Detección de Datos Influenciales y Multicolinealidad en el Modelo No Lineal

Juan Esteban Palacio Salazar

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2015

Detección de Datos Influenciales y Multicolinealidad en el Modelo No Lineal

Juan Esteban Palacio Salazar

Tesis presentada como requisito parcial para optar al título de:
Magister en Estadística

Director:
M.Sc Elkin Castaño Vélez

Línea de Investigación:
Modelamiento estadístico

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín, Colombia
2015

A mi familia

Resumen

El análisis de regresión es una herramienta ampliamente usada en el trabajo estadístico aplicado. Una de las características deseables respecto a un conjunto de observaciones es que cada uno de los elementos de este conjunto tenga un impacto aproximadamente igual dentro del proceso de estimación de un modelo estadístico. En el caso del modelo de regresión lineal múltiple, se espera que cada una de las filas de la matriz de diseño X y su respectiva variable respuesta y, tenga un impacto similar en la estimación del modelo. También es deseable que la matriz de datos X no presente problemas de multicolinealidad, es decir, que las estimaciones del modelo no pierdan precisión debido a relaciones lineales entre las columnas de la matriz de diseño o falta de variabilidad en las mismas. No obstante, cuando se trabaja con datos observacionales, esto difícilmente ocurre, puesto que los procesos de estimación suelen estar contaminados por datos extremos y, algunas veces, por presencia de multicolinealidad. En este trabajo, se analizan este tipo de problemas en el modelo no lineal, y el diagnóstico de éstos se realiza generalizando algunas metodologías que son utilizadas en el modelo lineal múltiple, utilizando en este caso, estimación por mínimos cuadrados no lineales. Finalmente, se desarrolla una interfaz gráfica en R donde son implementadas estas técnicas.

Palabras clave: Datos extremos, datos influenciales, multicolinealidad, mínimos cuadrados, modelo no lineal, interfaz gráfica.

Abstract

Regression analysis is a widely used tool in the statistical work applied. One of the desirable features over a set of observations is that each of the elements of this set has an impact approximately equal within the estimation process of a statistical model. In the case of multiple linear regression model, it is expected that each of the rows of the matrix design X and respective variable response y, have a similar impact in the model estimation. It is also desirable that the data matrix X collinearity issues not present. However, when working with observational data, this hardly occurs, since the estimation processes are often contaminated by outliers and, sometimes, by the presence of collinearity. In this paper, these problems are discussed in the nonlinear model, and diagnosis of these are done generalizing some methodologies that are used in the multiple linear model, using in this case, estimated by nonlinear least squares. Finally, a graphical interface in R where these techniques are implemented develops.

Key words: Outliers, influential data, collinearity, least squares, nonlinear model, graphical interface.

Contenido

. Resumen **vii**

1. Introducción **3**

2. Detección de Datos Extremos en el Modelo Lineal **6**
2.1. Conceptos Fundamentales 6
2.2. Eliminación de Una Sola Observación 7

3. Detección de Multicolinealidad en el Modelo Lineal **16**
3.1. Conceptos Fundamentales 16
3.2. Búsqueda Exploratoria de Multicolinealidad 17
3.3. Técnicas de Análisis Numérico en Detección de Multicolinealidad 19
3.4. Descomposición en Valores Singulares (SVD) de una Matriz 20
3.5. El Número de Condición 21
3.6. Índices de Condición 24
3.7. Descomposición de la Varianza de los Coeficientes de Regresión 25
3.8. Procedimiento Sugerido para Diagnosticar Multicolinealidad 28
3.9. Efectos Nocivos de la Multicolinealidad 29
3.10. Multicolinealidad Nociva vs Degradante 29

4. Estimación por Mínimos Cuadrados No Lineales **31**
4.1. Mínimos Cuadrados en Modelos No Lineales 32
4.2. Aproximación Lineal del Modelo No Lineal 33
4.3. Algunos Algoritmos Numéricos de Optimización 36
4.3.1. Algoritmo de Newton - Raphson 36
4.3.2. Algoritmo de Gauss - Newton 38

5. Detección de Datos Extremos y Multicolinealidad en el Modelo No Lineal **40**
5.1. Detección de Datos Extremos en el Modelo No Lineal 40
5.2. Detección de Multicolinealidad en el Modelo No Lineal 42
5.3. Aplicación 43
5.3.1. Detección de Datos Extremos en el Modelo No Lineal 43
5.3.2. Detección de Multicolinealidad en el Modelo No Lineal 47

5.4. Un Análisis de la Capacidad de la Metodología para Detectar Presencia de Datos Extremos o Relaciones de Dependencia Lineal en los Datos Originales 64
5.4.1. Detección de Observaciones Extremas en los Datos Originales 64
5.4.2. Detección de Dependencias Lineales en los Datos Originales 72

6. Conclusiones **79**

A. Anexo: Interfaz Gráfica en R **81**

Lista de Figuras

5-1. Comportamiento por pares de TN vs (NIN, TW) 44
5-2. Comportamiento Univariado de Respuesta Ultrasónica y Distancia del Metal 49
5-3. Gráfico de Dispersión de la Respuesta Ultrasónica vs Distancia del Metal . . 50
5-4. Matriz de Dispersión Variables Pseudo Regresoras 53
5-5. Comportamiento Univariado de Longitud de Raíz y Concentración 57
5-6. Longitud de Raíz vs Concentración 58
5-7. Matriz de Dispersión Variables Pseudo Regresoras 61
5-8. Matriz de Dispersión Variables Pseudo Regresoras 71

A-1. Aplicación en blanco 82
A-2. Tabla de datos y resumen descriptivo 83
A-3. Modelo no lineal y valores iniciales de parámetros a estimar 84
A-4. Estimación no lineal, variable respuesta transformada y pseudoregresores . . 85
A-5. Detección Datos Extremos 86
A-6. Detección Datos Extremos 87
A-7. Detección Datos Extremos 87
A-8. Detección Multicolinealidad 88

Lista de Tablas

3-1. Índices de condición y proporciones asociadas de varianza para cada parámetro estimado. 27

5-1. Indicadores para detección de datos atípicos en estimación de modelo no lineal para datos Stromberg. Las observaciones extremas son marcadas en negrita . 46
5-2. Primeros quince datos de Chwirut2 con respuesta ultrasónica y distancia del metal. 48
5-3. Medidas descriptivas para respuesta ultrasónica y distancia del metal. 48
5-4. Parámetros estimados y sus errores estándar asociados en el modelo no lineal para respuesta ultrasónica y distancia del metal. 49
5-5. Tabla de variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal. 51
5-6. Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal. 52
5-7. Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal. 53
5-8. Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal. 54
5-9. Primeros doce datos de longitud de raíz y concentración. 55
5-10.Medidas descriptivas para longitud de raíz y concentración. 56
5-11.Parámetros estimados y sus errores estándar asociados en el modelo no lineal para longitud de raíz y concentración. 56
5-12.Tabla de variables pseudo regresoras utilizadas en la linealización de modelo no lineal para longitud de raíz y concentración. 59
5-13.Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para longitud de raíz y concentración. 59
5-14.Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para longitud de raíz y concentración. 60
5-15.Índices de condición y descomposición de varianza para parámetros estimados en la linealización de modelo no lineal para longitud de raíz y concentración. 62
5-16.Parámetros estimados y sus errores estándar asociados en el modelo no lineal para longitud de raíz y concentración. 63

5-17.Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal alterno para longitud de raíz y concentración. 63
5-18.Medidas descriptivas Stromberg. 65
5-19.Observaciones extremas en datos originales. 66
5-20.Observaciones extremas (incluido dato artificial a 3 sd del vector de medias). 69
5-21.Primeros quince datos de Chwirut2 con respuesta ultrasónica, distancia del metal y distancia del metal doble. 74
5-22.Medidas descriptivas para respuesta ultrasónica, distancia del metal y distancia del metal doble. 74
5-23.Parámetros estimados y sus errores estándar asociados en el modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble. . 75
5-24.Tabla de variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble. 76
5-25.Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble. 76
5-26.Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble. 77
5-27.Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble. 77

1. Introducción

Los modelos de regresión lineal se usan frecuentemente en el análisis de datos de diversas áreas del conocimiento. En este tipo de estudios, es habitual encontrar observaciones con una influencia desproporcionada sobre los resultados del ajuste de un modelo lineal. La presencia de este tipo de características en los datos puede causar graves problemas en las estimaciones, que probablemente conduzcan a resultados inadecuados en los análisis. Debido a esto, han surgido una serie de procedimientos cuyo objetivo central se concentra en detectar la presencia de una o más observaciones altamente influyentes en la estimación mínimo cuadrática del vector de parámetros y su matriz de varianzas y covarianzas, así como del vector de predicciones, de los residuales y de otras medidas de interés (Cook, R.D., 1977).

Otro problema que puede causar daños en las estimaciones mínimo cuadráticas de un modelo lineal es la presencia de multicolinealidad, definida ésta como la pérdida de precisión en la estimación debida a una poca variabilidad observada en las variables regresoras o a la existencia de relaciones lineales entre ellas. Este problema produce una inflación artificial en la varianza de los coeficientes estimados del modelo, lo cual conlleva, desde luego, a la pérdida de precisión en los mismos. En el caso del modelo lineal, se han desarrollado, entre otros, procedimientos basados en análisis numérico, los cuales tienen como enfoque central determinar si existen o no problemas de multicolinealidad y, adicionalmente, en caso de que existan, indicar qué variables se ven afectadas por los mismos. (Belsley, Kuh and Welsch, 1982).

Por otra parte, uno de los supuestos asumidos en un modelo de regresión lineal es que la verdadera relación existente entre la variable dependiente y las variables independientes es de tipo lineal. Sin embargo, en muchas aplicaciones, asumir esta linealidad es bastante restrictivo, pues el verdadero tipo de relación, en muchos casos, tiene un comportamiento mejor explicado por una relación de tipo no lineal (Novales, 1993). Por tanto, en algunas situaciones particulares, resulta una práctica útil y adecuada la implementación de modelos que consideren relaciones de dependencia no lineales.

Teniendo en cuenta que al estimar un modelo de regresión no lineal mediante mínimos cuadrados no lineales, el algoritmo internamente utiliza una expansión de Taylor sobre la función no lineal para obtener una aproximación lineal de ésta, con la cual se estiman por mínimos cuadrados ordinarios los mismos parámetros, pero tomando como insumos una nueva va-

riable respuesta (obtenida a partir de ciertas transformaciones sobre la variable respuesta original) y una nueva matriz de diseño (construida a partir de las derivadas parciales de la función no lineal respecto a cada uno de los parámetros de interés), los problemas causados por datos extremos y/o multicolinealidad se presentan, en este caso, al estimar por mínimos cuadrados ordinarios esta forma lineal y por consiguiente, dichos problemas deben tenerse en cuenta y deben ser evaluados tomando como insumos esta nueva respuesta transformada y esta nueva matriz de diseño. En consecuencia, es posible extender, del modelo lineal al no lineal, las medidas de diagnóstico, tanto de datos extremos como de multicolinealidad.

En el capítulo 2 de este trabajo se presentan algunas de las medidas propuestas por Belsley, Kuh and Welsch (1982), las cuales, basadas en la eliminación de una única observación a la vez, apuntan a detectar la presencia de datos extremos en un conjunto de observaciones usadas para ajustar y estimar un modelo lineal. Asimismo, se presentan, para cada uno de los indicadores, los puntos de corte sugeridos por dichos autores a partir de los cuales una observación puede considerarse extrema.

El capítulo 3 presenta un recuento sobre el problema de multicolinealidad expuesto por Belsley, Kuh and Welsch (1982) en el modelo lineal. Esto incluye una definición básica del problema hecha por los autores, algunas de sus posibles causalidades, sus incidencias, algunos comentarios a tener en cuenta y un procedimiento propuesto por ellos que, basado en técnicas de análisis numérico sobre la matriz de diseño, tiene como objetivo central detectar la presencia de problemas de multicolinealidad en dicha matriz y adicionalmente, evaluar qué columnas de la misma provocan el problema y cómo éste finalmente afecta la precisión de los coeficientes estimados. Adicionalmente, se incluye una discusión que plantean los autores sobre el tipo de multicolinealidad que puede tenerse, siendo ésta nociva o degradante.

En el capítulo 4 se muestra de qué manera una función no lineal puede ser aproximada, mediante expansión de Taylor, a una de tipo lineal, en la cual los parámetros a estimar son exactamente los mismos que se tienen en el modelo no lineal original. En este capítulo se muestra de una forma detallada cómo se obtienen los insumos (variable respuesta y matriz de diseño) utilizados en la estimación de la aproximación lineal obtenida, los cuales son conseguidos a través de transformaciones sobre la variable respuesta inicial y derivadas parciales de la función original respecto a cada parámetro a estimar.

El capítulo 5, utilizando los resultados del capítulo 4, extiende las herramientas propuestas por Belsley, Kuh and Welsch (1982) para detectar datos extremos y problemas de multicolinealidad en el modelo lineal a modelos de tipo no lineal. Este capítulo incluye algunos apuntes importantes que deben tenerse en cuenta a la hora de entender qué papel juegan la presencia de datos extremos y/o problemas de multicolinealidad en la estimación de un modelo no lineal a través de mínimos cuadrados no lineales.

Este capítulo tiene un enfoque también práctico. En este capítulo se toman algunos ejemplos presentes en la literatura estadística donde se tiene un conjunto de datos reales y un modelo no lineal cuyo objeto es ser ajustado y estimado. Sobre la estimación por mínimos cuadrados ordinarios de la aproximación lineal obtenida es sobre la cual se realizan finalmente los diagnósticos concernientes a presencia de datos extremos y/o problemas de multicolinealidad. Estos diagnósticos se desarrollan siguiendo cuidadosamente la metodología propuesta en este trabajo. Por consiguiente, en cada caso se presentan los datos originales y una breve descripción de los mismos, la forma matemática del modelo no lineal, el proceso mediante el cual se obtienen los insumos necesarios utilizados por la aproximación lineal, el modelo estimado obtenido mediante dicha aproximación y, según sea el objetivo del ejemplo, el procedimiento de detección de datos extremos junto con sus respectivos análisis o la evaluación de multicolinealidad junto con las conclusiones correspondientes.

2. Detección de Datos Extremos en el Modelo Lineal

2.1. Conceptos Fundamentales

Las observaciones extremas o influenciales corresponden a aquellos subconjuntos de datos que parecen tener una influencia desproporcionada en la estimación de un modelo estadístico. Uno de los objetivos principales cuando se emprende la búsqueda de este tipo de observaciones es, además de detectarlas, determinar qué partes del modelo estadístico resultan más afectadas por la presencia de ellas. Los métodos utilizados en los análisis deben tener en cuenta tanto la variable respuesta (dependiente) como las variables explicativas (independientes). No usar ambos insumos puede llevar a la no detección de observaciones extremas multivariadas (Belsley, Kuh and Welsch, 1982).

Belsley Kuh and Welsch (1982) citan las más frecuentes fuentes de las cuales pueden provenir las observaciones extremas. En primer lugar, el registro incorrecto de información, ya sea en su origen o en su transcripción al computador, es una causa potencial de presencia de datos extremos. En segundo lugar, los errores observacionales, que pueden estar naturalmente ligados a la estructura de los datos. En tercer lugar, los datos atípicos pueden llevar a ocurrencias legítimas de valores extremos. Estos datos a menudo contienen información valiosa que mejora la eficiencia de la estimación. Incluso en esta situación beneficiosa, esos autores sugieren como una práctica estadística deseable aislar estos puntos extremos y cuantificar en qué medida las estimaciones de los parámetros son vulnerables a estas observaciones. Finalmente, una cuarta causa está relacionada con los datos podrían estar siendo generados por otro modelo diferente al especificado, lo cual puede conllevar a catalogar como atípica una observación que en realidad no lo sea.

El hecho de que un subconjunto pequeño de los datos pueda tener una influencia desproporcionada en los parámetros estimados o en las predicciones, es de interés para los analistas de regresión porque, en este caso, es muy posible que las estimaciones del modelo estén basadas principalmente en este subconjunto de datos y no en la gran mayoría (Kuh and Welsch, 1977).

Como una etapa previa a la estimación de una regresión múltiple, Belsley, Kuh and Welsch (1982) recomiendan examinar las distribuciones univariadas de cada variable con el fin de

detectar eventos extraños. Los diagramas de dispersión también pueden ser analizados para detectar observaciones extrañas hasta en dos dimensiones. Si bien hay claros beneficios al tratar las observaciones peculiares de esta manera, estos autores enfatizan en que diagnósticos de este tipo no pueden detectar observaciones extremas multivariadas, ni pueden determinar en qué manera estos datos influyen en el modelo estimado.

Después de estimar la regresión múltiple, la mayoría de los procedimientos de diagnóstico se basan en los residuales, en los valores ajustados (predichos), y en las variables explicativas. Pese a que estos métodos proporcionan información importante acerca de la calidad y validez de los resultados del modelo, su punto débil consiste en que no logran mostrar directamente cuál sería el modelo estimado si un subconjunto de los datos fuese modificado o no tenido en cuenta. Incluso si se detectan observaciones sospechosas a través de estos métodos, ellos no pueden cuantificar en qué medida su presencia ha afectado los coeficientes estimados, los errores estándar, y las pruebas estadísticas. Las técnicas presentadas por Belsley, Kuh and Welsch (1982) para diagnosticar datos influenciales se enfocan precisamente en suplir algunas de estas limitaciones. Estas técnicas, basadas en la eliminación de una única observación a la vez, son usadas en este trabajo como herramienta para detectar datos extremos en un modelo lineal.

2.2. Eliminación de Una Sola Observación

Las medidas de diagnóstico usadas en esta sección, propuestas por Besley, Kuh and Welsch (1982), tienen como objetivo final examinar cómo la supresión de una sola fila afecta los coeficientes estimados, los valores predichos (ajustados), los residuales y la estructura de covarianza estimada de los coeficientes. Estas cuatro salidas en un proceso de estimación son, desde luego, muy conocidas por los analistas de regresión múltiple y proporcionan un núcleo básico de las herramientas de diagnóstico.

Cada observación, por supuesto, está estrechamente relacionada con una única fila de la matriz de datos X y el elemento correspondiente de y. Una observación influyente es la que, de forma individual o en conjunto con otras observaciones, tiene un impacto demostrablemente más grande en los valores calculados de diversas estimaciones (coeficientes, errores típicos, valores t, etc) que la mayoría de las demás observaciones (Belsley, Kuh and Welsch, 1982). Una de las medidas intuitivas para examinar tal impacto es eliminar cada fila, una a la vez, y tener en cuenta el efecto resultante sobre los diversos valores calculados. La fila cuya eliminación produzca cambios relativamente grandes en los valores calculados, se considera influyente. De esta forma, siguiendo el trabajo de Belsley, Kuh and Welsch (1982), se presenta el procedimiento de eliminación de una fila, mirando a su vez el impacto que esto tiene en los coeficientes estimados y en los valores predichos (ajustados) $\hat{y}$, los residuales

$\epsilon = y - \hat{y}$, y los parámetros estimados en la matriz de varianzas y covarianzas.

Eliminación

Coeficientes y Valores Ajustados

Dado que los coeficientes estimados son de interés fundamental para los analistas de los modelos de regresión, inicialmente se analiza el cambio producido en éstos al suprimir la i-ésima fila de los datos. Si $\hat{\beta}$ representa el vector de parámetros estimados calculado con todos los datos; y $\hat{\beta}_{(i)}$ representa al mismo vector, omitiendo en la estimación la fila i; este cambio se calcula como:

$$\text{DFBETA}_i = \hat{\beta} - \hat{\beta}_{(i)} = \frac{(X^T X)^{-1} x_i^T \epsilon_i}{1 - h_i}$$

Donde

$$h_i = x_i (X^T X)^{-1} x_i^T, \quad i = 1, 2, 3, ..., n$$

$$\epsilon_i = y_i - \hat{y}_i, \quad i = 1, 2, 3, ..., n$$

Donde x_i es el vector fila que contiene los valores de la observación i en cada una de las variables consideradas en la matriz X. La cantidad h_i, discutida más adelante, se utiliza con frecuencia en las medidas de diagnóstico utilizadas en este trabajo.

Para evaluar si el cambio en $\hat{\beta}_j$, la componente j de $\hat{\beta}$, que resulta de la eliminación de la fila i, es grande o pequeño, es útil evaluar la varianza de $\hat{\beta}_j$, la cual está dada por $\hat{\sigma}^2 (X^T X)^{-1}_{jj}$, donde $(X^T X)^{-1}_{jj}$ es el elemento de la diagonal en la que se cruzan la fila j y la columna j de la inversa de la matriz $X^T X$.

Si se hace $C = (X^T X)^{-1} X^T$, entonces se tiene

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{c_{ji} \epsilon_i}{1 - h_i}, \quad i = 1, 2, 3, ..., n; j = 1, 2, 3, ..., p$$

donde p es la dimensión del vector β y c_{ji} es el elemento ubicado en la fila j y columna i de la matriz C, definida previamente.

