multicollinearity' problem, $p_1 = 0.9$ corresponds to a multicollinearity problem at the border line and $p_1 = 0.95$ is in the region of a medium multicollinearity problem for the case with only one dummy. The results of condition numbers based on 1000 simulated design matrices for the three values of p_1 are presented in figures 5 and 6 in three and one graphics, respectively.

We observe that the presence of second dummy variable increases the condition number of design matrix. The explanation is as follows. The linear combination of the two dummy variable is more similar to the intercept term than one dummy variable alone. In case of $p_1 = 0.9$, we expect a share of 90% 'ones', the second dummy variable may then fill some of the gaps with additional 'ones'. This leads to an increase in the degree of multicollinearity between the dummy variables and intercept term. As p_1 increases, the rate of increase of condition number declines. This is because if p_1 is small, then the second dummy variable has a greater chance to fill the linear combination with another 'one'. As p_1 increases, this chance become smaller because

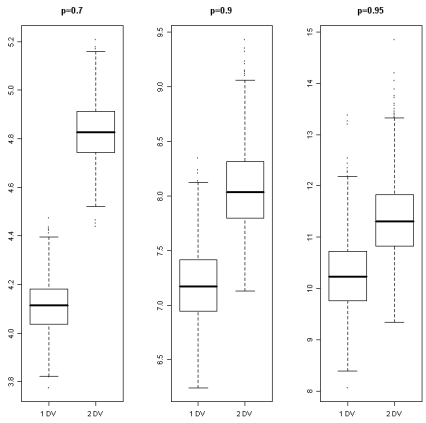

Figure 5: Results for the different values of p for $y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \varepsilon$

the first dummy variable contribute most of the 'ones' in linear combination.

We now add another categorical variable D^C to the model (19) to support this argument as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \beta_4 D^C + \varepsilon. \tag{20}$$

The additional dummy variable (D^C) is drawn from a Binomial distribution $B(1, 0.35)$, so we expect a smaller increase in the condition number since the third dummy variable has less 'ones' than the second dummy variable. The results are presented in figure 7. It compares the condition index of the cases with one, two and three dummy variables. Having a third dummy variable in the model increases the condition number of the design matrix. As the third dummy variable has a lower probability of success, so the increment in condition number is lower than the case of two dummy variables.

3.3 Interaction of two dependent categorial variables

Now we simulate a model in which the two categorical variables are modeled as dependent. An issue to be explored here is whether a dependent choice for the

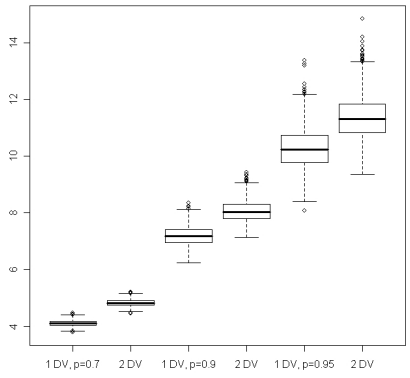

Figure 6: Results for the different values of p in one figure for $y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \varepsilon$

reference category can affect the stability of the design matrix or not.

The simulation set up is as earlier, a model with an intercept term, a quantitative continuous variable (X_1) and categorical variables represented by two dummy variables D^A and D^B as

$$y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \varepsilon. \tag{21}$$

We observe dummy variable D^A from a Binomial distribution $B(1, 0.5)$. Another dummy variable D^B is dependent on D^A and is modeled as follows. If a value of D^A is 1, then we draw the corresponding value of D^B from $B(1, p)$, else if a value of D^A is 0 we use $B(1, (1-p))$. Hence if the value of p is close to 1, then we expect a positive association between the two dummy variables. Such an example when $p = 0.9$ is as follows:

$D^A \backslash D^B$	'0'	'1'
'0'	462	45
'1'	53	440

If D^A has the value '1' we expect D^B to have the value '1' as well.

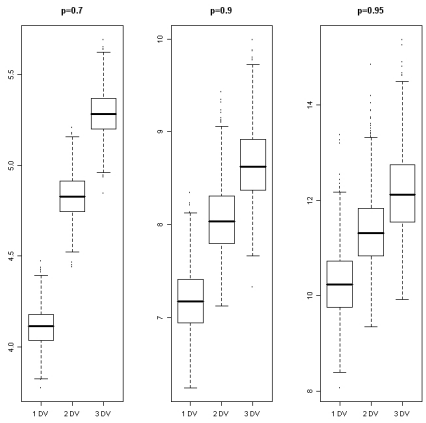

Figure 7: Results for 3 categorical variables for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \beta_4 D^C + \varepsilon$

If we choose the reference category for D^A different from D^B, then we have a negative association as follows:

$D^A \backslash D^B$	'0'	'1'
'0'	45	462
'1'	440	53

This is then simulated by using a small value of p and we choose $p = 0.1$. We simulate 100 experiments with $p = 0.1$ and $p = 0.9$, compute the condition numbers based on the outcomes from 100 experiments and then compare them together. In figure 8, we have described the results of the condition numbers for $p = 0.9$ and $p = 0.1$.

If we code the categorical variables such that they match in the 'ones', then we obtain smaller condition numbers compared to the case when they are coded in the other way. This means that a 'near' linear dependency with two dummy variables is less harmful in terms of the condition number than a 'near' linear dependency with two dummy variables and an intercept term.

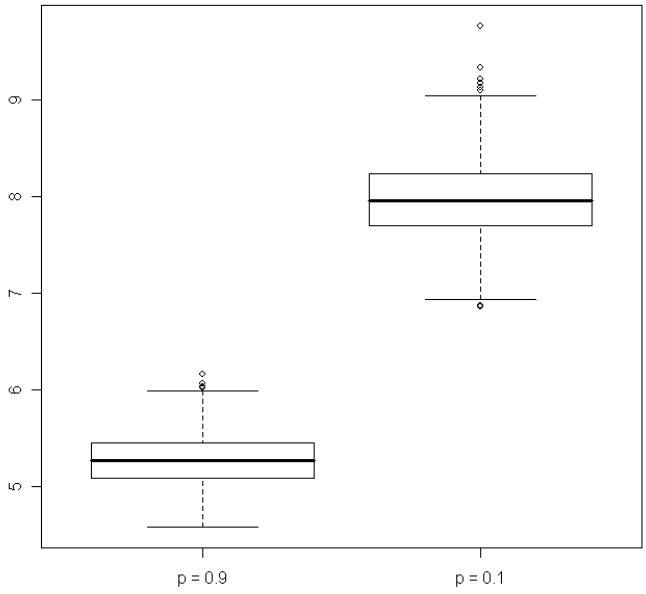

Figure 8: Dependent categorical variables under different types of coding the dummy variables for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 D^A + \beta_3 D^B + \beta_4 D^C + \varepsilon$

3.4 Influence of the choice of reference category on multicollinearity with two continuous variables

Next we consider a model where we have two continuous quantitative variables involved in a 'near' linear dependency and one dummy variable which may have a weak reference category. We term a category with low frequency as a weak reference category. For example, if the number of 'ones' are 95% in a category and only 5% 'zeros' in other category, then the category with '0' is called as a weak reference category.

The first continuous quantitative variable X_1 is drawn from an exponential distribution exp(2.3). The second continuous quantitative variable X_2 is a linear combination of X_1 and of a normal random variable with mean zero and standard deviation $0.085 s_{x1}$ where s_{x1}^2 is the sample variance of observations on X_1. The dummy variable

Coefficient:	Estimate	Standard error	$Pr(>\|t\|)$
Intercept	22.85	0.15	0
X_1	1.31	0.92	0.16
X_2	3.57	1.31	0.007
D	0.66	0.15	0

Table 2: Regression analysis output for $y = 23 + 1.5X_1 + 3X_2 + 0.5D + \varepsilon$

D takes two values '0' and '1'. The study variable y is generated by

$$y_i = 23 + 1.5X_{1i} + 3X_{2i} + 0.5D_i + \varepsilon_i, \quad (i = 1, \ldots, n) \tag{22}$$

where ε is a normally distributed random variable with zero mean and a standard deviation of 1.5. The sample size is 1000.