Además, en Mosteller and Tukey (1977), se muestra que

$$\text{var}(\hat{\beta}_j) = \hat{\sigma}^2 \sum_{i=1}^{n} c_{ji}^2, \quad j = 1, 2, 3, ..., p$$

De esta forma, la medida de cambio escalada puede ser definida como

$$\text{DFBETAS}_{ij} \equiv \frac{\widehat{\beta}_j - \widehat{\beta}_{j(i)}}{s_{(i)}\sqrt{(X^T X)^{-1}_{jj}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \frac{\epsilon_i}{s_{(i)}(1 - h_i)}, \quad i = 1, 2, 3, ..., n; j = 1, 2, 3, ..., p$$

Donde se reemplaza s^2, la estimación usual de σ^2, por

$$s_{(i)}^2 = \frac{1}{n - p - 1} \sum_{k \neq i} [y_k - x_k \hat{\beta}_{(i)}]^2, \quad i = 1, 2, 3, ..., n$$

siendo p, nuevamente, la dimensión del vector β, es decir, el número de parámetros del modelo.

Se puede mostrar que la posibilidad de encontrar un DFBETA grande se reduce en proporción directa al aumento del tamaño de la muestra, es decir, la eliminación de una observación debe tener un menor efecto si el tamaño de la muestra se hace más grande.

Grandes valores en $|\text{DFBETAS}_{ij}|$ indican observaciones que son influyentes en la determinación del j-ésimo coeficiente $\hat{\beta}_j$.

Belsley, Kuh and Welsch (1982) sugieren marcar como influyentes aquellas observaciones con DFBETAS asociados cuyo valor absoluto sea mayor a $2/\sqrt{n}$.

Otra forma de resumir los cambios en los coeficientes y, simultáneamente, comprender mejor los efectos en la predicción cuando se elimina una observación, es con el cambio en el ajuste, definido como

$$\text{DFFIT}_i = \hat{y}_i - \hat{y}_{i(i)} = x_i[\hat{\beta} - \hat{\beta}_{(i)}] = \frac{h_i \epsilon_i}{1 - h_i}, \quad i = 1, 2, 3, ..., n$$

Para efectos de escala, es natural dividir entre $\hat{\sigma}\sqrt{h_i}$, lo cual corresponde a la desviación estándar del valor ajustado, $\hat{y}_i$. Así

$$\text{DFFITS}_i = \sqrt{\left|\frac{h_i}{1 - h_i}\right|} \frac{\epsilon_i}{s_{(i)}\sqrt{1 - h_i}}, \quad i = 1, 2, 3, ..., n$$

Donde s^2 es la estimación de σ^2.

Belsley, Kuh and Welsch (1982) sugieren marcar como influyentes aquellas observaciones con DFFITS asociado cuyo valor absoluto sea mayor a $2\sqrt{p/n}$.

Es importante tener cuidado al tratar con las observaciones influyentes. Una observación de este tipo, después de ser detectada, puede ser legítimamente eliminada únicamente en el caso en que se demuestre que ésta es consecuencia de un error (Belsley, Kuh and Welsch,1982). Generalmente, no se justifica la acción y, cuando lo es, es apropiado tomar una acción menos radical que la eliminación definitiva como por ejemplo investigar más profundamente las causas que la produjeron y realizar un plan que disminuya los posibles efectos adversos que la presencia de dicha observación puede provocar.

Matriz de Proyección

En los indicadores usados como medidas de diagnóstico, h_i y ϵ_i son componentes fundamentales. Los h_i son los elementos de la diagonal de la matriz de proyección de mínimos cuadrados, también conocida como la matriz sombrero

$$H = X(X^T X)^{-1} X^T,$$

los cuales determinan los valores predichos, ya que

$$\hat{y} = X\hat{\beta} = Hy.$$

La influencia de la i-ésima respuesta y_i en el ajuste está directamente reflejada en su impacto en el correspondiente valor ajustado $\hat{y}_i$, y esta información, como se observa en la ecuación anterior, está contenida en h_i. Adicionalmente, se puede probar que los elementos h_i están relacionados con la distancia entre x_i y su media muestral, $\overline{x}$ (Rao, 1973).

Cuando hay dos o menos variables explicativas, los gráficos de dispersión revelan rápidamente los valores extremos, y no es difícil comprobar que éstos tienen grandes valores h_i asociados. Cuando $p > 2$, los gráficos de dispersión no pueden revelar outliers multivariados, los cuales están separados de la mayor parte de los puntos x pero no aparecen como valores extremos en gráficos univariados ni bivariados. Dado que los elementos de la matriz H se pueden interpretar como distancias, éstos proporcionan un punto de partida básico para revelar la presencia de outliers multivariados. Los elementos de la matriz de proyección, entonces, son herramientas de diagnóstico por sí mismos, además de ser componentes fundamentales de otras herramientas (Belsley, Kuh and Welsch, 1982).

Belsley, Kuh and Welsch (1982) prueban que

$$0 \leq h_i \leq 1, \quad i = 1, 2, 3, ..., n$$

Además, puesto que X es de rango completo,

$$\sum_{i=1}^{n} h_i = p$$

El tamaño promedio de un elemento de la diagonal es p/n. Cuando se diseña un experimento, es deseable usar datos con una influencia relativamente igual, es decir, que cada observación tenga un h_i asociado cercano al promedio p/n. Sin embargo, dada la variación natural de los datos, es necesario tener algún criterio para decidir cuándo un h_i es lo suficientemente diferente a p/n para justificar atención especial (Belsley, Kuh and Welsch, 1982).

Bajo el supuesto de que las variables explicativas son independientes y están distribuidas normalmente, es posible calcular la distribución exacta de ciertas funciones de h_i. Por ejemplo, Belsley, Kuh and Welsch (1982) muestran que la cantidad $(n-p)\left[h_i - \left(\frac{1}{n}\right)\right]/(1-h_i)(p-1)$ tiene una distribución F con $p-1$ y $n-p$ grados de libertad. Para $p > 10$ y $n-p > 50$, el percentil 95 para F es menor que 2 y por lo tanto, $2p/n$ es un buen punto de corte. De esta forma, la influencia de la i-ésima observación se considerará grande si su h_i asociado es mayor a $2p/n$. Estos resultados se deben utilizar únicamente con fines exploratorios, teniendo en cuenta que la independencia y el supuesto de normalidad no son válidos en la práctica (Belsley, Kuh and Welsch, 1982).

Residuales

El uso de los residuales de la regresión, $\epsilon_i = y_i - \hat{y}_i$, en un contexto de diagnóstico, es amplio. Entre sus diversos usos, los residuales se pueden utilizar para revelar aquellos datos sospechosos que afectan indebidamente los resultados de la regresión. Los residuales también se pueden usar para detectar presencia de heterocedasticidad y autocorrelación, y además, pueden proporcionar la base para mitigar estos problemas. Adicionalmente, también se suelen usar para probar la normalidad aproximada en el término de error (Belsley, Kuh and Welsch, 1982).

Para realizar diagnósticos a través de los residuales, es conveniente que éstos sean modificados de manera que se mejore su capacidad de detectar datos problemáticos. En Theil (1971) se prueba que:

$$\text{var}(\epsilon_i) = \sigma^2(1-h_i)$$

En consecuencia, con el fin de tener un indicador insensible a las unidades de medición utilizadas, es es natural sugerir que, en lugar de estudiar ϵ_i, se deben estudiar los residuales estandarizados, los cuales están dados por

$$\epsilon_{si} = \frac{\epsilon_i}{s\sqrt{1-h_i}}$$

Sin embargo, al reemplazar la estimación de σ con $s_{(i)}$, se obtiene un residual estudentizado (RSTUDENT)

$$\epsilon_i^* = \frac{\epsilon_i}{s_{(i)}\sqrt{1-h_i}}$$

el cual, en muchas situaciones prácticas, se relaciona estrechamente con la distribución t con $n-p-1$ grados de libertad. De esta forma, si el supuesto de normalidad es válido, se puede evaluar fácilmente la importancia de cada residual estudentizado por separado. Dado que cada ϵ_i^* está estrechamente ligado con h_i y, a su vez, los h_i no son independientes entre sí (recuerde $\sum_{i=1}^n h_i = p$), los ϵ_i^* tampoco lo son (Belsley, Kuh and Welsch, 1982).

Distancia de Cook

Esta medida, utilizada para detectar la presencia de observaciones influyentes en un conjunto de datos, tiene en cuenta tanto la localización del punto como la variable respuesta medida en él. La distancia de Cook para la i-ésima observación está dada por

$$d_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{ps^2} \leq 1, \quad i = 1, 2, 3, ..., n$$

donde $\hat{y}_{(i)}$ es el vector de predicciones obtenido al eliminar la observación i.

Cook and Weisberg (1982) prueban que esta distancia también se puede escribir como

$$d_i = \left[\frac{\epsilon_i}{\sqrt{s^2(1-h_i)}}\right]^2 \frac{h_i}{p(1-h_i)} \leq 1, \quad i = 1, 2, 3, ..., n$$

Observe que el primer factor de esta expresión es el residual estandarizado asociado a la respectiva observación.

Cook and Weisberg (1982) sugieren prestar especial atención a aquellas observaciones cuya distancia asociada sea mayor a 1.

Matriz de Covarianza

En este punto se compara la matriz de covarianza calculada con todos los datos, con la matriz de covarianza que resulta al eliminar la i-ésima fila. Entre los diversos métodos existentes para comparar dos matrices simétricas definidas positivas, calcular la razón de sus determinantes es uno de los más simples. Dado que estas dos matrices difieren sólo por la inclusión de la i-ésima fila en la suma de cuadrados y productos cruzados, valores del

cociente de determinantes cercanos a la unidad se pueden interpretar como señal de que las dos matrices de covarianza están cerca, o que la matriz de covarianza original es insensible a la eliminación de la fila i. Este análisis se basa sólo en la información de la matriz X e ignora el hecho de que el estimador s^2 de σ^2 también cambia con la eliminación de la i-ésima observación (Belsley, Kuh and Welsch, 1982). Esto último se puede incorporar comparando las matrices $s^2(X^TX)^{-1}$ y $s^2_{(i)}(X^T_{(i)}X_{(i)})^{-1}$, mediante la razón de sus determinantes.

$$\text{COVRATIO} = \frac{det\left\{s^2_{(i)}(X^T_{(i)}X_{(i)})^{-1}\right\}}{det\left\{s^2(X^TX)^{-1}\right\}}$$

$$\text{COVRATIO} = \frac{s^{2p}_{(i)}det\left\{(X^T_{(i)}X_{(i)})^{-1}\right\}}{s^{2p}det\left\{(X^TX)^{-1}\right\}}$$

Belsley, Kuh and Welsch (1982) prueban que:

$$\frac{det\left\{(X^T_{(i)}X_{(i)})^{-1}\right\}}{det\left\{(X^TX)^{-1}\right\}} = \frac{1}{1-h_i}$$

De esta forma, se puede deducir que

$$\text{COVRATIO} = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{\epsilon^{*2}_i}{n-p}\right]^p (1-h_i)}$$

Se tiene interés en aquellas observaciones que conlleven a valores de COVRATIO que no estén cerca de la unidad, ya que estas observaciones son posiblemente influyentes y merecen mayor atención.

Belsley, Kuh and Welsch (1982) proponen investigar aquellos puntos con $|\text{COVRATIO} - 1|$ cercanos o mayores a $3p/n$

Los indicadores presentados hasta aquí sólo consideran la eliminación de una única observación al tiempo. Conjuntos de datos en los cuales se presenten casos de enmascaramiento,

entendido éste como la existencia de dos o más observaciones atípicas similares entre sí, pueden pasar inadvertidos por estas técnicas, ya que al eliminar una de estas observaciones extremas, las estimaciones no variarán significativamente debido a la presencia de las demás.

Existen extensiones de estos procedimientos que consideran la eliminación de dos o más observaciones simultáneamente, a fin de detectar datos extremos enmascarados. Algunas de estas técnicas, adicionalmente, agrupan las observaciones a eliminar de manera óptima, de tal forma que si, por ejemplo, se desean eliminar tres observaciones al tiempo, no haya necesidad de analizar todas las posibles tripletas de datos.

Rousseeuw and Van Zomeren (1990) plantean una estrategia para detectar datos extremos multivariados en presencia de enmascaramiento. Estos autores proponen la distancia de Mahalanobis como una medida de detección de datos atípicos multivariados. Esta distancia mide qué tan lejos está una observación del centro de los datos (vector de medias) teniendo en cuenta la escala de los mismos (matriz de covarianza). No obstante, esta medida no logra advertir el problema de enmascaramiento, lo cual hace que algunos datos extremos no tengan necesariamente una distancia grande calculada. Esto ocurre debido al hecho de que estos parámetros de localización y escala no son robustos: un pequeño grupo de datos extremos puede atraer estas medidas en su dirección. La solución propuesta por los autores a este problema consiste en calcular la distancia tomando como insumos de localización y escala unas medidas más robustas que el vector de medias y la matriz de covarianzas, las cuales se basan en el mínimo volumen de un elipsoide introducido por Rousseeuw (1985).

En dicho trabajo, Rousseeuw and Van Zomeren también exponen una metodología cuyo propósito es detectar observaciones influyentes con problemas de enmascaramiento en la estimación de un modelo de regresión. Ellos proponen estimar una regresión donde se minimice la mediana de la suma de cuadrados del error. Posteriormente diseñan un gráfico donde se trazan los residuales estandarizados de esta regresión contra la distancia robusta mencionada en el párrafo anterior. Este gráfico es dividido en sectores, lo cuales permiten clasificar las observaciones en regulares, influenciales benignas, influenciales malignas y atípicas.

El estudio de estos métodos se deja como tarea recomendada para el lector o como posibles trabajos futuros, puesto que los alcances de este trabajo no cubren este tipo de problemas.

3. Detección de Multicolinealidad en el Modelo Lineal

3.1. Conceptos Fundamentales

No existe en la literatura una definición única de multicolinealidad. Literalmente, dos variables presentan multicolinealidad si los vectores de datos que las representan se ubican en la misma dirección (subespacio de dimensión uno). Más generalmente, existe multicolinealidad en un conjunto de k variables si los vectores que las representan conforman un subespacio de dimensión menor a k, es decir, si alguno de los vectores es una combinación lineal de otros. En la práctica, la multicolinealidad exacta (correlación perfecta entre variables explicativas) rara vez ocurre, lo cual no significa, sin embargo, que el problema no exista (Belsley, Kuh and Welsch, 1982).

Menos estrictamente, dos variables presentan multicolinealidad si se encuentran casi en la misma dirección, es decir, si el ángulo entre ellas es pequeño. La presencia de multicolinealidad entre dos variables es equivalente a la existencia de una alta correlación entre ellas. Este concepto se puede generalizar a más de dos variables diciendo que existe multicolinealidad si hay un alto coeficiente de determinación cuando una de las variables explicativas es utilizada como variable respuesta en una regresión donde las restantes variables se utilizan como regresoreas (Belsley, Kuh and Welsch, 1982).

Nótese que la definición dada hasta ahora sobre multicolinealidad está basada en características específicas de la matriz de datos X y no en aspectos estadísticos del modelo de regresión lineal $y = X\beta + \epsilon$. Es decir, la multicolinealidad es un problema originado en los datos, no un problema estadístico como tal. La multicolinealidad, entonces, es un problema estadístico que, sin embargo, como se probará más adelante, tiene una gran importancia en la precisión de la estimación (Belsley, Kuh and Welsch, 1982).

Intuitivamente, se puede entender el efecto nocivo resultante de tener datos con presencia de multicolinealidad al tener en cuenta el hecho de que variables colineales no aportan información muy diferente a la ya inherente en las otras variables. Se hace difícil, por tanto, cuantificar separadamente la influencia de dichas variables explicativas en la respuesta (Belsley, Kuh and Welsch, 1982).

Adicionalmente, los efectos nocivos que produce la multicolinealidad pueden presentarse, en muchos casos, debido a un escaso número de observaciones o a que las variables explicativas tenidas en cuenta en el estudio varían sobre un rango no muy amplio, es decir, presentan un coeficiente de variación pequeño.

La varianza de los coeficientes estimados se puede escribir como:

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}, \quad j = 1, 2, 3, ..., p$$

Donde R_j^2 es el coeficiente de determinación de la regresión de x_j sobre las demás variables explicativas. Esta ecuación muestra que mientras más fuerte sea la relación lineal entre x_j y las demás variables explicativas, más precisión pierden los estimadores de mínimos cuadrados, y que a menor variación de x_j, también habrá menos precisión en estas estimaciones.

De esta forma, la capacidad para diagnosticar multicolinealidad es fundamental por los efectos que esta tiene en la estimación por mínimos cuadrados de un modelo de regresión. Este diagnóstico, siguiendo a Belsley, Kuh and Welsch (1982), consta de dos elementos relacionados pero distintos: en primer lugar, detectar la presencia de dependencias lineales (multicolinealidad) entre las series de datos, y, adicionalmente, evaluar la medida en la cual estas relaciones afectan la precisión de los parámetros estimados.

Con el fin de abordar estos dos elementos en este trabajo, se utilizan algunos indicadores numéricos expuestos por Belsley, Kuh and Welsch (1982), cuyas magnitudes dan indicios sobre la presencia o no de una o más relaciones de multicolinealidad dentro de las columnas de la matriz de datos X; luego se proporcionan los medios para determinar, dentro del modelo de regresión lineal, el grado en que cada una de estas dependencias afecta la estimación de mínimos cuadrados de cada uno de los coeficientes de regresión. En la mayoría de los casos esta información permite determinar específicamente qué columnas de la matriz de datos están involucradas en qué relaciones de dependencia lineal, es decir, aislar las variables involucradas y determinar las relaciones en las que ellas participan.

3.2. Búsqueda Exploratoria de Multicolinealidad

Una de las primeras tareas emprendidas cuando se tiene como propósito realizar diagnósticos sobre presencia de problemas de multicolinealidad incluye, entre otras revisiones, analizar si

los signos asociados a las estimaciones son coherentes con lo esperado, buscar si se presentan variables explicativas consideradas *importantes* cuyo estadístico t sea pequeño, probar si las estimaciones de la regresión son demasiado sensibles a la eliminación de una fila de X. Cualquiera de estas condiciones se cita con frecuencia como indicio de multicolinealidad y, peor aún, la multicolinealidad se cita comúnmente como una explicación a estas condiciones. Desafortunadamente, ninguna de estas condiciones es necesaria ni suficiente para la existencia de multicolinealidad, necesitándose, de esta forma, técnicas más elaboradas, tanto para detectar la presencia de multicolinealidad como para evaluar sus posibles efectos dañinos (Belsley, Kuh and Welsch, 1982).

Otra acción exploratoria que no debe pasar inadvertida consiste en examinar la matriz de correlación $\mathbf{R}$ de las variables explicativas. Examinar esta matriz es un procedimiento comúnmente empleado. Mientras que un alto coeficiente de correlación entre dos variables explicativas puede suponer un posible problema de multicolinealidad, la ausencia de altas correlaciones no puede verse como evidencia de la no existencia del problema. Es posible que tres o más variables hagan parte de una única dependencia lineal, aunque un par de éstas no esté altamente correlacionado. La matriz de correlación es incapaz de diagnosticar una situación de este tipo (Belsley, Kuh and Welsch, 1982). Esta matriz tampoco permite detectar la existencia simultánea de varias relaciones de dependencia lineal entre las variables explicativas. Los procedimientos presentados por Belsley, Kuh and Welsch (1982) permiten descubrir dependencias por separado así como proporcionar información referente a la participación de cada variable en cada tipo de dependencia.

Otra medida indicadora de multicolinealidad la constituyen los factores de inflación de varianza, VIF_j, los cuales se calculan para cada variable regresora de la siguiente manera:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, 3, ..., p$$

Donde R_j^2 es el coeficiente de determinación de x_j regresada en las restantes variables explicativas. Un alto valor para VIF_j significa un R_j^2 cercano a la unidad, lo cual, como se citó previamente, es un indicio de multicolinealidad. Por tanto, esta medida es de alta utilidad como indicador general de multicolinealidad. Sus puntos débiles, como los de la matriz de correlación, se encuentran en su incapacidad para distinguir diferentes dependencias y en la falta de un límite significativo para distinguir entre los valores de VIF que pueden ser considerados altos y aquellos que no deben serlo (Belsley, Kuh and Welsch, 1982).

En resumen, ninguno de los enfoques anteriores es completamente efectivo para diagnosticar problemas de multicolinealidad o evaluar el daño causado por éstos. Una solución más ri-

gurosa al problema debe buscarse en otros lugares: Belsley, Kuh and Welsch (1982) presentan una metodología que tiene que ver directamente con las propiedades numéricas de la matriz de datos X, haciendo cálculos basados en la inestabilidad o sensibilidad de ésta. De esta forma, la base para un diagnóstico adecuado se fundamenta utilizando diversos conceptos desarrollados en el campo del análisis numérico, los cuales son capaces de poner, sobre la base de valores singulares (o valores propios), significado real a las medidas descritas anteriormente.