We then compare the situation with a weak reference, modeled by using $B(1, 0.9)$ for D, with the recoded model which is obtained by changing the reference category in (22).

The regression analysis output and Π matrix for the weak reference case are presented in tables 2 and 3, respectively. The deviations in the resulting estimates from the known β are quite large. The parameter b_1 is not significant at one percent level of significance. The inflated variances of the regression estimates turn a known significant parameter into an insignificant parameter. The related condition indices and variance decomposition factors are stated in table 3. We observe from table 3 that one condition index is just below the threshold for a medium degree of multicollinearity. This is associated with the intercept term and dummy variable. Then we have a condition index that is just above the threshold for a serious multicollinearity problem, which is associated with the 'near' linear dependency of X_1 and X_2. Next we explore what happens if we choose a different reference category. The results of regression analysis obtained after reverting the reference category are presented in table 4. The

Condition number	Intercept	Variance X_1	Decomposition X_2	Proportions D
1	0.007	0.001	0.001	0.008
2.43	0.037	0.003	0.003	0.056
7.91	0.956	0	0	0.937
30.3	0	0.996	0.996	0

Table 3: Multicollinearity analysis output for $y = 23 + 1.5X_1 + 3X_2 + 0.5D + \varepsilon$

Coefficient:	Estimate	Standard error	$Pr(>\|t\|)$
Intercept	23.01	0.07	0
X_1	1.31	0.92	0.16
X_2	3.57	1.31	0.007
D	-0.66	0.15	0

Table 4: Regression analysis output for $y = 23+1.5X_1+3X_2+0.5D+\varepsilon$ with changed reference category

Condition number	Intercept	Variance Decomposition Proportions X_1	X_2	D
1	0.045	0.001	0.001	0.021
1.75	0	0	0	0.895
2.821	0.955	0.003	0.003	0.084
27.38	0	0.996	0.996	0.001

Table 5: Multicollinearity analysis output of $y = 23 + 1.5X_1 + 3X_2 + 0.5D + \varepsilon$ with changed reference category

estimates are more precise now (see table 4). The standard error of the intercept term is lower and all other standard errors remain the same as in table 2. By changing the reference category, we change the sign of the parameter estimate associated with the dummy variable. In the earlier case, β_3 is $[E(y|D = 1) - E(y|D = 0)]$ and now the reference category is the opposite of the earlier case. Therefore the sign must be different. Note that β_3 is still $[E(y|D = 1) - E(y|D = 0)]$ but the interpretations of '0' and '1' are changed. The regression estimates associated with continuous quantitative variables are still inflated. The condition indices and variance decomposition proportions are presented in table 5. Now we have a condition index which is just below 30. The absence of the 'near' linear dependency of the dummy variable and intercept term therefore reduces the condition index. We have only problem now between X_1 and X_2. The additional dependency between D and intercept term is avoidable by the appropriate choice of the reference category.

4 Application to real data sets

Now we study the role of dummy variable and choice of reference category under multicollinearity in real data sets. We consider two real data sets and do the similar

analysis as in Section 3. The analysis and results are presented in the following subsections.

4.1 Example 1: Infant-Mortality

We consider a data set from Leinhardt and Wassermann (1979) which was used in Fox (1997) and is available in the R package 'car', R Development Core Team (2007).

It contain data on infant-mortality (y) per 1000 live births of 105 nations around the world. The influence factors are whether the country is an oil exporting country (factor A), the regions (Africa, America, Asia and Europe) (factor B) and the per-capita income (in U.S. Dollars) X.

First we consider the model with 2 categorical variables as:

$$y = \beta_0 + \beta_1 D^A + \beta_2 D_1^B + \beta_3 D_2^B + \beta_4 D_3^B + \beta_5 X + \varepsilon \tag{23}$$

Note that B has four categories and hence we need three dummy variables to represent them as D_1^B, D_2^B and D_3^B. The bar plot of the two categorial variables indicates whether there are weak categories or not. Weak categories are those classes which have low frequencies. They may give rise to multicollinearity with the intercept term if chosen as reference category. The regions seems to be well balanced and the relative frequencies of all the categories are larger then 15%.

There are not much oil exporting countries in the sample as figure 10 shows. Only 8% of the countries in the sample are oil exporting countries. There is no association between the region and an oil exporting country, only Europe has no oil exporting country in the sample.

If we choose the countries with oil export as a reference category, then we expect a 'near' dependency between the corresponding dummy variable and intercept term. Assuming homoscedastic errors, we obtain the regression results for (23) in table 6.

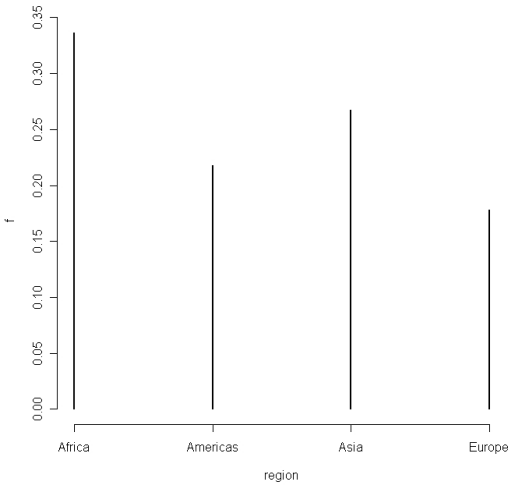

Figure 9: Barplot of the regions in infant-mortality data

The intercept term in this case is the mean mortality for a non oil exporting country in Africa with income zero. It gives a reference for the interpretation of dummy variables. From the output in table 6, we observe that exporting oil lowers the infant mortality. An African state which does not export oil has a higher infant mortality than a non-African state who does export oil, as the comparison of b_1 and intercept shows. It interprets that since an African state does not export oil, so they have low income and less profit. In turn, they spend less on the welfare of the people and hence they have higher infant mortality rate.

The variance decomposition proportions shows a medium degree of multicollinearity between the dummy variable for oil and intercept term. Only the last row is printed in table 7.

Now if we choose a different reference category for the oil exporting countries, then we have no problem at all and the variance of intercept term is more precise. Since the reference category for oil is different, so the interpretation of intercept term estimate is also different, see table 8. Table 8 shows that not exporting oil increases

Coefficients:	Estimate	Standard Error	$Pr(>\|t\|)$
Intercept	215.2	29.7	0
oil-yes	-78.3	28.9	0.01
region-America	-83.7	21.8	0
region-Asia	-45.9	20.1	0.02
region-Europe	-101.5	30.7	0.001
income	-0.005	0.007	0.48

Table 6: Regression analysis output of infant-mortality data

Condition-		Variance	Decomposition	Proportions		
number	Intercept	oil-yes	region-America	region-Asia	region-Europe	income
9.25	0.96	0.93	0.02	0.04	0	0.02

Table 7: Multicollinearity analysis output of infant-mortality data

Coefficients:	Estimate	Standard error	$Pr(>\|t\|)$
Intercept	136.8	13.6	0
oil-no	78.3	28.9	0.01
region-America	-83.7	21.8	0
region-Asia	-45.9	20.1	0.02
region-Europe	-101.5	30.7	0.001
income	-0.005	0.007	0.48

Table 8: Regression analysis output of infant-mortality data with changed reference category