3.3. Técnicas de Análisis Numérico en Detección de Multicolinealidad

Algunos trabajos, proporcionados especialmente por analistas numéricos, han desarrollado un conjunto muy útil de herramientas para diagnosticar de una forma rigurosa el problema de multicolinealidad. Estos trabajos no se han caracterizado precisamente por haberse enfocado de manera directa en diagnosticar problemas de multicolinealidad, sino por haberse dirigido a temas relacionados. Los analistas numéricos se han interesado, por ejemplo, en estudiar las propiedades de una matriz A en un sistema lineal de ecuaciones $Az = c$, de tal manera que esta matriz permita obtener una solución numéricamente estable para z. El enfoque utilizado en estos trabajos está estrechamente relacionado con el problema de multicolinealidad cuando se tiene una estimación por mínimos cuadrados ordinarios, puesto que el estimador obtenido por este método es una solución para el sistema lineal $(X^TX)\beta = X^Ty$, cuya matriz de varianza y covarianza está dada por $\sigma^2(X^TX)^{-1}$. De esta forma, el problema de multicolinealidad en la matriz de datos X puede estudiarse a través de una matriz $A = X^TX$, cuyo mal *acondicionamiento* causa que tanto la solución para β como la matriz de varianzas y covarianzas de este sean numéricamente inestables. Así pues, las técnicas usadas por los analistas numéricos finalmente terminan relacionándose con la comprensión de los problemas estadísticos asociados a la presencia de multicolinealidad (Belsley, Kuh and Welsch, 1982).

Las técnicas basadas en análisis numérico se caracterizan por tener un enfoque centrado en analizar qué columnas de una matriz pueden ser descartadas haciendo menores sacrificios para un análisis posterior. Este enfoque es poco frecuente cuando el análisis es hecho desde un punto de vista estadístico. En este tipo de casos, el análisis se concentra en determinar cuáles variables deben permanecer o cuáles pueden ser borradas de acuerdo a pruebas de significancia estadísticas, no actuando en función de la estabilidad numérica (Belsley, Kuh and Welsch, 1982). Sin embargo, como se verá más adelante, el conjunto de técnicas de análisis numérico ofrece elementos fundamentales para diagnosticar problemas de multicolinealidad.

Por ejemplo, Silvey (1969) demuestra que la multicolinealidad puede ser señalada por la presencia de un valor propio pequeño en X^TX. No obstante, el trabajo de Silvey no establece un límite a partir del cual pueda considerarse que un valor propio es pequeño. Belsley, Kuh and Welsch (1982) proponen una solución a este inconveniente basada en el número de condición, el cual se definirá más adelante. Adicionalmente, el trabajo de Silvey proporciona una base para cuantificar la participación de cada variable explicativa en cada una de las relaciones de dependencia lineal detectadas. Esta propuesta consiste en descomponer la varianza estimada de cada coeficiente de regresión, de manera que se pueda identificar la degradación de cada uno de ellos debida a posibles relaciones de dependencia lineal. Esto se verá de forma detallada en la sección 3.7.

De esta manera, siguiendo paso a paso el procedimiento propuesto por Belsley, Kuh and Welsch (1982), en este trabajo, inicialmente, se aplican las técnicas de análisis numérico sugeridas por Silvey (1969) para proporcionar un conjunto de indicadores (índices de condición, definidos en la sección 3.6) que muestran la presencia de una o más relaciones de dependencia lineal entre las columnas de X y, finalmente, se adapta la descomposición de varianza de Silvey, de manera que se pueda combinar con los índices anteriores para detectar las variables aleatorias que están involucradas en las relaciones de dependencia y para evaluar el grado en que las varianzas de los coeficientes de regresión están siendo afectadas por la presencia de estas dependencias.

3.4. Descomposición en Valores Singulares (SVD) de una Matriz

Cualquier matriz X de dimensión nxp, considerada como una matriz de n observaciones sobre p variables, puede descomponerse como:

$$X = UDV^T$$

Donde $U^TU = V^TV = I_p$ y D es una matriz diagonal no negativa con elementos $\mu_k, k = 1, 2, ..., p$, llamados los valores singulares de X (Belsley, Kuh and Welsch, 1982).

La SVD está estrechamente relacionada con los conceptos de valores y vectores propios, pero tiene diferencias útiles. Note que $X^TX = VD^2V^T$, por lo que V es una matriz ortogonal que diagonaliza X^TX y por tanto los elementos de la diagonal de D^2, los cuadrados de los valores singulares, son los valores propios de la matriz simétrica X^TX. Además, las

columnas ortogonales de V son los vectores propios de X^TX (Belsley, Kuh and Welsch, 1982).

La SVD de X, por tanto, ofrece información que incluye la dada por el sistema de valores y vectores propios de X^TX. Belsley, Kuh and Welsch plantean tres razones básicas en favor del uso de la SVD. En primer lugar, hacen notar que se aplica directamente a la matriz de datos X, la cual es el objeto de atención, y no a la matriz X^TX de productos cruzados. En segundo lugar, prueban que la noción de un número de condición de X (definido en la sección 3.5) está bien definida en términos de los valores singulares de X y no de las raíces cuadradas de los valores propios de X^TX. Y tercero, argumentan que mientras que el sistema de valores y vectores propios y la SVD de una matriz de datos son matemáticamente equivalentes, computacionalmente no lo son. La mayoría de los algoritmos existentes permiten calcular la SVD de X con mucha mayor estabilidad numérica que el sistema de valores y vectores propios de X^TX, particularmente en el caso en que X esté mal *acondicionada*. Como cuestión práctica, entonces, siguiendo las razones dadas por los autores, el diagnóstico de multicolinealidad que se lleva a cabo utiliza un algoritmo estable para la SVD de X, en lugar de un algoritmo para determinar los valores y vectores propios de X^TX.

En la mayoría de las aplicaciones, las dependencias lineales entre las columnas de X no son exactas, por lo cual los valores singulares iguales a cero rara vez ocurren. No obstante, debe tenerse en cuenta que cada relación lineal aproximada entre las columnas de X se manifestará en un valor singular μ_k pequeño (Belsley, Kuh and Welsch, 1982).

La cuestión ahora es determinar qué es pequeño. Aunque en última instancia el juicio sobre qué es o no pequeño sigue siendo empírico, el número de condición de la matriz X, presentado a continuación, ayuda en gran medida a responder esta pregunta.

3.5. El Número de Condición

Una medida para estudiar el *acondicionamiento* de una matriz, que es intuitiva y evita las limitaciones de las técnicas expuestas en la sección 3.2, es producida por la descomposición en valores singulares. Como se verá a continuación, se considerará razonable pensar que una matriz $\mathbf{A}$ está mal *acondicionada* si el producto entre su norma espectral y la de $\mathbf{A}^{-1}$ es grande. Esta medida, llamada el número de condición de $\mathbf{A}$, proporciona información resumida sobre las posibles dificultades que se encuentran al realizar diversos cálculos basados en $\mathbf{A}$; cuanto mayor sea el número de condición, más mal *acondicionada* estará la matriz dada (Belsley, Kuh and Welsch, 1982). Estos autores muestran que el número de condición proporciona una medida de la sensibilidad de la solución del vector $\mathbf{z}$ al sistema lineal $Az = c$, cuando se realizan pequeños cambios en los elementos de $\mathbf{A}$ y $\mathbf{c}$.

A continuación se hace un breve repaso sobre algunos conceptos mencionados por Belsley, Kuh and Welsch (1982) respecto al número de condición de una matriz, con el fin de hacer comprender la importancia que este medida tiene en el contexto de las soluciones de sistemas de ecuaciones lineales.

Inicialmente, la norma euclidiana de un vector **z** de dimensión n, denotada $\|z\|$, está definida como:

$$\|z\| = \sqrt{z^T z}$$

Como una generalización de la norma euclidiana a una matriz cuadrada **A** de dimensión nxn, la norma espectral, denotada $\|A\|$, se define como:

$$\|A\| = \sup_{\|z\|=1} \|Az\|$$

Belsley, Kuh and Welsch (1982) muestran que $\|A\| = \mu_{max}$, es el máximo valor singular de **A**. Similarmente, si **A** es cuadrada, $\|A^{-1}\| = 1/\mu_{min}$. Además, como la norma euclidiana, la norma espectral tiene las características de una verdadera norma, es decir, posee las siguientes propiedades:

1. $\|\lambda A\| = |\lambda| \, \|A\|$ para λ real y todo **A**.
2. $\|A\| = 0$ si y solo si $A = 0$, la matriz de ceros.
3. $\|A + B\| \leq \|A\| + \|B\|$ para matrices **A** y **B** de mn.

Y, adicionalmente, la norma espectral obedece a las siguientes relaciones:

4. $\|Az\| \leq \|A\| \, \|z\|$ que sigue directamente de la definición.
5. $\|AB\| \leq \|A\| \, \|B\|$ para todo **A** y **B** proporcionales.

La norma espectral es un insumo indispensable en un análisis de *acondicionamiento* de un sistema de ecuaciones lineales $Az = c$ (**A** cuadrada de dimensión nxn y no singular) con solución $z = A^{-1}c$. Interesa cuantificar cuánto cambia el vector **z** cuando se hacen pequeñas

perturbaciones en los elementos de **A** o **c**, denotadas δA y δc, respectivamente (Belsley, Kuh and Welsch, 1982).

En el caso donde **A** es fijado pero **c** cambia en δc, se tiene $\delta z = A^{-1}\delta c$, o

$$\|\delta z\| \leq \left\|A^{-1}\right\| \|\delta c\|$$

Por el empleo de la propiedad **4)** para el sistema de ecuaciones, se tiene

$$\|c\| \leq \|A\| \|z\|$$

Organizando ambas expresiones, se tiene

$$\frac{\|\delta z\|}{\|z\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\delta c\|}{\|c\|}$$

Es decir, la magnitud $\|A\| \|A^{-1}\|$ proporciona una cota para el cambio relativo en la magnitud del vector solución **z** que puede resultar de un cambio relativo en la longitud de **c**. Un resultado similar se tiene para las perturbaciones de los elementos de la matriz **A**. Belsley, Kuh and Welsch (1982) demuestran que:

$$\frac{\|\delta z\|}{\|z + \delta z\|} < \|A\| \left\|A^{-1}\right\| \frac{\|\delta A\|}{\|A\|}$$

Debido a su utilidad en este contexto, la magnitud $\|A\| \|A^{-1}\|$ es presentada por Belsley, Kuh and Welsch (1982) como el número de condición de la matriz no singular **A** y es denotado como $\kappa(A)$. Las expresiones anteriores muestran que $\kappa(A)$ proporciona una medida de la sensibilidad potencial de la solución de un sistema lineal de ecuaciones (que incluyen las ecuaciones de mínimos cuadrados normales) a los cambios en los elementos **c** y **A** del sistema lineal.

El *acondicionamiento* de cualquier matriz cuadrada **A** se puede cuantificar, entonces, a través de su número de condición $\kappa(A)$, definido como el producto entre su máximo valor singular de **A** (su norma espectral) y el máximo valor singular de su inversa. Este concepto es extendido a una matriz rectangular (Harville, 1997) y se puede calcular sin recurrir a la inversa. Desde la SVD, $X = UDV^T$, Belsley, Kuh and Welsch (1982) muestran que la inversa generalizada $\mathbf{X}^+$ de **X** es VD^+U^T, donde D^+ es la inversa generalizada de **D** y es simplemente **D** con sus elementos diagonales diferentes de cero invertidos. Por lo tanto, los valores singulares de $\mathbf{X}^+$ son los inversos de **X**, y el máximo valor singular de $\mathbf{X}^+$ es el recíproco del mínimo valor singular de **X** (Belsley, Kuh and Welsch, 1982). Entonces, para cualquier matriz **X** de *nxp*, el número de condición se define como

$$\kappa(X) = \frac{\mu_{max}}{\mu_{min}} \geq 1$$

Se puede mostrar que el número de condición de cualquier matriz con columnas ortonormales es la unidad, y por tanto $\kappa(X)$ alcanza su límite inferior en este tipo de casos. Por otra parte, $\kappa(X) = \kappa(X^+)$, por lo tanto, el número de condición tiene la propiedad altamente deseable de proporcionar la misma información si se trata de **X** o de su inversa generalizada (Belsley, Kuh and Welsch, 1982).

3.6. Índices de Condición

Inicialmente, debe tenerse en cuenta que una dependencia lineal exacta entre las columnas de X genera un cero en los valores singulares. La extensión de esta propiedad acerca de las dependencias lleva a Belsley, Kuh and Welsch (1982) a sugerir que la presencia de dependencias lineales aproximadas se verá reflejada en valores singulares (o valores propios) pequeños. Esta sugerencia no incluye una medida que determine qué es o no pequeño. El grado de mal *acondicionamiento* de **X** depende de qué tan pequeño sea el mínimo valor singular con relación al máximo; es decir, μ_{max} proporciona una referencia para la pequeñez. Para esto, dichos autores definen

$$\eta_k = \frac{\mu_{max}}{\mu_k}, \quad k = 1, 2, 3, ..., p$$

los k-ésimos índices de condición de la matriz **X** de dimensión *nxp*.

Por supuesto, $\eta_k \geq 1$ para toda k, donde el límite inferior necesariamente ocurre para alguna k. El valor más grande para η_k es también el número de condición de la matriz dada. Un valor singular que es pequeño con respecto a la referencia μ_{max}, entonces tiene un alto índice de condición.

De esta forma, hay tantas dependencias lineales aproximadas entre las columnas de una matriz de datos $\mathbf{X}$ como altos índices de condición hay (valores singulares pequeños con respecto a μ_{max}) (Belsley, Kuh and Welsch, 1982).

Cabe destacar que no se ha redirigido el problema de determinar cuánto pequeño es pequeño simplemente a determinar cuánto grande es grande. Como se vio anteriormente, tomados por sí solos, los valores singulares (o valores propios) no arrojan ninguna luz sobre el *acondicionamiento* de una matriz de datos: pueden haber casos donde matrices de datos bien *acondicionadas* tengan valores singulares bajos. Determinar cuándo un valor singular es pequeño, entonces, no necesariamente es suficiente para determinar la presencia de una relación lineal aproximada que cause una matriz de datos mal *acondicionada*. No obstante, estudiar el número de condición para determinar cuándo un valor singular es pequeño en comparación con μ_{max} (o, equivalentemente, determinar cuándo un índice de condición es alto) sí está directamente relacionado con este problema (Belsley, Kuh and Welsch, 1982)

Adicionalmente, incluso si existe el término *grande* cuando se trata de índices de condición, no hay una base a priori para determinar qué tan grande debe ser un índice de condición para que haya evidencia de multicolinealidad o, más importante aún, evidencia de datos cuyas dependencias lineales perjudiquen las estimaciones de regresión. Lo que ha de ser considerado como un índice de condición grande es una cuestión que debe determinarse empíricamente. Belsley, Kuh and Welsch (1982) sugieren que las relaciones débiles se asocian con índices de condición entre 5 y 10, mientras que las relaciones entre moderadas y fuertes se asocian con índices de condición entre 30 y 100.

3.7. Descomposición de la Varianza de los Coeficientes de Regresión

Resumiendo lo concluido en la sección 3.6, los valores singulares pequeños con respecto a μ_{max} en una matriz $\mathbf{X}$, son interpretados como indicadores de dependencias lineales aproximadas (entre las columnas de $\mathbf{X}$) asociadas a esos valores singulares. A continuación, siguiendo el trabajo de Belsley, Kuh and Welsch (1982), se muestra cómo la varianza estimada de cada coeficiente de regresión puede ser descompuesta en una suma de términos, cada uno asociado con un único valor singular, proporcionando así un medio para cuantificar el grado en

que las dependencias lineales aproximadas (que tienen altos índices de condición) degradan (se convierten en una parte predominante) cada varianza estimada. Esta descomposición, citando a dichos autores, proporciona una conexión entre el análisis numérico de la matriz de datos $\mathbf{X}$, que se concreta con la descomposición en valores singulares, y la precisión del posterior análisis de regresión que usa a dicha matriz $\mathbf{X}$ como la matriz de datos, que se concreta en la matriz de varianza y covarianza de $\hat{\beta}$.

La matriz de varianza y covarianza del estimador de mínimos cuadrados $\hat{\beta} = (X^T X)^{-1} X^T y$ es $\sigma^2 (X^T X)^{-1}$, donde σ^2 es la varianza común de ξ en el modelo lineal $y = X\beta + \xi$. Usando la SVD, $X = UDV^T$, la matriz de varianza y covarianza de $\hat{\beta}$, var$(\hat{\beta})$, puede ser escrita como:

$$\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 V D^{-2} V^T$$

o, la k-ésima componente de $\hat{\beta}$,

$$\text{var}(\hat{\beta}_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\mu_j^2}$$

Donde los μ_j son los valores singulares y $V \equiv [v_{ij}]$.

Esta ecuación descompone a var$(\hat{\beta}_k)$ en una suma de componentes, cada una asociada con uno de los p valores singulares μ_j (o valores propios μ_j^2). Como las μ_j^2 aparecen en el denominador, aquellas componentes asociadas con dependencias lineales aproximadas (es decir, pequeños μ_j) serán relativamente grandes en comparación con las demás. Esto sugiere, pues, que una proporción inusualmente alta de la varianza de dos o más coeficientes, concentrada en los componentes asociados con el mismo valor singular pequeño, proporciona evidencia de que una dependencia lineal aproximada está causando problemas (Belsley, Kuh and Welsch, 1982).

Defina la (k, j) ésima proporción de descomposición de varianza como la proporción de la varianza del k-ésimo coeficiente de regresión asociada con el j-ésimo componente de su descomposición. Estas proporciones se pueden calcular de la siguiente forma:

Primero sea

$$\phi_{kj} = \frac{v_{kj}^2}{\mu_j^2}, \quad j,k = 1,2,3,...,p$$

y

$$\phi_k = \sum_{j=1}^{p} \phi_{kj}$$

De esta forma, las proporciones de descomposición de varianza se calculan como:

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k}, \quad k,j = 1,2,3,...,p$$

Una manera rápida y sencilla de presentar estos resultados puede ser tal como se sugiere en la Tabla 3.1.

Valor Singular Asociado	$var(\hat{\beta}_1)$	$var(\hat{\beta}_2)$	...	$var(\hat{\beta}_p)$
μ_1	π_{11}	π_{12}	...	π_{1p}
μ_2	π_{21}	π_{22}	...	π_{2p}
...	...	...	...	...
μ_p	π_{p1}	π_{p2}	...	π_{pp}

Tabla 3-1.: Índices de condición y proporciones asociadas de varianza para cada parámetro estimado.

Note que π_{jk} hace uso de la información obtenida por SVD en dependencias aproximadas, como una manera aplicada directamente para examinar sus efectos en la precisión de las estimaciones de regresión (Belsley, Kuh and Welsch, 1982).

En la descomposición de varianza dada para cada coeficiente de regresión, μ_j's pequeños conducen a magnitudes grandes en las correspondientes componentes. Sin embargo, no todas las componentes de var$(\hat{\beta}_k)$ se verán necesariamente afectadas por μ_j's pequeños, porque los v_{kj}^2's en el numerador pueden ser incluso más pequeños. En el caso extremo donde $v_{kj} = 0$, var$(\hat{\beta}_k)$ no se verá afectado por dependencias aproximadas entre las columnas de $\mathbf{X}$ que puedan causar μ_j's muy pequeños (Belsley, Kuh and Welsch, 1982).

3.8. Procedimiento Sugerido para Diagnosticar Multicolinealidad

Belsley, Kuh and Welsch (1982) sugieren un procedimiento práctico para probar la presencia de una o más dependencias lineales aproximadas entre las columnas de una matriz de datos **X** y, adicionalmente, evaluar el grado en que cada una de estas dependencias degrada la precisión de las estimaciones de regresión que se hacen con base en dicha matriz de datos.

Estos autores sugieren que un medio apropiado para diagnosticar los daños causados por la presencia de multicolinealidad debe cumplir con la siguiente condición doble:

1. Un valor singular que indique tener un índice de condición alto, y el cual simultáneamente esté asociado con

2. Una alta proporción en la descomposición de la varianza estimada de dos o más coeficientes de regresión.

Los índices de condición considerados de gran tamaño (mayores a 30) en **1.** identifican el número de dependencias lineales aproximadas entre las columnas de la matriz de datos **X**. Por otra parte, la determinación en **2.** de proporciones grandes en la descomposición de varianza (mayores a 0.5) asociadas con un alto índice de condición identifican aquellas variables que están involucradas en las dependencias correspondientes, y la magnitud de estas proporciones en conjunto con un alto índice de condición proporciona una medida del grado en que la correspondiente estimación de regresión ha sido afectada por la presencia de multicolinealidad (Belsley, Kuh and Welsch, 1982).

A Tener en Cuenta

Inicialmente parece que la concentración de varianza de cualquier coeficiente de regresión en alguno de sus componentes puede indicar que la multicolinealidad está causando problemas. Sin embargo, dado que se requieren dos o más variables para crear una dependencia lineal aproximada, debe ocurrir que dos o más varianzas se vean afectadas negativamente con altas proporciones en su descomposición, asociadas con un único valor singular (es decir, una única dependencia aproximada) (Belsley, Kuh and Welsch, 1982).

Debido al hecho de que dos o más columnas de **X** deben participar en una dependencia lineal aproximada, la degradación de una estimación debido a la multicolinealidad puede ser observada solo cuando un valor singular μ_j está asociado con una proporción grande de la

varianza de dos o más coeficientes (Belsley, Kuh and Welsch, 1982).