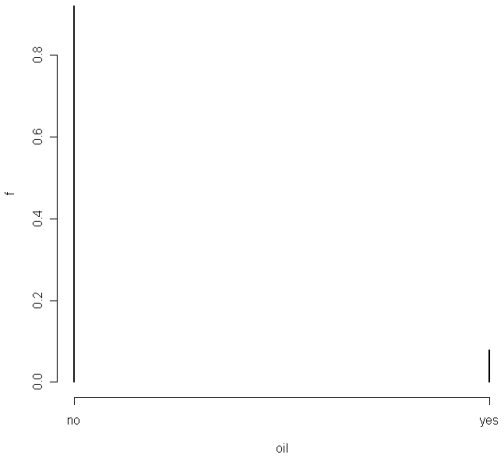

Figure 10: Barplot for the oil exporting countries in infant-mortality data

Condition		Variance	Decomposition	Proportions		
number	Intercept	oil-no	region-America	region-Asia	region-Europe	income
3.93	0.49	0.03	0.51	0.47	0.72	0.3

Table 9: Multicollinearity analysis output of infant-mortality data with changed reference category

the infant mortality relative to the baseline of oil exporting states, all other things equal. Note that the standard error of intercept term is less then half than the value of the standard error of intercept term in the model (11) with the different choice of reference category for oil as in table 6. The condition indices and variance decomposition proportions are stated in table 9. Only the last row is printed in table 9. We observe from table 9 that the choice of reference category lowers the condition index. Now there is no multicollinearity at all.

4.2 Example 2: Prestige of occupations

The data set is taken from Duncan (1961). The data was also used by Fox (1997) and is available in R. The data was collected in 1950. Here we study the relationship between categorial and continuous variables which causes multicollinearity. It illustrates how the choice of a weak category as a reference may affect the multicollinearity measures in this case.

The occupation prestige (y) was measured as the percent of raters in a NORC[1] study which rated the prestige of occupations as excellent or good. A categorical variable is used for the type of occupation with the values - professional and managerial ($prof$), white-collar class (wc) and blue-collar class (bc). So we have 3 categories which are represented by two dummy variables D_1 and D_2. The percentage of males in a occupation earning US$ 3500 or more was used as a measure for income (X_1) and the share of males in a occupation with high-school diploma as a measure for education (X_2).

The model for the prestige of occupations is

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X_1 + \beta_4 X_2 + \varepsilon. \tag{24}$$

First we analyze the frequencies of the occupation type from figure 11. Most people in the sample are in blue-collar class. Only 13.3% persons in the sample are in the white-collar class. Our recommended choice for a reference category is the blue-collar class which has a relative frequency of 46.7%.

Another problematic issue can be, that whether the occupation type is associated with income or education. For example, the persons in professional and managerial jobs earn a higher income than the persons in white-collar jobs. This is illustrated in figure 12. The group effect seems to be stronger for education than for income.

[1]National Opinion Research Center

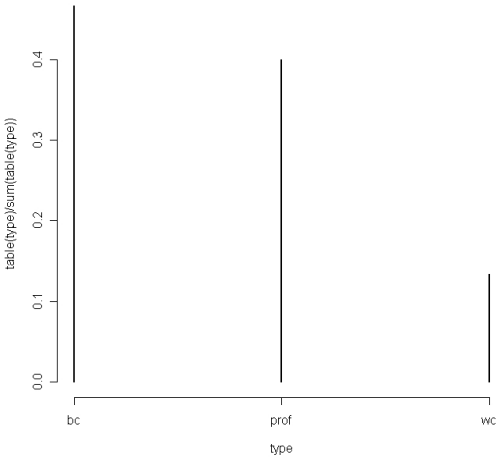

Figure 11: Barplot of the occupational types

Again we choose the naive reference category and take white-collar as reference category, because it has the lowest relative frequency from figure 11. The intercept term then is the mean prestige of a white-collar worker given all quantitative variables are zero. The results are presented in table 10 and we observe that having a professional or managerial job increases the prestige of the job more then having a blue-collar job. When we compare the prestige of the professional and blue-collared jobs, then the intercept term represents the mean prestige level for white-collar jobs with income and education equal to zero. In this case, we have the lowest prestige in the reference category (wc) followed by blue collared jobs and by professionals. This can be considered as an intuitive guideline to choose the reference category in this case. But as it turns out, this increases the degree of multicollinearity as well as the variance of intercept term.

Now consider the multicollinearity diagnostic for this data which is presented in table 11. We observe from table 11 that there is a multicollinearity between the intercept term and dummy variable for blue-collar occupation and also with the eduction. The intercept term together with the dummy variable and education form

Coefficients:	Estimate	Standard Error	$Pr(>\|t\|)$
Intercept	-14.8	8.18	0.07
type-*prof*	31.3	5.07	0
type-*bc*	14.66	6.11	0.02
income	0.6	0.09	0
education	0.34	0.11	0.004

Table 10: Regression analysis output of 'prestige of occupation' data with *wc* as reference category

Condition		Variance	Decomposition	Proportions	
number	Intercept	type-*prof*	type-*bc*	income	education
1	0.002	0.010	0.002	0.008	0.003
1.8	0.002	0.039	0.073	0.001	0.001
4.9	0	0.73	0.082	0.251	0.008
7.6	0.053	0.209	0.078	0.724	0.276
13.8	0.942	0.012	0.765	0.016	0.712

Table 11: Multicollinearity analysis output of 'prestige of occupation' data with *wc* as reference category

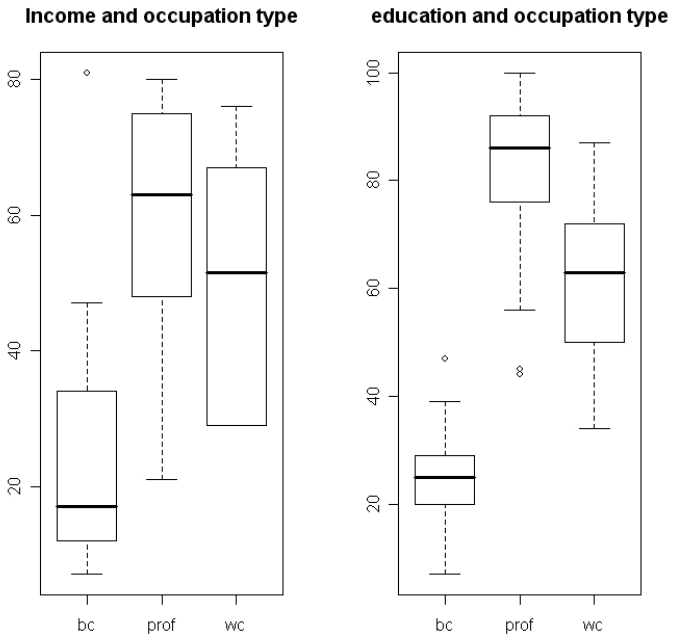

Figure 12: Boxplot of the income and eduction for the occupational types

a weak 'near' linear dependency.

We can get the intercept term out of that dependency if we use the blue-collared occupation as a reference category, since it has the highest frequency. The interpretation of intercept term is the mean prestige level for blue-collar jobs with income and education equal to zero. Here the baseline level of the prestige lies in the middle, white-collar jobs have lower prestige and professional jobs have higher prestige, all other things equal, see table 12. Again we get a more precise estimate for the intercept term and the standard error of intercept term is less then half of the value as from table 10. The results about the multicollinearity diagnostic for this case are presented in table 13. Now we observe from table 13 that the medium multicollinearity problem is reduced on the basic association between occupation type and education. The intercept term is now of no problem anymore. Thus, we expect the standard errors of regression estimator of type-$prof$ and education to be little inflated.