3.9. Efectos Nocivos de la Multicolinealidad

Los efectos nocivos resultantes de una regresión basada en datos con problemas de multicolinealidad pueden ser de dos tipos: computacionales y estadísticos (Belsley, Kuh and Welsch, 1982).

Estadísticamente, el problema provocado por la presencia de multicolinealidad en una matriz de datos es la disminución en la precisión de las estimaciones estadísticas derivadas de esos datos, es decir, la multicolinealidad causa un incremento artificial en las varianzas. Este problema se debe a que cuando los datos están mal *acondicionados*, algunas series de datos son combinaciones casi lineales de otras y, por tanto, añaden muy poca información nueva, independiente de la información estadística que puede ser obtenida.

Es importante resaltar que las varianzas infladas son bastante perjudiciales cuando la regresión se usa como base para pruebas de hipótesis, estimación y predicción. Muchos analistas de regresión lineal enfrentan la situación en la que una prueba de significancia no llega a un resultado concluyente a causa de una varianza innecesariamente alta inducida por datos colineales, o que un intervalo de confianza o de pronóstico resulta inútilmente grande, lo que refleja la falta de datos correctamente *acondicionados* que puedan permitir la estimación apropiada de estos intervalos.

Estos problemas en datos multicolineales pueden ser resueltos a través de la introducción de nuevos datos bien *acondicionados*. En muchas aplicaciones, sin embargo, no se tiene disponibilidad de nuevos datos, o éstos pueden ser adquiridos pero con un gran costo en tiempo y esfuerzo. La utilidad de contar con herramientas de diagnóstico que indiquen la presencia de multicolinealidad e incluso poder aislar las variables involucradas es evidente, por tanto, para que el investigador puede determinar si el esfuerzo para corregir la multicolinealidad vale la pena (Belsley, Kuh and Welsch, 1982).

3.10. Multicolinealidad Nociva vs Degradante

Inicialmente, es importante señalar que la multicolinealidad no siempre es perjudicial. Además, puede mostrarse que combinaciones lineales específicas de los coeficientes de regresión estimados pueden ser bien determinadas incluso si los coeficientes individualmente no lo son. Si, de manera particular, el objetivo del análisis de centra únicamente en los parámetros

estimados no afectados o en combinaciones lineales bien determinadas de los coeficientes estimados, no existe ningún problema (Belsley, Kuh and Welsch, 1982). En un caso menos extremo, y por tanto más útil, se tiene que la descomposición de la varianza estimada para el k-ésimo coeficiente de regresión, $\text{var}(\hat{\beta}_k)$ es $s^2 \sum_j v_{kj}^2/\mu_j^2$, donde s^2 es la varianza estimada del error. Si s^2 es suficientemente pequeña, puede que, de manera particular, las $\text{var}(\hat{\beta}_k)$'s sean pequeñas para propósitos de pruebas específicos a pesar de haber grandes componentes en los términos v_{kj}^2/μ_j^2 resultantes de dependencias lineales aproximadas. Así, por ejemplo, si el interés del análisis se centra en determinar si un coeficiente dado es significativamente positivo y se llega, incluso en presencia de multicolinealidad, a aceptar esta hipótesis con base en la prueba t, la multicolinealidad no ha causado un problema. Por supuesto, los pronósticos resultantes o estimaciones puntuales pueden tener intervalos de confianza más amplios de lo necesario para satisfacer un análisis más ambicioso, pero para el pronóstico de la prueba de significancia inicialmente propuesta, la multicolinealidad no ha causado prácticamente ningún daño (Belsley, Kuh and Welsch, 1982).

Proporcionar evidencia de que la multicolinealidad ha dañado la estimación, sin embargo, es absolutamente más complicado. Para esto, hay que mostrar, por ejemplo, que un intervalo de predicción que es demasiado amplio para un fin determinado podría ser adecuadamente reducido si se ponen condiciones estadísticas para *acondicionar* los datos (o que un intervalo de confianza puede ser reducido adecuadamente, o la precisión de un estimador puede aumentar apropiadamente). Hasta la fecha ningún procedimiento proporciona dicha información. No obstante, si el investigador tiene información de que **1)** hay fuertes dependencias lineales entre los datos, y **2)** las varianzas de los parámetros (o intevalos de confianza basados en éstas) que son de interés para él tienen una gran magnitud a causa de la presencia de datos multicolineales, de manera que la multicolinealidad es potencialmente dañina, entonces él debe emprender un largo camino para decidir si se justifican los costos de la acción correctiva. Adicionalmente, dicha información ayuda a indicar cuándo las varianzas de interés no están siendo negativamente afectadas y se podría continuar sin emprender reparaciones adicionales. La información anterior es, desde luego, proporcionada por el número de condición y las proporciones de descomposición de varianza usadas en los dos frentes de diagnóstico sugeridos anteriormente. Así que cuando se ha cumplido esta condición conjunta, se dice que los coeficientes de regresión afectados se han degradado (pero no necesariamente dañado) por la presencia de multicolinealidad; degradado en el sentido de que la magnitud de la varianza estimada está determinada principalmente por la presencia de una relación lineal. Por lo tanto, existe la presunción de que los intervalos de confianza, intervalos de predicción y estimadores puntuales podrían perfeccionarse, si es necesario, introduciendo un mejor *acondicionamiento* a los datos (Belsley, Kuh and Welsch, 1982).

4. Estimación por Mínimos Cuadrados No Lineales

El modelo de regresión lineal se puede escribir como

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \xi_i, \quad i = 1, 2, 3, ..., n$$

o en su forma matricial

$$y = X\beta + \xi$$

El estimador de mínimos cuadrados ordinarios para el parámetro (o vector de parámetros) β está dado por la expresión matricial

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Donde $\mathbf{X}$ es una matriz de dimensión $n\text{x}(k+1) = n\text{x}p$, que tiene en sus columnas las n observaciones de cada unas de las $k+1 = p$ variables predictoras (cada una con un parámetro a estimar asociado), y y es un vector columna, formado por las n observaciones de la variable dependiente.

Considérese ahora el caso en que la relación de dependencia es de la forma

$$y = f(x; \beta) + \xi$$

donde $f(x;\beta)$ es una función no lineal en las componentes del vector β.

Existen casos donde la función $f(x;\beta)$ es no lineal únicamente en las variables explicativas, con lo cual un cambio de variable permite transformar este modelo en uno de tipo lineal. Sin embargo, no es objetivo de este trabajo abordar este caso.

4.1. Mínimos Cuadrados en Modelos No Lineales

Cuando la relación de dependencia entre la variable dependiente y y las variables regresoras $x_1, x_2, ..., x_k$ es no lineal en los parámetros, el procedimiento de mínimos cuadrados no lineales se enfoca en resolver el problema de optimización

$$min_\beta \sum_{i=1}^{n}(y_i - f(x_i;\beta))^2 = min_\beta \sum_{i=1}^{n}\xi_i^2$$

donde y_i es la i-ésima observación de la variable respuesta, x_i es un vector fila con la i-ésima observación en cada una de las variables predictoras y $f(x_i;\beta)$ es la función no lineal evaluada en x_i.

Buscar este óptimo se traduce, al aplicar reglas de derivación, en resolver el sistema de ecuaciones

$$\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)^T y = \left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)^T f(x,\beta),$$

donde la matriz gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)$ tiene n filas (una para cada observación) y k columnas (determinadas por cada una de las variables predictoras sobre las que se debe derivar parcialmente la función no lineal evaluada en la observación i), mientras que $y = [y_1, y_2, ..., y_n]^T$ y $f(x,\beta) = f(x_1, x_2, ..., x_n;\beta) = [f(x_1;\beta), f(x_2;\beta), ..., f(x_n;\beta)]^T$ son vectores columna de dimensión n. Este sistema puede tener múltiples soluciones, o no tenerla (Seber and Wild, 2003).

La matriz de covarianzas del parámetro estimado está dada por

$$\text{var}(\hat{\beta}) = \sigma_\xi^2 \left[\left(\frac{\partial f(x;\beta)}{\partial \beta} \right)^T \left(\frac{\partial f(x;\beta)}{\partial \beta} \right) \right]^{-1}$$

donde la matriz gradiente $\left(\frac{\partial f(x;\beta)}{\partial \beta}\right)$ tiene n filas y k columnas, mientras que $\hat{\beta}$ es un vector de dimensión k. En el caso del modelo lineal, esta matriz de varianzas y covarianzas se se reduce a la expresión $\sigma_\xi^2[X^T X]^{-1}$

4.2. Aproximación Lineal del Modelo No Lineal

Un primer acercamiento a la estimación del modelo no lineal consiste en obtener, mediante la expansión de Taylor, la mejor aproximación lineal de la función $f(x_i;\beta)$ alrededor de un estimador inicial para β, denotado como β^* [1], y estimar por mínimos cuadrados ordinarios el modelo lineal resultante. La aproximación de Taylor de primer orden para la función no lineal se puede escribir como

$$f(x_i;\beta) \approx f(x_i;\beta^*) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta} \right)_{\beta=\beta^*} (\beta - \beta^*), \quad i = 1, 2, 3, ..., n$$

De esta forma, el modelo no lineal se puede expresar como

$$y_i \approx f(x_i;\beta^*) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta} \right)_{\beta=\beta^*} (\beta - \beta^*) + \xi_i, \quad i = 1, 2, 3, ..., n$$

Que también se puede escribir como

$$y_i - f(x_i;\beta^*) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta} \right)_{\beta=\beta^*} \beta^* = \left(\frac{\partial f(x_i;\beta)}{\partial \beta} \right)_{\beta=\beta^*} \beta + \xi_i, \quad i = 1, 2, 3, ..., n$$

[1] Existen diversos procedimientos para seleccionar adecuadamente valores iniciales en un proceso de optimización numérica. Estos métodos tienen en cuenta desde la experiencia del especialista que conoce la naturaleza de los datos y del modelo, hasta valores elegidos convenientemente para simplificar la función objetivo a optimizar. No obstante, estos métodos no son abordados en este trabajo dado que no hacen parte del alcance del mismo.

De esta forma, se obtiene el modelo lineal

$$y_i^* \approx \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\beta^*} \beta + \xi_i, \quad i = 1, 2, 3, ..., n$$

donde

$$y_i^* = y_i - f(x_i;\beta^*) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\beta^*} \beta^*, \quad i = 1, 2, 3, ..., n$$

De esta forma, dado un valor inicial β^* para el estimador, se puede construir la variable y_i^*, así como las k variables que componen el valor del gradiente de la función $f(x_i;\beta)$ en el punto $\beta = \beta^*$. Las *realizaciones* de estas nuevas variables están en función de las observaciones muestrales de y_i y x_i y del vector de valores iniciales β^*. A continuación, se estima a través de mínimos cuadrados ordinarios el modelo lineal que tiene como variable respuesta a y_i^* y como variables explicativas las componentes de la matriz gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\beta^*}$, con lo cual se obtiene una nueva estimación para β (Seber and Wild, 2003).

El nuevo estimador obtenido por mínimos cuadrados ordinarios para β, denotado como $\widetilde{\beta}$, sería:

$$\widetilde{\beta} = \left[\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)\right]^{-1}_{\beta=\beta^*} \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T_{\beta=\beta^*} y_i^*$$

$$\widetilde{\beta} = \left[\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)\right]^{-1}_{\beta=\beta^*} \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T_{\beta=\beta^*} \left(y_i - f(x_i;\beta^*) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\beta^*} \beta^*\right)$$

$$\widetilde{\beta} = \beta^* + \left[\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)\right]^{-1}_{\beta=\beta^*} \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)^T_{\beta=\beta^*} \xi_i^*$$

donde $\xi_i^* = y_i - f(x_i;\beta^*)$ son los nuevos residuales obtenidos con la estimación inicial β^*. Esta expresión proporciona la nueva estimación $\widetilde{\beta}$ a partir de β^*. Con esta nueva estimación $\widetilde{\beta}$ como valor inicial, se repite el proceso de manera idéntica. Finalmente, tras repetir este

proceso varias veces, se obtiene la estimación definitiva para β, denotada como $\widehat{\beta}$.

A partir de $\widehat{\beta}$ y de las observaciones muestrales para y_i y x_i, es posible construir el modelo lineal que tiene como variable respuesta a y_i^* y como variables explicativas a las componentes de la matriz gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}}$. Este modelo lineal tiene una importancia trascendental puesto que es sobre éste sobre quien se realizarán todos los diagnósticos para detectar presencia de valores extremos y/o problemas de multicolinealidad.

Este modelo puede escribirse como

$$y_i - f(x_i;\widehat{\beta}) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \widehat{\beta} = \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \beta + \xi_i, \quad i = 1, 2, 3, ..., n$$

O, similarmente

$$y_i^* = \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \beta + \xi_i, \quad i = 1, 2, 3, ..., n$$

Con el fin de simplificar un poco la notación, el modelo se puede expresar como

$$y_i^* = W\beta + \xi_i, \quad i = 1, 2, 3, ..., n$$

donde

$$W = \left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}, \quad i = 1, 2, 3, ..., n; j = 1, 2, 3, ..., k$$

donde n es el número de observaciones y k el número de parámetros a estimar. La matriz de diseño de este modelo lineal, en este caso W, como se explicó en los capítulos 2 y 3, tiene una importancia fundamental en el diagnóstico sobre presencia de datos extremos y problemas de multicolinealidad.

4.3. Algunos Algoritmos Numéricos de Optimización

En esta sección se presentan los algoritmos de estimación numérica propuestos por Newton - Rapshon y Gauss - Newton, los cuales son comúnmente usados en procesos de estimación no lineal.

4.3.1. Algoritmo de Newton - Raphson

Inicialmente se dispone de una estimación $\widehat{\theta}_n$ del mínimo $\widehat{\theta}$ de una función $F(\theta)$ con derivada continua. Si se considera una pequeña vecindad del punto $\widehat{\theta}_n$, el valor numérico de F en un punto de dicho entorno puede aproximarse mediante un desarrollo en serie de Taylor de orden 2

$$F(\theta) \cong M(\theta) = F(\widehat{\theta}_n) + [\nabla F(\widehat{\theta}_n)]^T(\theta - \widehat{\theta}_n) + \frac{1}{2}(\theta - \widehat{\theta}_n)^T[\nabla^2 F(\widehat{\theta}_n)](\theta - \widehat{\theta}_n)$$

donde $\nabla F(\widehat{\theta}_n)$ y $\nabla^2 F(\widehat{\theta}_n)$ denotan, respectivamente, el gradiente y matriz hessiana de la función $F(\theta)$ evaluados en el punto $\widehat{\theta}_n = \theta$. La actual estimación, $\widehat{\theta}_n$, se puede mejorar reemplazándola por el vector que minimice la expresión cuadrática anterior. Esto ocurre en el punto en que

$$\frac{\partial M(\theta)}{\partial \theta} = \nabla F(\widehat{\theta}_n) + [\nabla^2 F(\widehat{\theta}_n)](\theta^* - \widehat{\theta}_n) = 0$$

es decir

$$\widehat{\theta}_{n+1} = \theta^* = \widehat{\theta}_n - [\nabla^2 F(\widehat{\theta}_n)]^{-1}\nabla F(\widehat{\theta}_n)$$

Esta expresión constituye el algoritmo de Newton - Raphson y sugiere cómo aproximarse al valor desconocido del vector $\widehat{\theta}$ a partir de un vector $\widehat{\theta}_n$ suficientemente próximo a él. Debe tenerse en cuenta que el punto obtenido como nueva estimación, denotado θ^*, minimiza realmente el valor de la función F en el entorno de $\widehat{\theta}_n$ si la matriz hessiana $\nabla F(\widehat{\theta}_n)$ es definida positiva. Esto se garantiza si F es convexa en el punto $\widehat{\theta}_n$, es decir, si dicho punto estaba ya suficientemente próximo a un mínimo local de la función F (Seber and Wild, 2003).

Este algoritmo se utiliza de un modo iterativo, utilizando la nueva estimación como punto de partida en cada etapa del algoritmo y llevando a cabo iteraciones hasta que se satisfagan

los criterios de convergencia previamente estipulados.

El algoritmo de Newton - Raphson puede utilizarse para obtener numéricamente el estimador mínimo cuadrático en el caso de un modelo de regresión no lineal en los parámetros. La función objetivo está dada por

$$F(\theta) = S(\beta) = \sum_{i=1}^{n} [y_i - f(x_i; \beta)]^2$$

Se tiene como propósito encontrar el vector de coeficientes $\widehat{\beta}$ que minimiza la expresión $S(\beta)$. En este caso, $S(\beta)$ representa a la función genérica $F(\theta)$ tenida en cuenta por Newton - Raphson. De esta forma

$$\widehat{\beta}_n \cong \beta_{n-1} - [\nabla^2 F(\beta_{n-1})]^{-1} \nabla F(\beta_{n-1})$$

Para aplicar el algoritmo, se deben utilizar las expresiones para el gradiente y la matriz hessiana de $S(\beta)$, los cuales son, respectivamente

$$\nabla F(\theta) = \frac{\partial S(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} \left[[y_i - f(x_i; \beta)] \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \right]$$

$$\nabla^2 F(\theta) = \frac{\partial^2 S(\beta)}{\partial \beta \partial \beta^T} = 2 \sum_{i=1}^{n} \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right)^T - 2 \sum_{i=1}^{n} \left[[y_i - f(x_i; \beta)] \frac{\partial^2 f(x_i; \beta)}{\partial \beta \partial \beta^T} \right]$$

Al sustituir estas dos expresiones en el algoritmo de Newton - Raphson, se tiene

$$\widehat{\beta}_n = \widehat{\beta}_{n-1} - \left(\frac{\partial^2 S(\beta)}{\partial \beta \partial \beta^T} \right)^{-1}_{\widehat{\beta}_{n-1}} \left(\frac{\partial S(\beta)}{\partial \beta} \right)_{\widehat{\beta}_{n-1}}$$

$$\widehat{\beta}_n = \widehat{\beta}_{n-1} + \sum_{i=1}^{n} \left[\left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right)^T - \frac{\partial^2 f(x_i; \beta)}{\partial \beta \partial \beta^T} \xi_i \right]^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \xi_i \right]$$

donde

$$\xi_i = y_i - f(x_i; \beta)$$

4.3.2. Algoritmo de Gauss - Newton

Este algoritmo es una variante del algoritmo de Newton - Raphson, especialmente útil cuando se tiene como propósito estimar por mínimos cuadrados un modelo no lineal. La función objetivo, nuevamente, es

$$F(\theta) = S(\beta) = \sum_{i=1}^{n} [y_i - f(x_i; \beta)]^2$$

Volviendo a la expresión dada para $\widehat{\beta}_n$ por el algoritmo de Newton - Raphson, el algoritmo de Gauss - Newton consiste en ignorar el término que contiene la segunda derivada de $f(x_i; \beta)$ en la hessiana. Si la función $f(x_i; \beta)$ es aproximadamente cuadrática en el vector β, entonces el término que contiene la segunda derivada es aproximadamente proporcional a la suma de los residuales, por lo que su contribución a la hessiana es muy pequeña (Novales, 1993).

Con esta aproximación, se sustituye la hessiana por la matriz simétrica, definida positiva

$$\left[\sum_{i=1}^{n} \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right)^T \right]$$

por lo que el algoritmo de Gauss - Newton resulta

$$\widehat{\beta}_n = \widehat{\beta}_{n-1} + \left[\sum_{i=1}^{n} \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right)^T \right]_{n-1}^{-1} \left[\sum_{i=1}^{n} \left(\frac{\partial f(x_i; \beta)}{\partial \beta} \right) \xi_i \right]_{n-1}$$

donde

$$\xi_i = y_i - f(x_i; \beta)$$

Esta expresión sugiere la forma en que se debe utilizar el algoritmo: dado un estimador inicial $\widehat{\beta}_{n-1}$ del vector β, se obtienen los residuales correspondientes a dicha estimación y se estima la regresión auxiliar de $\widehat{\xi}_i$ sobre el vector gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)$, evaluado en $\widehat{\beta}_{n-1}$. Este vector gradiente tiene la misma dimensión que β, por lo que se tendrán k coeficientes δ estimados en la regresión auxiliar, que serán las correcciones a introducir en la estimación inicial $\widehat{\beta}_{n-1}$, para obtener una nueva estimación $\widehat{\beta}_n$

$$\widehat{\beta}_n = \widehat{\beta}_{n-1} + \widehat{\delta} = \widehat{\beta}_{n-1} + \left[\left(\frac{\partial f(\widehat{\beta}_{n-1})}{\partial \beta}\right)^T \left(\frac{\partial f(\widehat{\beta}_{n-1})}{\partial \beta}\right)\right]^{-1} \left(\frac{\partial f(\widehat{\beta}_{n-1})}{\partial \beta}\right)^T \widehat{\xi}$$

Este resultado puede utilizarse, a su vez, como estimación inicial en la siguiente etapa del algoritmo.

5. Detección de Datos Extremos y Multicolinealidad en el Modelo No Lineal

5.1. Detección de Datos Extremos en el Modelo No Lineal

Teniendo en cuenta que $\widehat{\beta}$, el vector de estimadores definitivos (obtenido en la última i teración, una vez cumplidos todos los criterios de convergencia establecidos) para el conjunto de parámetros considerados en el modelo no lineal, es calculado por mínimos cuadrados ordinarios a partir de un modelo lineal que tiene como variable dependiente a $y_i^* = y_i - f(x_i;\widehat{\beta}) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \widehat{\beta}$ y como variables regresoras al conjunto de k componentes del gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}} (j = 1, 2, 3, ..., k)$, los problemas que este estimador pueda tener debido a la presencia de observaciones extremas en los datos, deben evaluarse teniendo como base a y_i^* y a $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}$. Estos dos insumos se pueden reconstruir a partir del vector $\widehat{\beta}$ y de las observaciones muestrales para y_i y x_i. De esta forma, las medidas utilizadas para diagnosticar la presencia de observaciones extremas son, desde luego, las utilizas por Belsley, Kuh and Welsch (1982) para el modelo lineal.

Tomando en consideración que

$$W = \left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}, \quad j = 1, 2, 3, ..., k$$

es una matriz gradiente de n filas y k columnas definida en el capítulo 4, las medidas utilizadas para diagnosticar la presencia de datos extremos son, como en el caso del modelo lineal, las siguientes

$$h_i = W_i(W^T W)^{-1} y_i^{*T}$$

donde W_i es un vector fila con la i-ésima observación en cada una de las k componentes del gradiente W.

$$\epsilon_i = y_i^* - \hat{y}_i^*$$

$$\epsilon_{si} = \frac{\epsilon_i}{s\sqrt{1-h_i}}$$

$$\epsilon_i^* = \frac{\epsilon_i}{s_{(i)}\sqrt{1-h_i}}$$

$$\text{DFBETA}_i = \frac{(W^TW)^{-1}W_i^T\epsilon_i}{1-h_i}$$

$$\text{DFBETAS}_{ij} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^n c_{jk}^2}}\frac{\epsilon_i}{s_{(i)}(1-h_i)}$$

Donde

$$C = (W^TW)^{-1}W^T$$

$$\text{DFFIT}_i = \frac{h_i\epsilon_i}{1-h_i}$$

$$\text{DFFITS}_i = \sqrt{\left[\frac{h_i}{1-h_i}\right]}\frac{\epsilon_i}{s_{(i)}\sqrt{1-h_i}}$$

$$\text{COVRATIO} = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{\epsilon_i^{*2}}{n-p}\right]^p (1-h_i)}$$

Los puntos de corte para estas medidas a partir de los cuales se debe considerar como extrema la observación correspondiente son los mismos propuestos por Belsley, Kuh and Welsch (1982) para el caso del modelo lineal.

Es de fundamental importancia tener en cuenta que las observaciones extremas detectadas por estos indicadores pueden no ser atípicas respecto a los datos originales que componen el modelo de regresión no lineal: en cambio, sí lo son con respecto a las realizaciones de la matriz de diseño $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}$ y de la variable independiente y_i^* que conforman el modelo lineal dado por $y_i^* = \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \beta + \xi_i$ y sobre el cual se estima por mínimos cuadrados ordinarios el vector de parámetros β.

Por último, debe tenerse presente que estas medidas tienen las mismas limitaciones mencionadas en el capítulo 2 cuando hay presencia de casos de enmascaramiento en los datos.

5.2. Detección de Multicolinealidad en el Modelo No Lineal

Inicialmente, recordando algunos conceptos básicos del capítulo 3, los problemas de multicolinealidad en la estimación de un modelo de regresión lineal están asociados a las características numéricas de la matriz de diseño X y tales problemas se diagnostican y estudian tomando como insumo a dicha matriz. Como también se analizó en el capítulo 3, la presencia de multicolinealidad en los datos provoca un crecimiento artificial en la varianza de los coeficientes estimados, haciendo que éstos pierdan precisión. Ahora bien, teniendo en cuenta que el estimador $\widehat{\beta}$ de mínimos cuadrados en un modelo de regresión no lineal se obtiene a partir de un modelo lineal en el que se utilizan como variable respuesta a $y_i^* = y_i - f(x_i;\widehat{\beta}) + \left(\frac{\partial f(x_i;\beta)}{\partial \beta}\right)_{\beta=\widehat{\beta}} \widehat{\beta}$ y como matriz de diseño a las k componentes del gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}} (j = 1, 2, 3, ..., k)$, el diagnóstico de multicolinealidad que afecta la precisión de dicho estimador debe hacerse, en este caso, tomando como insumo la matriz de diseño de dicho modelo lineal, es decir, a $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}$.

La matriz $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}$ puede reconstruirse a partir del vector $\widehat{\beta}$ y de las observaciones muestrales obtenidas para x_i

Dada su trascendental importancia y para tener mayor simplicidad en la notación, es útil recordar que esta matriz es denotada como

$$W = \left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}, \quad j = 1, 2, 3, ..., k$$

La matriz de correlación **R**, calculada en el caso lineal sobre **X**, para explorar posibles relaciones de dependencia lineal entre las variables regresoras del modelo, es ahora calculada sobre la matriz de pseudo datos **W**.

La descomposición en valores singulares aplicada en el modelo lineal a **X** es aplicada en el modelo no lineal a **W**, obteniéndose de esta manera los valores singulares, el número de condición y los índices de condición.

De la misma forma que se usó **X** en el modelo lineal como insumo para calcular la descomposición de varianza de los parámetros estimados y determinar cómo se afectaba ésta debido a las relaciones de dependencia lineal, se usa **W** en el modelo no lineal con el mismo propósito.

El problema de multicolinealidad, en este caso, no se interpreta como la existencia de relaciones de dependencia lineal entre las variables regresoras originales x_i sino entre las columnas de la matriz gradiente $\left(\frac{\partial f(x_i;\beta)}{\partial \beta_j}\right)_{\beta=\widehat{\beta}}$, las cuales sirven como insumo para la estimación de $\widehat{\beta}$ por mínimos cuadrados ordinarios.

5.3. Aplicación

5.3.1. Detección de Datos Extremos en el Modelo No Lineal

En este punto, interesa aplicar la metodología descrita para detección de datos influyentes y atípicos en el modelo no lineal a un problema real. Con este fin, se consideran los datos de Stromberg (1993), referentes a la concentración de Nitrógeno, recogidos por la Agencia de Protección del Medio Ambiente de Estados Unidos en 29 lagos de la Florida. Stromberg utiliza un modelo no lineal para relacionar la media anual de concentración total de nitrógeno (TN) como variable respuesta, con el afluente promedio de concentración de nitrógeno (NIN) y el tiempo de retención del agua (TW), como variables predictoras.

Dado que se tienen tres variables, el comportamiento de éstas por pares es presentado en la Figura **5-1**.

La forma funcional de esta relación es la siguiente

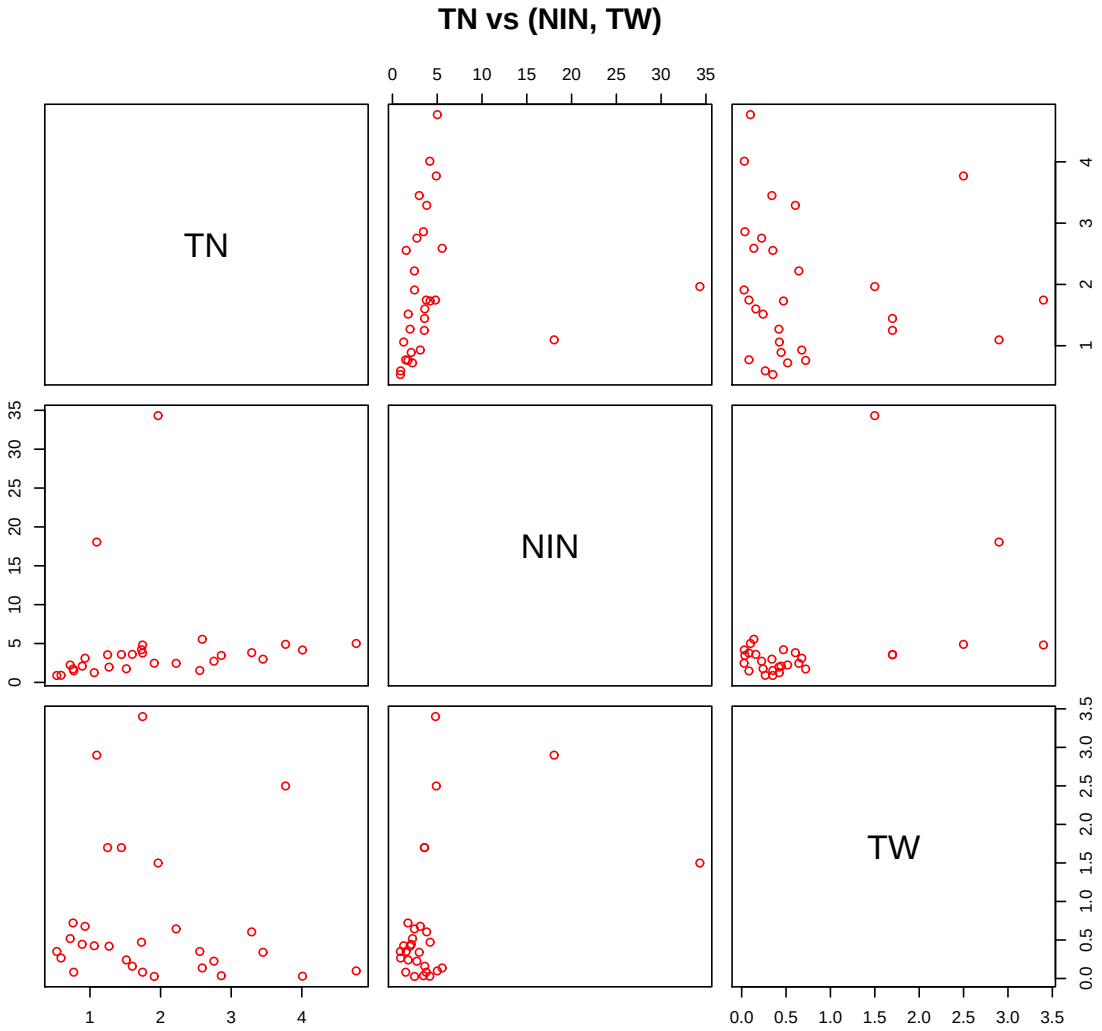

Figura 5-1.: Comportamiento por pares de TN vs (NIN, TW)

$$\mathrm{TN}_i = \frac{\mathrm{NIN}_i}{1 + \delta \mathrm{TW}_i^{\beta}} + \xi_i, \quad i = 1, 2, 3, ..., 29$$

Donde δ y β son parámetros desconocidos.

Stromberg, adicionalmente, sugiere que los puntos 10 y 23 presentan un comportamiento verdaderamente atípico respecto a las demás observaciones.

Teniendo como insumos la forma matemática del modelo no lineal y el conjunto de datos,

se emplea un procedimiento previamente diseñado en R, el cual recibe esta información y calcula las medidas propuestas para detección de datos atípicos e influyentes; realizando internamente todo el proceso de estimación no lineal, construcción de pseudo datos, aproximación lineal y cálculo de medidas de diagnóstico.

Para este caso particular, teniéndose $p = 2$ y $n = 29$, se debe prestar atención especial a aquellas observaciones con DFBETAS cuyo valor absoluto sea mayor a $2/\sqrt{n} = 2/\sqrt{29} = 0.3714$, con DFFITS con valor absoluto mayor a $2\sqrt{p/n} = 2\sqrt{2/29} = 0.5252$, con h_i asociado mayor a $2p/n = 2(2/29) = 0.1379$ o con COVRATIO mayor a $3p/n + 1 = 3(2/29) + 1 = 1.2069$ o menor a $3p/n - 1 = 3(2/29) - 1 = 0.7931$. Es preciso mencionar que aquella observación que sobrepase al menos uno de los límites anteriores, debe recibir atención especial.

Los resultados obtenidos con dicho procedimiento son mostrados en la Tabla 5.1.

Los resultados muestran que la observación 1 tiene un fuerte impacto en la estimación del parámetro β del modelo, pues esta observación tiene asociado un DFBETAβ significativamente grande, lo cual indica que la estimación de este parámetro es altamente susceptible a la presencia o no de esta observación. Este dato también tiene un impacto grande en la estimación de la matriz de varianzas y covarianzas de los coeficientes estimados, puesto que su COVRATIO presenta un valor lejano de la unidad. Adicionalmente, el hecho de que esta observación presente un DFFITS y un h_i por fuera de los puntos de corte establecidos, hace que sea altamente influyente en las predicciones del modelo.

Por otra parte, la inclusión de la observación 2 tiene un fuerte impacto en la precisión de los coeficientes estimados. Esto se explica en el hecho de que al incluir esta observación en la estimación del modelo, la matriz de varianzas y covarianzas tiene un cambio importante, lo cual se refleja en un COVRATIO bastante lejano de la unidad.

La observación 16 tiene, al igual que la observación 2, tiene también un impacto importante en la precisión de los coeficientes estimados, puesto que su COVRATIO también sobrepasa los puntos de corte. Dicha observación también tiene una influencia significativa en el cálculo de los valores predichos del modelo, lo cual se deduce por su alto DFFITS asociado.

Finalmente, observación 23 tiene un gran impacto en la estimación de los parámetros β y δ del modelo, pues esta observación tiene asociados tanto un DFBETAβ como un DFBETASδ significativamente grandes, lo cual indica que la estimación de ambos parámetros es altamente susceptible a la presencia o no de esta observación particular. La observación 23 también tiene un impacto importante en la estimación de la matriz de varianzas y covarianzas de los coeficientes estimados, puesto que su COVRATIO asociado presenta un valor por fuera de los puntos de corte. Adicional a todo lo anterior, el hecho de que este dato presente

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	e_i^*	h_i
1	0.259	**-0.830**	**-0.932**	**1.220**	−1.565	**0.262**
2	−0.233	−0.229	0.299	**0.557**	3.217	0.009
3	−0.034	0.031	0.051	1.089	0.364	0.019
4	−0.069	−0.047	0.077	1.024	0.883	0.008
5	−0.315	0.121	0.365	0.885	1.812	0.039
6	−0.012	0.003	0.013	1.101	0.086	0.021
7	0.098	−0.274	−0.314	1.158	−0.855	0.119
8	−0.057	−0.038	0.063	1.044	0.728	0.008
9	−0.243	0.297	0.425	0.931	1.690	0.059
10	−0.054	−0.058	0.072	1.186	0.224	0.094
11	0.067	−0.332	−0.358	1.031	−1.274	0.073
12	−0.026	0.022	0.038	1.071	0.430	0.008
13	−0.006	0.005	0.008	1.100	0.056	0.020
14	−0.095	0.190	0.232	1.094	0.861	0.068
15	−0.026	0.049	0.061	1.098	0.363	0.027
16	−0.047	0.197	0.215	**1.264**	0.488	**0.162**
17	−0.001	0.041	0.149	0.976	1.235	0.014
18	−0.002	0.004	0.005	1.086	0.057	0.007
19	−0.004	0.004	0.006	1.083	0.086	0.005
20	0.002	−0.024	−0.025	1.072	−0.374	0.004
21	0.001	−0.001	−0.001	1.096	−0.007	0.016
22	−0.064	−0.076	0.091	0.964	1.253	0.005
23	**7.250**	**4.153**	**-7.785**	**2.454**	−3.729	**0.813**
24	−0.108	0.126	0.184	0.929	1.491	0.015
25	0.008	−0.040	−0.043	1.077	−0.405	0.011
26	0.004	−0.002	−0.005	1.098	−0.035	0.018
27	−0.028	0.020	0.038	1.162	0.135	0.073
28	0.004	−0.042	−0.044	1.085	−0.348	0.016
29	−0.019	0.003	0.020	1.079	0.259	0.006

Tabla 5-1.: Indicadores para detección de datos atípicos en estimación de modelo no lineal para datos Stromberg. Las observaciones extremas son marcadas en negrita

un DFFITS y un h_i por fuera de los límites establecidos por Belsley, Kuh and Welsch (1982), hace que también sea altamente influyente en los valores predichos del modelo.

Los resultados obtenidos son consecuentes con Stromberg en cuanto al dato 23, pues él había anunciado previamente que ésta había sido una observación extrema en su estudio sobre la concentración de nitrógeno en los 29 lagos de la Florida, mientras que las medidas de detección de datos extremos usadas en este trabajo también la detectaron reiteradamente. Respecto a la otra observación extrema anunciada por Stromberg, la cual fue marcada con

el número 10, ninguna de las medidas usadas en este trabajo la detectó.

5.3.2. Detección de Multicolinealidad en el Modelo No Lineal

Ejemplo Respuesta Ultrasónica vs Distancia del Metal

Se tienen los resultados de un estudio hecho por NIST (National Institute of Standards and Technology) que involucra calibración ultrasónica. La variable respuesta es la respuesta ultrasónica y la variable predictora es la distancia del metal (Chwirut and NIST, 1979).

Se propone un modelo no lineal para relacionar estas dos variables, dado por la siguiente expresión

$$y = f(x;\beta) + \xi$$

donde

$$f(x;\beta) = \frac{e^{-\beta_1 x}}{\beta_2 + \beta_3 x}$$

Siendo β_1, β_2 y β_3 parámetros desconocidos.

Se tienen 54 observaciones referentes a estas dos variables, las cuales están almacenadas en la librería NISTnls de R bajo el nombre Chwirut2. El encabezado de estos datos se presenta en la Tabla 5.2.

Donde RespUltr es la variable respuesta y DistMetal la regresora.

Análisis Exploratorio sobre Variables Originales

Inicialmente se hace un análisis exploratorio sobre la respuesta ultrasónica y sobre la distancia del metal. En la Tabla 5.3 se presentan algunas medidas descriptivas para estas dos variables.

Cabe destacar que no se tienen coeficientes de variación llamativamente pequeños.

ID	RespUltr	DistMetal
1	92.9000	0.500
2	57.1000	1.000
3	31.0500	1.750
4	11.5875	3.750
5	8.0250	5.750
6	63.6000	0.875
7	21.4000	2.250
8	14.2500	3.250
9	8.4750	5.250
10	63.8000	0.750
11	26.8000	1.750
12	16.4625	2.750
13	7.1250	4.750
14	67.3000	0.625
15	41.0000	1.250
...	...	...

Tabla 5-2.: Primeros quince datos de Chwirut2 con respuesta ultrasónica y distancia del metal.

	RespUltr	DistMetal
Media	31.5641	2.5463
Desv Estándar	26.3130	1.7034
Mediana	20.1000	2.3750
Coef. Variación	0.8837	0.6690
Mínimo	3.7500	0.5000
Máximo	92.9000	6.0000

Tabla 5-3.: Medidas descriptivas para respuesta ultrasónica y distancia del metal.

El comportamiento univariado de las dos variables consideradas se muestra detalladamente en la Figura **5-2**.

Por otra parte, la Figura **5-3** muestra el comportamiento de la respuesta ultrasónica en presencia de la distancia del metal. Este gráfico muestra evidencia suficiente para considerar que la relación entre estas dos variables es de tipo no lineal.

Estimación Modelo No Lineal

Luego de ajustar el modelo no lineal propuesto, se procede a estimar sus respectivos parámetros. Las estimaciones puntuales junto con sus errores estándar, sus valores t y sus respectivas

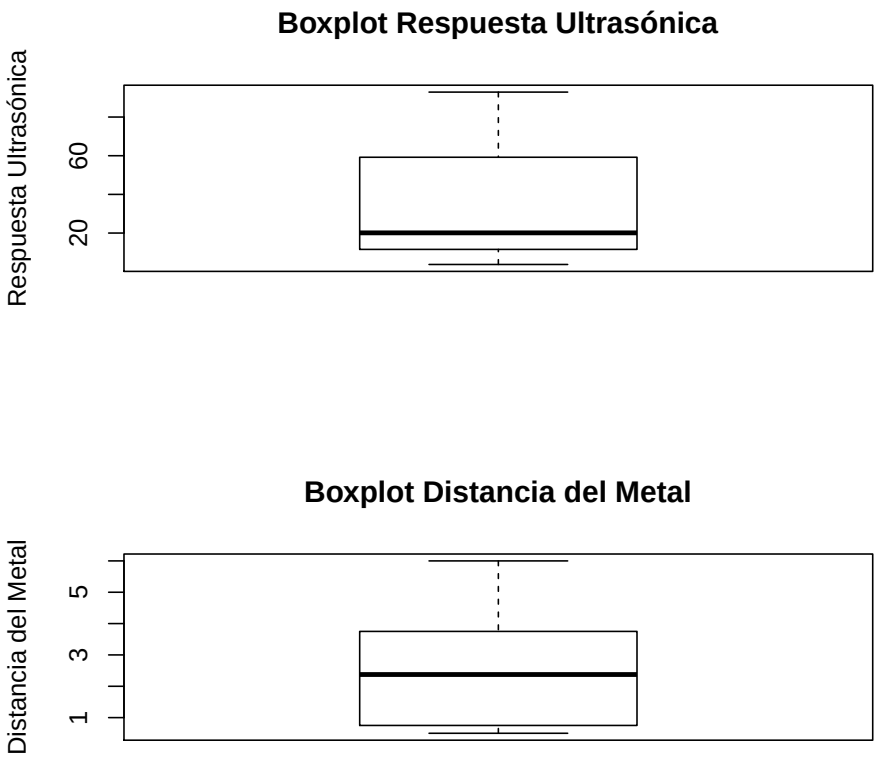

Figura 5-2.: Comportamiento Univariado de Respuesta Ultrasónica y Distancia del Metal

pruebas de significancia univariadas, se presentan en la Tabla 5.4.