Coefficients:	Estimate	Standard error	$Pr(>\|t\|)$
Intercept	-0.19	3.71	0.96
type-*prof*	16.66	6.99	0.02
type-*wc*	-14.66	6.11	0.02
income	0.6	0.09	0.02
education	0.34	0.11	0.004

Table 12: Regression analysis output of 'prestige of occupation' data with *bc* as reference category

Condition		Variance	Decomposition	Proportions	
number	Intercept	type-*prof*	type-*wc*	income	education
1	0.01	0.006	0.006	0.008	0.003
1.9	0.001	0.022	0.335	0	0
3.6	0.31	0.125	0.178	0.001	0.002
6.5	0.209	0.093	0.075	0.977	0.032
10.828	0.471	0.754	0.406	0.014	0.963

Table 13: Multicollinearity analysis output of 'prestige of occupation' data with *bc* as reference category

The worst case of choice of reference category in this example is if we choose the professional occupations as reference category. The dummy variable for white-collared occupation and intercept term form a 'near' linear dependency with eduction. The condition index in this case is 15.

5 Summary

We have considered an issue related to the problem of multicollinearity in the presence of categorical variable as explanatory variable in the context of linear regression analysis. The role of dummy variables and the choice of reference category is analyzed through different linear models to see their effect in the problem of multicollinearity. In a simple case of one dummy variable, we have demonstrated that how the choice of reference category affects the multicollinearity. A closed form of condition number is also obtained in this case as a function of a collinearity increasing factor. It is difficult to get such a closed form expression in the general case as there can be many possible combinations of dummy and quantitative variables in linear regression models. So such an issue is explored more in detail by choosing various combinations of dummy and quantitative variables. It is found that the presence of dummy variable and the choice of reference category can be a cause of multicollinearity. Also, the situation of multicollinearity can be averted by changing the reference category. We have demonstrated this issue by simulation as well as through the application of two real data sets.

References

Belsley, D.A., "Demeaning Conditioning Diagnostics through Centering," *The American Statistican*, 1984, *38* (2), 73–77.

___ , *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, 1 ed., John Wiley & Sons, Inc. New York, 1991.

___ **, E. Kuh, and R.E. Welsch**, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, 1 ed., John Wiley & Sons, Inc. New York, 2004.

Duncan, O. D., "A socioeconomic index for all occupations," in A. J. Reiss, ed., *Occupation and Social Status*, Free Press, 1961.

Fox, J., "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association*, 1992, *87* (417), 178–183.

___ , *Applied Regression, Linear Models and Related Models*, Sage Publications, 1997.

Hendrickx, J., B. Belzer, M. Grotenhuis, and J. Lammers, "Collinearity involving ordered and unordered categorical variables," RC33 conference in Amsterdam August 2004. http://www.xs4all.nl/ jhckx/perturb/perturb.pdf.

Leinhardt, S. and S.S. Wassermann, "Explorattory data analysis: An introduction to selected methods," in K. Schssler, ed., *Sociological Methology*, Jossey-Bass, 1979.

Marquardt, D. W., "A Critique on some ridge regression Methods: Comment 'You should standardize the predictor vairables in your regression modells'," *Journal of the American Statistical Association*, 1980, *75* (369), 87–91.

R Development Core Team, *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing 2007. ISBN 3-900051-07-0.

Rao, C.R. and M.B. Rao, *Matrix Algebra and its Applications to Statistics and Econometrics*, 1 ed., World Scientific Publishing, Singapore, 1998.

___ , **H. Toutenburg, Shalabh, and C. Heumann**, *Linear Models and Generalizations - Least Squares and Alternatives*, 3rd ed., Springer, 2008.

Silvey, S. D., "Multicollinearity and Imprecise Estimation," *Journal of the Royal Statistical Society. Series B*, 1969, *31* (3), 539–552.

Steward, G.W., "Collinearity and Least Squares Regression," *Statistical Science*, 1987, *2* (1), 68–84.