	Estimación	**Error Estándar**	**Valor t**	**Valor p**
β_1	0.1666	0.0383	4.3490	$6.56e-05$
β_2	0.0052	0.0007	7.7530	$3.54e-10$
β_3	0.0122	0.0015	7.9390	$1.81e-10$

Tabla 5-4.: Parámetros estimados y sus errores estándar asociados en el modelo no lineal para respuesta ultrasónica y distancia del metal.

Según las pruebas de significancia individuales, los parámetros considerados en el modelo son estadísticamente significativos.

Variables Pseudo - Regresoras

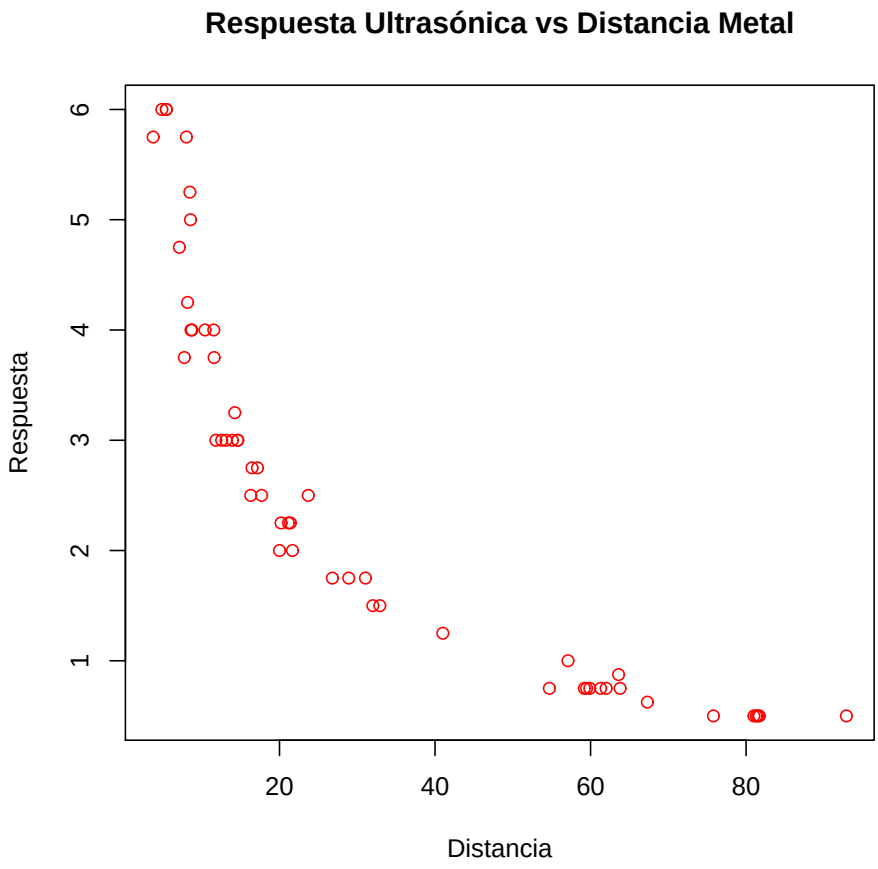

Figura 5-3.: Gráfico de Dispersión de la Respuesta Ultrasónica vs Distancia del Metal

Ahora bien, la linealización utilizada en la estimación de este modelo, toma como insumo una nueva matriz de diseño, denotada **W**, cuyas columnas son las derivadas parciales de la función no lineal con respecto a cada uno de los parámetros a estimar. Así pues, la matriz **W** tiene la siguiente estructura.

$$W = [W_1, W_2, W_3]_{\beta=\hat{\beta}}$$

donde W_j $(j = 1, 2, 3)$ son vectores de dimensión n que tienen en su entrada i $(i = 1, 2, 3, ..., n)$ la respectiva derivada parcial evaluada en la realización de la observación i.

En este caso, se tiene:

$$W_1 = \frac{\partial f(x;\beta)}{\partial \beta_1} = \frac{-xe^{-\beta_1 x}}{(\beta_2 + \beta_3 x)}$$

$$W_2 = \frac{\partial f(x;\beta)}{\partial \beta_2} = \frac{-e^{-\beta_1 x}}{(\beta_2 + \beta_3 x)^2}$$

$$W_3 = \frac{\partial f(x;\beta)}{\partial \beta_3} = \frac{-xe^{-\beta_1 x}}{(\beta_2 + \beta_3 x)^2}$$

Al evaluar estas derivadas en los parámetros estimados y en la matriz de diseño original, se obtiene la matriz de diseño que sirve como insumo al modelo linealizado, la cual contiene unas nuevas variables regresoras, conocidas como pseudo regresoras. El encabezado de esta matriz se presenta en la Tabla 5.5.

ID	W_1	W_2	W_3
1	−40.9279	−7282.3234	−3641.1617
2	−48.8906	−2823.5440	−2823.5440
3	−49.4738	−1069.7332	−1872.0331
4	−39.5820	−208.0747	−780.2802
5	−29.4084	−68.1680	−391.9662
6	−47.8789	−3463.9605	−3030.9655
7	−47.5870	−650.7055	−1464.0873
8	−42.3564	−291.8674	−948.5690
9	−31.7545	−87.7188	−460.5239
10	−46.3598	−4329.2996	−3246.9747
11	−49.4738	−1069.7332	−1872.0331
12	−45.0869	−424.9908	−1168.7248
13	−34.2425	−114.6499	−544.5869
14	−44.1414	−5535.3659	−3459.6037
15	−49.8719	−1960.2912	−2450.3641
...	...	...	...

Tabla 5-5.: Tabla de variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal.

Teniendo en cuenta que esta nueva matriz de diseño se utiliza en la estimación del modelo lineal, sobre dicha matriz se deben hacer todas las pruebas de *adecuación* a fin de determinar si entre sus columnas se presentan relaciones de dependencia lineal que puedan conllevar a problemas de multicolinealidad.

Análisis Exploratorio sobre Variables Pseudo - Regresoras

A continuación, se hace un análisis exploratorio sobre las variables pseudo regresoras contenidas en las columnas de **W**. Los resultados se presentan en la siguiente Tabla 5.6.

	W_1	W_2	W_3
Media	−42.4976	−1928.7245	−1775.1354
Desv Estándar	6.2480	2439.3633	1151.9034
Mediana	−43.7359	−586.3693	−1384.5851
Coef. Variación	−0.1470	−1.2648	−0.6489
Mínimo	−49.9505	−7282.3234	−3641.1617
Máximo	−28.2900	−60.3981	−362.3886

Tabla 5-6.: Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal.

Ninguna de las variables presenta un coeficiente de variación llamativamente pequeño, lo cual hace pensar que es poco probable que se presenten problemas de mala adecuación debido a poca variación en alguna de las variables.

Búsqueda de Multicolinealidad

La búsqueda de multicolinealidad en la matriz **W** inicia con la exploración de posibles relaciones lineales entre diferentes pares de sus columnas. Aunque este análisis no detecta relaciones de dependencia lineal en la que participen simultáneamente más de dos variables, es un buen inicio para identificar patrones sospechosos que luego han de ser corroborados con un análisis más riguroso. Para identificar rápidamente tendencias sospechosas entre pares de variables, se construye la matriz de dispersión. Esta matriz es mostrada en la Figura **5-4**.

Una medida que cuantifica la cantidad de relación lineal existente entre cada par de variables es, por supuesto, la matriz de correlación, la cual se presenta en la Tabla 5.7.

Llama la atención la alta correlación presentada entre W_2 y W_3, lo cual da indicios de la existencia de al menos una relación lineal aproximada entre las columnas de la matriz de diseño, lo cual se puede transformar en problemas de multicolinealidad.

Una prueba más rigurosa para detectar relaciones de dependencia lineal, presentada en este trabajo, consiste en calcular los índices de condición de la matriz **W** y asociar a cada uno de

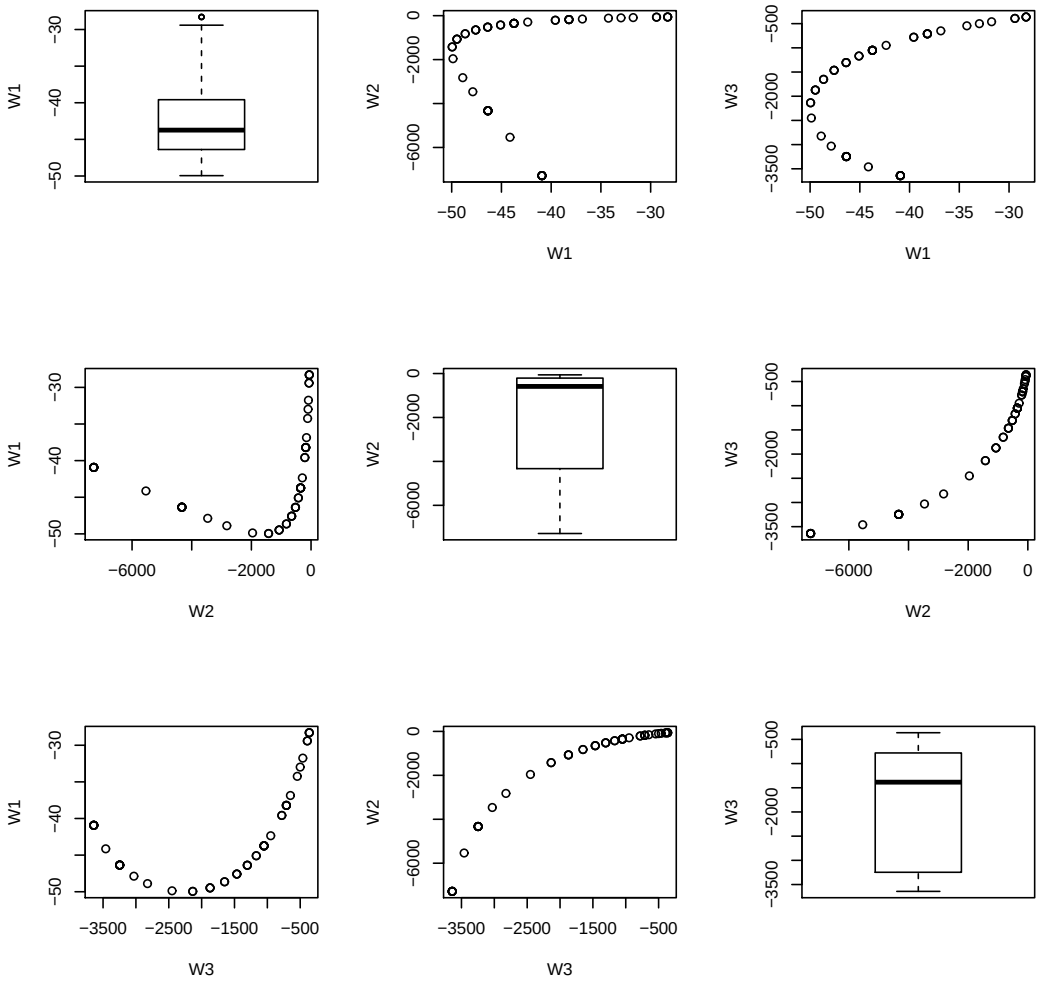

Figura 5-4.: Matriz de Dispersión Variables Pseudo Regresoras

	W_1	W_2	W_3
W_1	1.0000	0.2116	0.5083
W_2	0.2116	1.0000	0.9363
W_3	0.5083	0.9363	1.0000

Tabla 5-7.: Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal.

ellos una proporción de varianza de cada uno de los parámetros estimados. Grandes índices de condición están asociados a relaciones de dependencia lineal aproximadas, siendo cada uno de ellos producto de una relación de dependencia. Adicionalmente, grandes proporciones de varianza en algunos parámetros asociadas a un índice de condición grande, significan que

en esa relación de dependencia están involucradas las variables asociadas a dichos parámetros.

Los resultados correspondientes a este ejemplo, se presentan en la Tabla 5.8.

Índice de Condición	$var(\hat{\beta}_1)$	$var(\hat{\beta}_2)$	$var(\hat{\beta}_3)$
1.000	0.013	0.005	0.005
1.620	0.165	0.007	0.000
13.359	0.821	0.987	0.995

Tabla 5-8.: Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal para respuesta ultrasónica y distancia del metal.

Se observan grandes proporciones de varianza de cada uno de los parámetros estimados asociadas a un solo índice de condición. Aunque según los límites propuestos por Belsley, Kuh and Welsch (1982), este índice de condición no es considerablemente grande, se puede pensar que la matriz de diseño contiene una relación lineal asociada a este índice, en la cual participan simultáneamente las tres columnas de la matriz de diseño.

Ejemplo Longitud de Raíz vs Concentración

Se investiga la acción conjunta de mezclas de ácidos fenólicos en la inhibición del crecimiento de perennial ryegrass (un tipo de semilla de hierba) y su significado en investigación alelopática. Se estudia la toxicidad del ácido ferúlico (antioxidante de origen vegetal que se puede encontrar naturalmente en las aceitunas y en las moras) cuando se aplica al perennial ryegrass en diferentes rangos de concentración.

La variable respuesta, y, es la longitud de la raíz, medida en centímetros, y la variable predictora, x, es la concentración de ácido ferúlico, medida en mM. En total se tienen 24 observaciones para cada unas de las variables consideradas. Los datos son tomados de Inderjit, Streibig and Olofsdotter (2002).

Para explicar la relación entre estas dos variables se propone un modelo de tipo no lineal, el cual está dado por la siguiente expresión

$$y = f(x;\theta) + \xi$$

donde

$$f(x;\theta) = \theta_3 + \frac{\theta_4 - \theta_3}{1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}}$$

Siendo θ_2, θ_3, θ_4 y θ_5 parámetros desconocidos.

Algunos de estos datos se presentan en la Tabla 5.9.

ID	**LongitudRaíz**(y)	**Concentración**(x)
1	7.5800	0.01
2	8.0000	0.01
3	8.3285	0.01
4	7.2500	0.01
5	7.3650	0.01
6	7.9625	0.01
7	8.3556	0.94
8	6.9143	0.94
9	7.7500	0.94
10	6.8714	1.88
11	6.4500	1.88
12	5.9222	1.88
...	...	...

Tabla 5-9.: Primeros doce datos de longitud de raíz y concentración.

Donde y es la variable respuesta y x la regresora.

Análisis Exploratorio sobre Variables Originales

Inicialmente se hace un análisis descriptivo univariado de las variables medidas. En la Tabla 5.10 se presentan algunas medidas de interés, tanto para la longitud de la raíz como para la concentración aplicada.

Se destaca que ambas variables presentan un coeficiente de variación razonablemente alto.

La Figura **5-5** muestra el comportamiento univariado de estas dos variables.

La Figura **5-6** muestra el comportamiento de la longitud de la raíz en presencia de los diferentes niveles de concentración. Este gráfico muestra evidencia suficiente para considerar

	LongitudRaíz	Concentración
Media	4.3272	7.3863
Desv Estándar	3.2857	9.9778
Mediana	5.0778	2.8150
Coef. Variación	0.7593	1.3508
Mínimo	0.2200	0.0100
Máximo	8.3555	30.0000

Tabla 5-10.: Medidas descriptivas para longitud de raíz y concentración.

que la relación entre estas dos variables es de tipo no lineal.

Estimación Modelo No Lineal

Luego de ajustar el modelo no lineal propuesto, se procede a estimar sus respectivos parámetros. Las estimaciones puntuales junto con sus errores estándar, sus valores t y sus respectivas pruebas de significancia univariadas, se presentan en la Tabla 5.11.

	Estimación	Error Estándar	Valor t	Valor p
θ_2	2.9822	0.4584	6.5050	$2.43e-06$
θ_3	0.4814	0.2095	2.2980	0.0325
θ_4	7.7930	0.1897	41.0730	$< 2e-16$
θ_5	3.0580	0.1858	16.4560	$4.31e-13$

Tabla 5-11.: Parámetros estimados y sus errores estándar asociados en el modelo no lineal para longitud de raíz y concentración.

Según las pruebas de significancia individuales, los parámetros considerados en el modelo son estadísticamente significativos.

Variables Pseudo - Regresoras

Ahora bien, la linealización utilizada en la estimación de este modelo toma como insumo una nueva matriz de diseño, denotada **W**, cuyas columnas son las derivadas parciales de la función no lineal con respecto a cada uno de los parámetros a estimar. Así pues, la matriz **W** tiene la siguiente estructura

$$W = [W_1, W_2, W_3, W_4]_{\theta=\hat{\theta}}$$

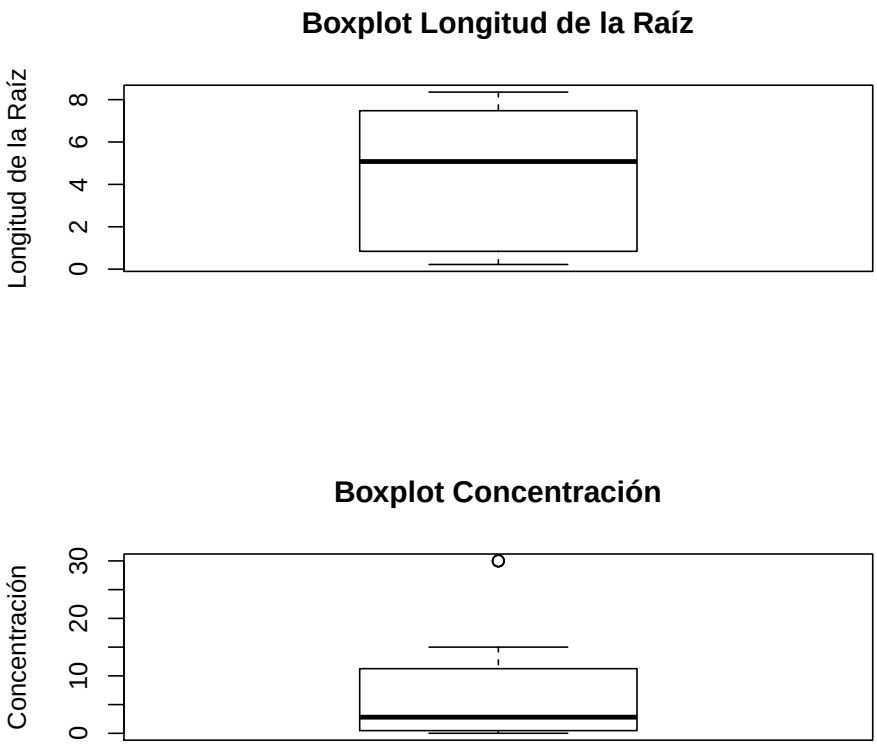

Figura 5-5.: Comportamiento Univariado de Longitud de Raíz y Concentración

donde W_j $(j = 1, 2, 3, 4)$ son vectores de dimensión n que tienen en su entrada i $(i = 1, 2, 3, ..., n)$ la respectiva derivada parcial evaluada en la realización de la observación i.

En este caso, se tiene:

$$W_1 = \frac{\partial f(x;\theta)}{\partial \theta_2} = -\frac{(\theta_4 - \theta_3)\left(\frac{x}{\theta_5}\right)^{\theta_2} log\left(\frac{x}{\theta_5}\right)}{\left(1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}\right)^2}$$

$$W_2 = \frac{\partial f(x;\theta)}{\partial \theta_3} = 1 - \frac{1}{\left(1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}\right)}$$

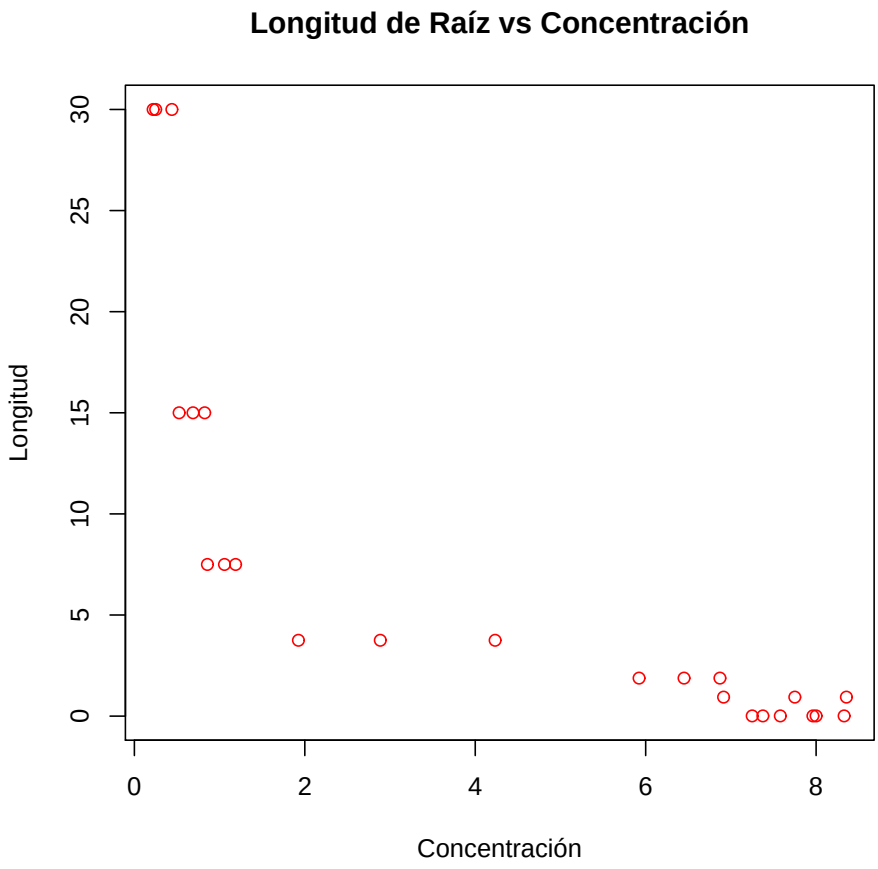

Figura 5-6.: Longitud de Raíz vs Concentración

$$W_3 = \frac{\partial f(x;\theta)}{\partial \theta_4} = \frac{1}{\left(1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}\right)}$$

$$W_4 = \frac{\partial f(x;\theta)}{\partial \theta_5} = \frac{\theta_2(\theta_4 - \theta_3)\left(\frac{x}{\theta_5}\right)^{\theta_2 - 1}\left(\frac{x}{\theta_5^2}\right)}{\left(1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}\right)^2}$$

Al evaluar estas derivadas en los parámetros estimados y en la matriz de diseño original, se obtiene la matriz de diseño que sirve como insumo al modelo linealizado, la cual contiene unas nuevas variables regresoras, conocidas como pseudo regresoras. El encabezado de esta matriz se presenta en la Tabla 5.12.

ID	W_1	W_2	W_3	W_4
1	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
2	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
3	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
4	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
5	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
6	$1.6200e-06$	$3.8717e-08$	$99.9999e-01$	$2.7607e-07$
7	$2.4130e-01$	$2.8807e-02$	$97.1193e-01$	$1.9949e-01$
8	$2.4130e-01$	$2.8807e-02$	$97.1193e-01$	$1.9949e-01$
9	$2.4130e-01$	$2.8807e-02$	$97.1193e-01$	$1.9949e-01$
10	$5.4715e-01$	$1.8988e-01$	$81.0119e-01$	$1.0969e+00$
11	$5.4715e-01$	$1.8988e-01$	$81.0119e-01$	$1.0969e+00$
12	$5.4715e-01$	$1.8988e-01$	$81.0119e-01$	$1.0969e+00$
...	...	...	...	...

Tabla 5-12.: Tabla de variables pseudo regresoras utilizadas en la linealización de modelo no lineal para longitud de raíz y concentración.

Dado que esta nueva matriz de diseño será el insumo utilizado en la estimación del modelo lineal, sobre dicha matriz se deben hacer todas las pruebas de *adecuación* a fin de determinar si entre sus columnas se presentan relaciones de dependencia lineal que puedan causar problemas de multicolinealidad.

Análisis Exploratorio sobre Variables Pseudo - Regresoras

A continuación, se hace un análisis exploratorio sobre las variables pseudo regresoras contenidas en las columnas de **W**. Los resultados se presentan en la siguiente Tabla 5.13.

	W_1	W_2	W_3	W_4
Media	−0.0082	0.4740	0.5260	0.4278
Desv Estándar	0.2888	0.4445	0.4445	0.5841
Mediana	−0.0092	0.4187	0.5818	0.1303
Coef. Variación	−35.3485	0.9376	0.8450	1.3653
Mínimo	−0.3954	0.0001	0.0011	0.0001
Máximo	0.5471	0.9989	1.0000	1.6273

Tabla 5-13.: Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para longitud de raíz y concentración.

Ninguna de las variables presenta un coeficiente de variación llamativamente pequeño, lo cual

hace pensar que es poco probable que se presenten problemas de mala adecuación debido a poca variación en alguna de las variables.

Búsqueda de Multicolinealidad

La búsqueda de multicolinealidad en la matriz **W**, nuevamente, inicia con la exploración de posibles relaciones lineales entre diferentes pares de sus columnas. Aunque este análisis tiene la limitación, como se ha mencionado anteriormente, de no detectar relaciones de dependencia lineal en las que participan simultáneamente más de dos variables, es un buen inicio para identificar patrones sospechosos que luego han de ser corroborados con un análisis más riguroso. Para identificar rápidamente tendencias sospechosas entre pares de variables, se construye la matriz de dispersión. Esta matriz es mostrada en la Figura **5-7**.

Una medida que cuantifica la cantidad de relación lineal existente entre cada par de variables es, por supuesto, la matriz de correlación, la cual se presenta en la Tabla 5.14.

	W_1	W_2	W_3	W_4
W_1	1.0000	−0.5747	0.5747	−0.0417
W_2	−0.5747	1.0000	−1.0000	0.0583
W_3	0.5747	−1.0000	1.0000	−0.0583
W_4	−0.0417	0.0583	−0.0583	1.0000

Tabla 5-14.: Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para longitud de raíz y concentración.

Llama la atención la correlación perfecta presentada entre las variables W_2 y W_3, lo cual da indicios serios de problemas de adecuación en la matriz de diseño, que se pueden transformar en problemas de multicolinealidad.

Esta correlación perfecta entre W_2 y W_3 se deriva de la forma matemática del modelo no lineal, ya que al diferenciar éste respecto a W_2 y W_3, se producen unas derivadas parciales estrechamente relacionadas, cuya suma es matemáticamente igual a 1.

Es decir, si

$$f(x;\theta) = \theta_3 + \frac{\theta_4 - \theta_3}{1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}}$$

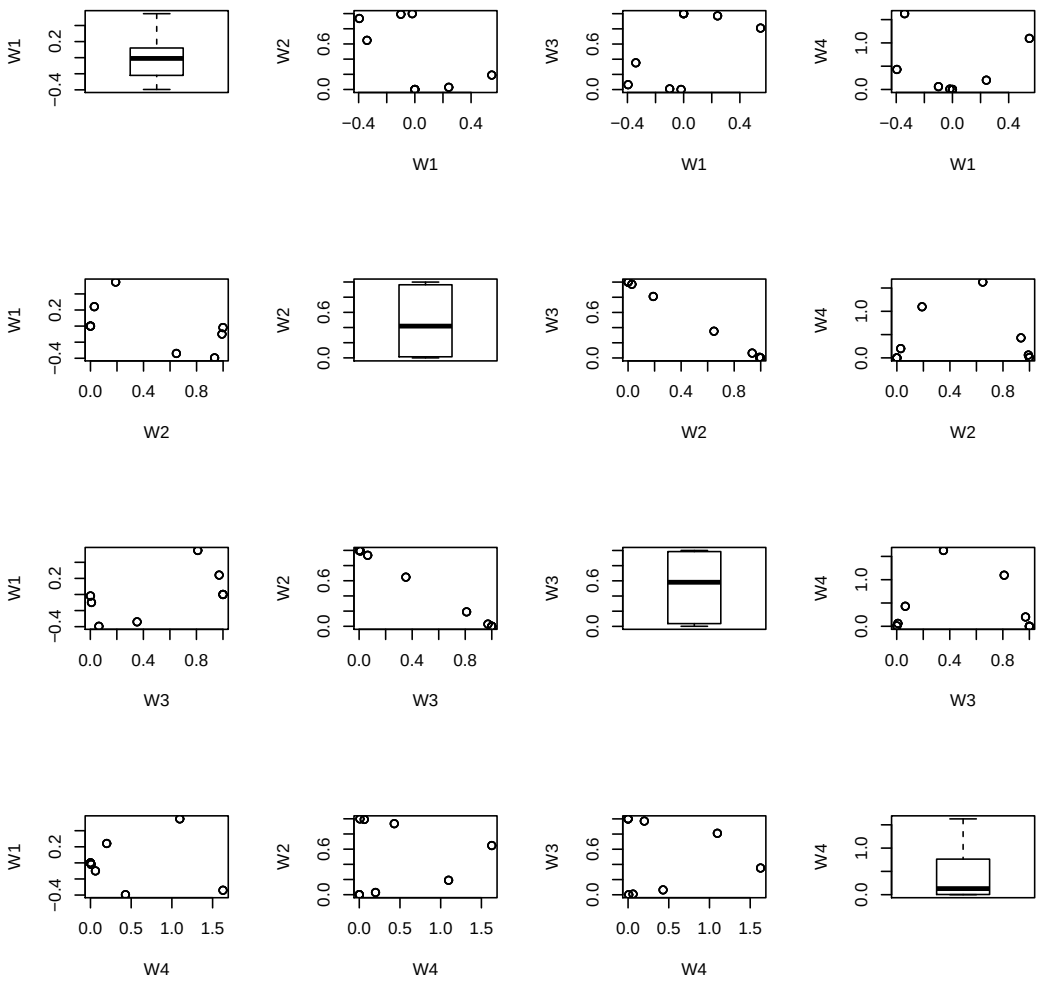

Figura 5-7.: Matriz de Dispersión Variables Pseudo Regresoras

Se puede probar que

$$\frac{\partial f(x;\theta)}{\partial \theta_3} + \frac{\partial f(x;\theta)}{\partial \theta_4} = 1 + \frac{1}{1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}} - \frac{1}{1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}} = 1$$

Esto produce una relación lineal exacta entre dos de las columnas de la matriz de diseño. Por consiguiente, independiente de los datos que se tengan, la forma matemática del modelo no lineal lleva a un serio problema de multicolinealidad en la aproximación lineal utilizada.

Igual que en el ejemplo anterior, se presentan a continuación los índices de condición de la matriz **W** junto con las proporciones de varianza de cada parámetro estimado asociadas a cada respectivo índice. Por supuesto, grandes índices de condición están asociados a relaciones de dependencia lineal aproximadas, siendo cada uno de ellos producto de una relación de dependencia. Finalmente, grandes proporciones de varianza en algunos parámetros asociadas a un índice de condición grande, indican que en esa relación de dependencia están involucradas las variables asociadas a dichos parámetros.

Los resultados correspondientes a este ejemplo, se presentan en la Tabla 5.15.

Índice de Condición	$var(\hat{\theta}_2)$	$var(\hat{\theta}_3)$	$var(\hat{\theta}_4)$	$var(\hat{\theta}_5)$
1.000	0.064	0.000	0.000	0.000
1.573	0.001	0.000	0.000	0.000
2.123	0.935	0.000	0.000	0.000
14315890151333272.000	0.000	1.000	1.000	1.000

Tabla 5-15.: Índices de condición y descomposición de varianza para parámetros estimados en la linealización de modelo no lineal para longitud de raíz y concentración.

Como era de esperarse, se observa un índice de condición inusualmente grande, el cual tiene asociadas enormes proporciones de la varianza estimada de los coeficientes θ_2, θ_3 y θ_4. De esta manera, existe una fuerte relación lineal entre las columnas de la matriz de diseño, en la cual participan, además de las columnas θ_2 y θ_3, la columna θ_4, lo cual no había sido posible detectar inicialmente a través de la matriz de correlación.

Haciendo $\theta_1 = \theta_4 - \theta_3$, la relación entre la longitud de la raíz y la concentración puede ser escrita de la siguiente forma:

$$y = f(x;\theta) + \xi$$

donde

$$f(x;\theta) = \theta_3 + \frac{\theta_1}{1 + \left(\frac{x}{\theta_5}\right)^{\theta_2}}$$

Siendo θ_1, θ_2, θ_3 y θ_5 parámetros desconocidos.

Las estimaciones de esta re especificación se presentan en la Tabla 5.16.

	Estimación	**Error Estándar**	**Valor t**	**Valor p**
θ_1	7.3116	0.2981	24.5240	$< 2e-16$
θ_2	2.9822	0.4584	6.5050	$2.43e-06$
θ_3	0.4814	0.2095	2.2980	0.0325
θ_5	3.0580	0.1858	16.4560	$4.31e-13$

Tabla 5-16.: Parámetros estimados y sus errores estándar asociados en el modelo no lineal para longitud de raíz y concentración.

De la Tabla 5.11 observe que $\hat{\theta}_4 - \hat{\theta}_3$ reproducen exactamente el $\hat{\theta}_1$ presentado en esta última tabla.

Esta re especificación de la forma no lineal del modelo evita el problema presentado con la parametrización original ya que no conduce a variables pseudo regresoras con dependencias lineales exactas entre sí.

Al calcular los índices de condición y al descomponer la varianza, se obtienen los resultados presentados en la Tabla 5.17.

Índice de Condición	$var(\hat{\theta}_1)$	$var(\hat{\theta}_2)$	$var(\hat{\theta}_3)$	$var(\hat{\theta}_5)$
1.000	0.043	0.005	0.043	0.072
1.410	0.016	0.475	0.007	0.060
2.096	0.078	0.142	0.060	0.787
4.047	0.863	0.377	0.889	0.081

Tabla 5-17.: Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal alterno para longitud de raíz y concentración.

Al no presentarse ningún índice de condición considerablemente grande, no hay indicios suficientes para pensar en la existencia de problemas de multicolinealidad bajo esta re parametrización.

5.4. Un Análisis de la Capacidad de la Metodología para Detectar Presencia de Datos Extremos o Relaciones de Dependencia Lineal en los Datos Originales

Inicialmente, es importante precisar que la metodología utilizada en este trabajo tiene como objetivo diagnosticar problemas de precisión en los coeficientes estimados de un modelo de regresión no lineal causados por la presencia de datos extremos o relaciones de multicolinealidad en el proceso de estimación de los mismos. No obstante, la presencia de estos problemas no debe investigarse en el modelo no lineal directamente, sino en el modelo lineal que se utiliza como insumo para hacer la estimación de dichos coeficientes a través mínimos cuadrados ordinarios. De esta forma, la presencia de datos extremos o problemas de multicolinealidad debe estudiarse directamente en dicho modelo lineal. Este modelo, haciendo un breve repaso, tiene una variable respuesta no observada directamente pero si relacionada estrechamente con la respuesta original del modelo no lineal y tiene una matriz de diseño tampoco observada directamente, que depende tanto de la forma funcional del modelo no lineal como de la matriz de diseño original.

No obstante lo anterior, es interesante analizar hasta qué punto esta metodología podría detectar presencia de datos extremos en la información original o relaciones de dependencia lineal en la matriz de diseño utilizada en el modelo no lineal. Con este fin, se presentan dos casos que pretenden evaluar este alcance: en uno de ellos, partiendo de uno de los ejemplos presentados en este capítulo, se agrega a la información original una serie de datos extremos artificiales con el fin de evaluar a partir de qué punto la metodología podría detectarlos y marcarlos como atípicos; en el segundo caso, se presenta una variación de uno de los modelos no lineales utilizados en el este capítulo, en el cual se incorpora a la función no lineal una variable predictora adicional que es combinación lineal de la que se tenía previamente y, finalmente, se evalúa el impacto que esto tiene en los resultados de los índices de condición y su respectiva participación en la varianza estimada de cada coeficiente.

5.4.1. Detección de Observaciones Extremas en los Datos Originales

Teniendo en cuenta que los indicadores utilizados por Belsley, Kuh and Welsch (1982) para diagnosticar presencia de datos extremos en un modelo lineal están basadas simultáneamente en la variable respuesta y en las variables regresoras utilizadas en dicho modelo, al utilizar estos mismos indicadores sobre la variable respuesta transformada y las variables pseudo regresoras utilizadas en el modelo linealizado obtenido a partir del modelo no lineal, tales indicadores pasan a depender, adicionalmente, de la forma no lineal del modelo, al ser W, la matriz de pseudo diseño, función de las derivadas parciales de dicha función no lineal.

Adicional a esto, al ser la fila i de la matriz W función de la fila i de X, una observación (o fila) extrema en el conjunto de datos originales, puede también serlo en el conjunto de datos transformados.

Tomando en consideración lo anterior y aunque, como se explicó anteriormente, las observaciones extremas detectadas por estos indicadores pueden no ser atípicas respecto a los datos originales que componen el modelo de regresión no lineal (en cambio, sí lo son con respecto a las realizaciones de la matriz de diseño W y de la variable independiente y^* que conforman el modelo lineal dado por $y^* = W\beta + \xi$ y sobre el cual se estima por mínimos cuadrados ordinarios el vector de parámetros β), es interesante analizar hasta qué punto la metodología detecta observaciones extremas respecto al conjunto de datos originales.

Con el fin de evaluar el desempeño de la metodología ante situaciones en las cuales se evidencia la presencia de observaciones atípicas en los datos originales, se propone incorporar artificialmente una serie de observaciones extremas en los datos de Stromberg presentados en el este capítulo, es decir, agregar a las 29 mediciones sobre los lagos de la Florida un conjunto de observaciones que sean evidentemente extremas respecto al comportamiento general de los datos. Estas observaciones se agregan una a la vez, calculándose para cada caso las medidas de diagnóstico. En este caso, se propone agregar observaciones que se van alejando del vector de medias con un tamaño de paso de 0.25 desviaciones estándar, iniciando con una distancia de 0.25 desviaciones estándar y terminando con una distancia de 3. De esta forma, es posible determinar a partir de qué momento cuáles de estas medidas empiezan a sobrepasar los límites propuestos por Belsley, Kuh and Welsch (1982), lo cual permite evaluar qué tan sensible es la metodología a marcar estos valores como atípicos.

El vector de medias y de desviaciones estándar de las 29 mediciones hechas sobre los lagos es presentado en la Tabla 5.18.

	TN	**NIN**	**TW**
Media	1.9238	4.5661	0.7257
Desv Estándar	1.1242	6.4994	0.8939

Tabla 5-18.: Medidas descriptivas Stromberg.

Cuando se calculan las medidas de diagnóstico con los datos originales, estos son los resultados

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.26	**-0.83**	**-0.93**	**1.22**	0.41	**0.26**
2	−0.23	−0.23	0.30	**0.56**	0.03	0.01
16	−0.05	0.20	0.21	**1.26**	0.02	**0.16**
23	**7.25**	**4.15**	**-7.79**	**2.45**	**20.51**	**0.81**

En este caso, las observaciones 1, 2, 16 y 23 son marcadas como extremas. La observación 1 presenta altos valores en sus respectivos DFBETASβ, DFFITS y h_i, las observaciones 2 y 16 presentan COVRATIO altos, mientras que la observación 23 es marcada claramente por todos los indicadores. Esta observación 23 también fue calificada como extrema por Stromberg.

Estas observaciones se presentan en la Tabla 5.19.

	TN	**NIN**	**TW**
Observación 1	2.59	5.54	0.14
Observación 2	3.77	4.89	2.50
Observación 16	4.77	5.01	0.10
Observación 23	1.97	34.32	1.50

Tabla 5-19.: Observaciones extremas en datos originales.

Dato a 0.25 sd

Este dato se genera sumando dentro de cada entrada del vector de medias 0.25 veces la desviación estándar asociada a dicha entrada (o variable). Por ejemplo, si la media de la variable TN es 1.9238 y su desviación estándar es 1.1242, el nuevo valor añadido artificialmente para esta variable es $1.9238 + 0.25(1.1242) = 2.2049$. Análogamente, este es el procedimiento utilizado para añadir las demás observaciones a diferentes desviaciones estándar de distancia. Al agregar una nueva observación que dista 0.25 desviaciones estándar del vector de medias, los resultados obtenidos, mostrados a continuación, no parecen variar demasiado.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.28	−0.87	**-0.96**	1.19	0.44	**0.25**
2	−0.22	−0.23	0.30	**0.57**	0.03	0.01
16	−0.04	0.17	0.18	**1.25**	0.02	0.15
23	**6.82**	**3.86**	**-7.42**	**2.10**	**18.39**	**0.79**

Dato a 0.5 sd

La incorporación de una nueva observación a 0.5 desviaciones estándar del vector de medias, tampoco provoca un cambio drástico en los indicadores.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.29	−0.85	**-0.93**	1.21	0.41	**0.26**
2	−0.22	−0.22	0.30	**0.58**	0.03	0.01
16	−0.05	0.19	0.20	**1.26**	0.02	0.16
23	**6.67**	**3.69**	**-7.24**	**1.93**	**17.43**	**0.77**

Dato a 0.75 sd

Los indicadores tampoco varían demasiado con la incorporación de una nueva observación a 0.75 desviaciones estándar del vector de medias. Solo llama la atención que la observación 1 empieza a mostrar un COVRATIO que la marca como extrema.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.29	−0.81	**-0.89**	**1.23**	0.38	**0.26**
2	−0.22	−0.22	0.29	**0.59**	0.03	0.01
16	−0.06	0.21	0.23	**1.26**	0.03	0.16
23	**6.61**	**3.55**	**-7.19**	**1.77**	**16.79**	**0.76**

Dato a 1 sd

Adicionando una nueva observación a una desviación estándar del vector de medias, no se alteran de manera notable los resultados.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.28	−0.76	**-0.84**	**1.25**	0.34	**0.26**
2	−0.21	−0.21	0.29	**0.60**	0.03	0.01
16	−0.07	0.23	0.25	**1.27**	0.03	0.17
23	**6.60**	**3.44**	**-7.16**	**1.63**	**16.31**	**0.75**

Dato a 1.25 sd

Los indicadores tampoco varían demasiado con la incorporación de una nueva observación a 1.25 desviaciones estándar del vector de medias. Solo llama la atención que la observación 1 deja de tener un DFFITs alto, por lo cual deja de ser influyente en el cálculo de las predicciones del modelo.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.28	−0.71	−0.79	**1.27**	0.34	**0.26**
2	−0.21	−0.21	0.28	**0.62**	0.03	0.01
16	−0.08	0.26	0.28	**1.27**	0.03	0.18
23	**6.59**	**3.34**	**-7.14**	**1.63**	**15.91**	**0.75**

Dato a 1.5 sd

La adición de una observación a 1.5 desviaciones estándar del vector de medias no produce tampoco un cambio significativo en los resultados de los indicadores.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.27	−0.73	−0.73	**1.29**	0.26	**0.26**
2	−0.21	−0.20	0.28	**0.63**	0.03	0.01
16	−0.10	0.28	0.30	**1.27**	0.05	0.18
23	**6.59**	**3.24**	**-7.12**	**1.38**	**15.51**	**0.73**

Dato a 1.75 sd

Al adicionar una observación a 1.75 desviaciones estándar del vector de medias, los resultados no varían significativamente con respecto a los anteriores.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.26	−0.59	−0.66	**1.31**	0.22	**0.25**
2	−0.20	−0.19	0.27	**0.65**	0.03	0.01
16	−0.11	0.31	0.33	**1.27**	0.06	0.18
23	**6.58**	**3.13**	**-7.10**	**1.27**	**15.07**	**0.72**

Dato a 2 sd

La inclusión en los datos originales de una observación artificial a dos desviaciones estándar del vector de medias hace que ésta sea marcada como extrema por la metodología pues presenta un alto DFFITs, lo cual significa que tiene una alta incidencia en las predicciones del modelo. En este caso, la metodología, diseñada para marcar observaciones extremas en los pseudo datos del modelo linealizado, es también capaz de identificar esta observación atípica en los datos del modelo no lineal.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.24	−0.53	−0.60	**1.32**	0.18	**0.25**
2	−0.20	−0.19	0.27	**0.67**	0.03	0.01
16	−0.12	0.33	0.36	**1.27**	0.07	0.19
23	**6.55**	**3.00**	**-7.05**	1.17	**14.58**	**0.71**
30	−0.69	−0.68	**0.95**	0.82	0.38	0.13

Las observaciones extremas que resultan de la agregación de esta nueva observación se presentan en la Tabla 5.20.

	TN	**NIN**	**TW**
Observación 1	2.59	5.54	0.14
Observación 2	3.77	4.89	2.50
Observación 16	4.77	5.01	0.10
Observación 23	1.97	34.32	1.50
Observación 30	6.54	31.24	4.40

Tabla 5-20.: Observaciones extremas (incluido dato artificial a 3 sd del vector de medias).

Dato a 2.25

La inclusión de una observación artificial a 2.25 desviaciones estándar del vector de medias no provoca un cambio sustancial con respecto a los resultados anteriores, aunque este nuevo dato incluido empieza a ser marcado como extremo también por el COVRATIO, lo que significa que tiene un alto impacto en el cálculo de la matriz de covarianza de los coeficientes estimados.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.22	−0.47	−0.53	**1.33**	0.14	**0.24**
2	−0.19	−0.18	0.26	**0.69**	0.03	0.01
16	−0.14	0.36	0.39	**1.26**	0.08	0.19
23	**6.50**	**2.85**	**-6.99**	1.06	**14.00**	**0.69**
30	−0.79	−0.84	**1.13**	**0.78**	0.52	0.15

Dato a 2.5 sd

Al adicionar una nueva observación a 2.5 desviaciones estándar del vector de medias, siguen marcándose como extremas las mismas observaciones, aunque el dato 30 empieza a ser calificado como extremo también por el DFBETASβ, lo cual puede interpretarse como una alta

incidencia de esta observación en el cálculo de este coeficiente de regresión.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.20	−0.41	−0.46	**1.33**	0.11	**0.23**
2	−0.19	−0.17	0.25	**0.71**	0.03	0.01
16	−0.16	0.38	0.42	**1.26**	0.09	0.19
23	**6.43**	**2.66**	**-6.90**	0.96	**13.34**	**0.67**
30	−0.90	**-1.02**	**1.34**	**0.74**	0.71	0.17

Dato a 2.75 sd

Al adicionar una nueva observación a 2.75 desviaciones estándar del vector de medias, el dato 30 empieza a ser calificado como extremo también por el DFBETASδ y por la distancia de Cook, lo cual indica que esta observación empieza a ser cada vez más extrema con respecto al conjunto de datos originales.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.18	−0.35	−0.39	**1.33**	0.08	**0.22**
2	−0.18	−0.16	0.24	**0.73**	0.02	0.01
16	−0.18	0.41	0.45	**1.24**	0.10	0.19
23	**6.33**	**2.44**	**-6.78**	0.87	**12.57**	**0.65**
30	**-1.03**	**-1.23**	**1.61**	**0.70**	**0.97**	0.20

Dato a 3 sd

Agregando una observación a 3 desviaciones estándar del vector de medias, los indicadores siguen marcando como extremas las observaciones 1, 2, 16, 23 y 30. La observación 30 empieza a ser marcada como extrema por el h_i, lo cual indica que en este punto este dato supera los umbrales propuestos por Belsley, Kuh and Welsch (1982) para todos los indicadores con que se puede diagnosticar la presencia de observaciones extremas en un conjunto de datos utilizados en un modelo de regresión lineal.

ID	DFBETASδ	DFBETASβ	DFFITS	COVRATIO	DistCook	h_i
1	0.16	−0.28	−0.32	**1.33**	0.05	**0.21**
2	−0.17	−0.15	0.23	**0.75**	0.02	0.01
16	−0.21	0.43	0.47	**1.23**	0.11	0.19
23	**6.20**	**2.17**	**-6.63**	**0.77**	**11.69**	**0.63**
30	**-1.18**	**-1.50**	**1.93**	**0.67**	**1.33**	**0.23**

Con fines ilustrativos se presenta un gráfico en el cual se muestran los datos de Stromberg originales junto con la rejilla de datos artificiales que son añadidos, uno a la vez, para determinar a partir de qué momento las medidas propuestas por Belsley, Kuh and Welsch (1982) los marcan como datos atípicos. Dichos datos se presentan en la Figura **5-8**.

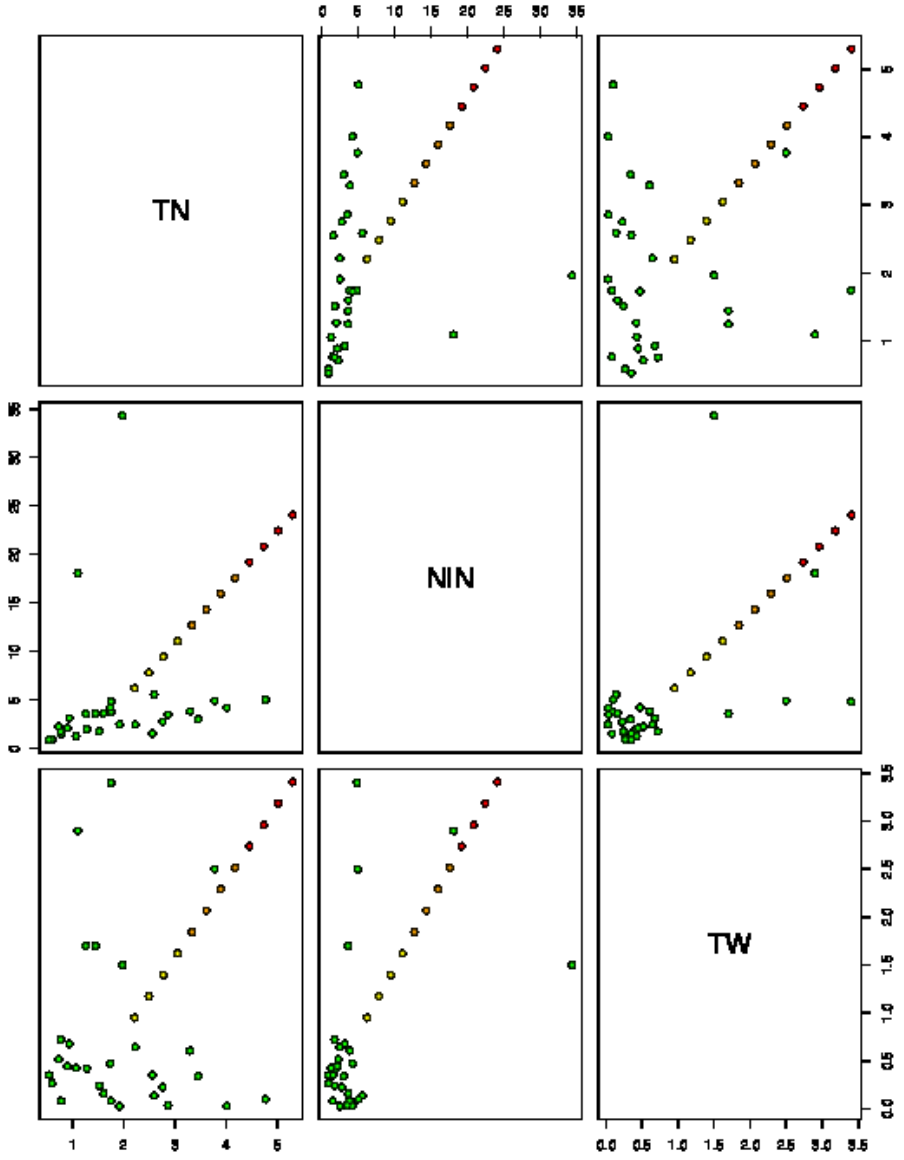

Figura 5-8.: Matriz de Dispersión Variables Pseudo Regresoras

Los datos originales son marcados con verde, las observaciones distanciadas del vector de medias entre 0.25 y 1 desviación estándar se encuentran en amarillo, aquellas entre 1.25 y 2 desviaciones estándar se marcan con naranjado y, finalmente, aquellas observaciones entre

2.25 y 3 desviaciones estándar son pintadas de rojo. Observe que las observaciones amarillas, en general, parecieran hacer parte de los datos originales y, en efecto, cuando la observación 30 está a máximo 1 desviación estándar del vector de medias, ninguna de las medidas de diagnóstico la marca como atípica. Por otra parte, las observaciones en color naranja tienden ya a diferenciarse de los datos originales aunque aún no alcanzan una distancia suficiente para ser marcadas como atípicas por las medidas de diagnóstico utilizadas. Finalmente, las observaciones marcadas con rojo sí presentan un claro distanciamiento de los datos originales, lo cual es detectado por las medidas de diagnóstico que marcan como atípicos aquellos datos distanciados al menos 2 desviaciones estándar del vector de medias.

A manera de conclusión, en este caso, la presencia en los datos originales de una observación a dos desviaciones estándar del vector de medias hace que la metodología empiece a marcar a dicha observación como extrema debido a que uno de los indicadores utilizados la marca como tal, mientras que una observación a una desviación estándar adicional del vector de medias, es marcada ya como extrema claramente por todos los indicadores utilizados en este trabajo. En este caso, la metodología, pese a estar diseñada para marcar datos extremos en el modelo linealizado, debido a la asociación que existe entre los insumos de éste (variables respuesta y matriz de diseño) y los del modelo no lineal, es también capaz de marcar observaciones extremas este último.

5.4.2. Detección de Dependencias Lineales en los Datos Originales

La metodología utilizada por Belsley, Kuh and Welsch (1982) para diagnosticar problemas de multicolinealidad en la estimación de un modelo lineal está enfocada exclusivamente en las propiedades numéricas de la matriz de diseño X. Para evaluar el impacto que este tipo de problemas generan en la precisión de la estimación de los coeficientes de regresión de un modelo no lineal debe evaluarse la matriz de diseño W del modelo lineal a partir del cual dichos coeficientes son obtenidos por mínimos cuadrados ordinarios. Teniendo en cuenta que las columnas de W se calculan como las derivadas parciales de la función no lineal evaluadas en la matriz de diseño original X, estas dos matrices de diseño podrían tener algún tipo de relación entre sí. Si esto fuese verdad, significaría que posibles relaciones de dependencia lineal en la matriz original X podrían conllevar a situaciones similares en la matriz W, lo cual supondría que la metodología, al detectar problemas de multicolinealidad en la matriz W, podría también detectarlos indirectamente en la matriz X.

Teniendo en cuenta lo anterior y aunque, como se explicó anteriormente, las relaciones de dependencia lineal detectadas por la metodología pueden no existir como tal en la matriz de diseño utilizada en el modelo de regresión no lineal (en cambio, sí existen en la matriz de diseño W utilizada para estimar por mínimos cuadrados ordinarios el vector de parámetros

β), es interesante analizar hasta qué punto la metodología detecta relaciones de dependencia lineal en las variables regresoras originales.

Con el propósito de analizar el comportamiento de la metodología ante una situación en la cual se presenta una relación de dependencia lineal exacta en la matriz de diseño original, se hace una modificación en una de las aplicaciones presentadas en el este capítulo, de tal manera que se incorpora en el modelo no lineal una variable regresora adicional linealmente dependiente de la ya existente. De esta forma, partiendo de una evidente relación de dependencia lineal en la matriz de diseño original, se aplica la metodología con el fin de determinar si esta dependencia tiene alguna incidencia en la precisión de las estimaciones de los coeficientes de regresión, es decir, si este problema de multicolinealidad se transmite a la matriz de diseño W del modelo lineal.

Así pues, se toman los resultados del estudio hecho por NIST (National Institute of Standards and Technology) que involucra calibración ultrasónica. La variable respuesta es la respuesta ultrasónica y las variable predictoras son la distancia del metal y una combinación lineal de esta (Chwirut and NIST, 1979).

El modelo no lineal, en este caso, está dado por la relación

$$y = \frac{e^{-\beta_1 x_1}}{\beta_2 + \beta_3 x_2} + \xi$$

Donde β_1, β_2 y β_3 son parámetros desconocidos.

Las 54 observaciones medidas en estas dos variables se encuentran originalmente en la librería NISTnls de R bajo el nombre Chwirut2. Algunos de estos datos se presentan en la Tabla 5.21.

Observe que la variable predictora incorporada artificialmente se calcula como dos veces la variables predictora original.

Análisis Exploratorio sobre Variables Originales

Algunas medidas descriptivas para estas variables se presentan en la Tabla 5.22.

Cabe destacar que no se tienen coeficientes de variación llamativamente pequeños.

Estimación Modelo No Lineal

ID	RespUltr	DistMetal	2 x DistMetal
1	92.9000	0.500	1.000
2	57.1000	1.000	2.000
3	31.0500	1.750	3.500
4	11.5875	3.750	7.500
5	8.0250	5.750	11.500
6	63.6000	0.875	1.750
7	21.4000	2.250	4.500
8	14.2500	3.250	6.500
9	8.4750	5.250	10.500
10	63.8000	0.750	1.500
11	26.8000	1.750	3.500
12	16.4625	2.750	5.500
13	7.1250	4.750	9.500
14	67.3000	0.625	1.250
15	41.0000	1.250	2.500
...	...	...	...

Tabla 5-21.: Primeros quince datos de Chwirut2 con respuesta ultrasónica, distancia del metal y distancia del metal doble.

	RespUltr	DistMetal	2 x DistMetal
Media	31.5641	2.5463	5.0926
Desv Estándar	26.3130	1.7034	3.4069
Coef. Variación	0.8837	0.6690	0.6690
Mínimo	3.7500	0.5000	1.0000
Máximo	92.9000	6.0000	12.0000

Tabla 5-22.: Medidas descriptivas para respuesta ultrasónica, distancia del metal y distancia del metal doble.

Luego de ajustar el modelo no lineal propuesto, se estiman sus respectivos parámetros. Las estimaciones puntuales junto con sus errores estándar, sus valores t y sus respectivas pruebas de significancia univariadas, se presentan en la Tabla 5.23.

Según las pruebas de significancia individuales, los parámetros considerados en el modelo son estadísticamente significativos.

Variables Pseudo - Regresoras

Ahora bien, el modelo lineal toma como insumo una nueva matriz de diseño, denotada W,

	Estimación	**Error Estándar**	**Valor t**	**Valor p**
β_1	0.0833	0.0192	4.3490	$6.56e-05$
β_2	0.0052	0.0007	7.7530	$3.54e-10$
β_3	0.0122	0.0015	7.9390	$1.81e-10$

Tabla 5-23.: Parámetros estimados y sus errores estándar asociados en el modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble.

cuyas columnas son las derivadas parciales de la función no lineal con respecto a cada uno de los parámetros a estimar. Esta matriz W tiene la siguiente estructura.

$$W = [W_1, W_2, W_3]_{\beta=\hat{\beta}}$$

donde W_j $(j = 1, 2, 3)$ son vectores de dimensión n que tienen en su entrada i $(i = 1, 2, 3, ..., n)$ la respectiva derivada parcial evaluada en la realización de la observación i.

En este caso, se tiene:

$$W_1 = \frac{\partial f(x;\beta)}{\partial \beta_1} = \frac{-x_1 e^{-\beta_1 x_1}}{(\beta_2 + \beta_3 x_2)}$$

$$W_2 = \frac{\partial f(x;\beta)}{\partial \beta_2} = \frac{-e^{-\beta_1 x_1}}{(\beta_2 + \beta_3 x_2)^2}$$

$$W_3 = \frac{\partial f(x;\beta)}{\partial \beta_3} = \frac{-x e^{-\beta_1 x_1}}{(\beta_2 + \beta_3 x_2)^2}$$

Al evaluar estas derivadas en los parámetros estimados y en la matriz de diseño original X, se obtiene la matriz de diseño que sirve como insumo al modelo lineal, la cual contiene unas nuevas variables regresoras, conocidas como pseudo regresoras. Algunas filas de esta matriz se presentan en la Tabla 5.24.

Análisis Exploratorio sobre Variables Pseudo - Regresoras

A continuación, se hace un análisis exploratorio sobre las variables pseudo regresoras contenidas en las columnas de W. Los resultados se presentan en la siguiente Tabla 5.25.

ID	W_1	W_2	W_3
1	−81.8557	−7282.3234	−3641.1617
2	−97.7813	−2823.5440	−2823.5440
3	−98.9479	−1069.7332	−1872.0331
4	−79.1638	−208.0747	−780.2802
5	−58.8166	−68.1680	−391.9662
6	−95.7579	−3463.9605	−3030.9655
7	−95.1741	−650.7055	−1464.0873
8	−84.7126	−291.8674	−948.5690
9	−63.5087	−87.7188	−460.5239
10	−92.7196	−4329.2996	−3246.9747
11	−98.9477	−1069.7332	−1872.0331
12	−90.1738	−424.9908	−1168.7248
13	−68.4848	−114.6499	−544.5869
14	−88.2828	−5535.3659	−3459.6037
15	−99.7438	−1960.2912	−2450.3641
...	...	...	...

Tabla 5-24.: Tabla de variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble.

	W_1	W_2	W_3
Media	−84.9950	−1928.7245	−1775.1354
Desv Estándar	12.4962	2439.3633	1151.9034
Coef. Variación	−0.1470	−1.2648	−0.6489
Mínimo	−99.9010	−7282.3234	−3641.1617
Máximo	−56.5798	−60.3981	−362.3886

Tabla 5-25.: Medidas descriptivas para variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble.

Ninguna de las variables presenta un coeficiente de variación llamativamente pequeño, lo cual hace pensar que es poco probable que se presenten problemas de mala adecuación debido a poca variación en alguna de las variables.

Búsqueda de Multicolinealidad

La búsqueda de multicolinealidad en la matriz W inicia con la exploración de posibles relaciones lineales entre diferentes pares de sus columnas. Aunque este análisis no detecta relaciones de dependencia lineal en la que participen simultáneamente más de dos variables,

es un buen inicio para identificar patrones sospechosos que luego han de ser corroborados con un análisis más riguroso. Una medida que cuantifica la cantidad de relación lineal existente entre cada par de variables es, por supuesto, la matriz de correlación, la cual se presenta en la Tabla 5.26.

	W_1	W_2	W_3
W_1	1.0000	0.2116	0.5083
W_2	0.2116	1.0000	0.9363
W_3	0.5083	0.9363	1.0000

Tabla 5-26.: Matriz de correlación entre variables pseudo regresoras utilizadas en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble.

Una prueba más rigurosa para detectar relaciones de dependencia lineal, presentada en este trabajo, consiste en calcular los índices de condición de la matriz W y asociar a cada uno de ellos una proporción de varianza de cada uno de los parámetros estimados. Grandes índices de condición están asociados a relaciones de dependencia lineal aproximadas, siendo cada uno de ellos producto de una relación de dependencia. Adicionalmente, grandes proporciones de varianza en algunos parámetros asociadas a un índice de condición grande, significan que en esa relación de dependencia están involucradas las variables asociadas a dichos parámetros.

Los resultados correspondientes a este ejemplo, se presentan en la Tabla 5.27.

Índice de Condición	$var(\hat{\beta}_1)$	$var(\hat{\beta}_2)$	$var(\hat{\beta}_3)$
1.000	0.013	0.005	0.005
1.620	0.165	0.007	0.000
13.359	0.821	0.987	0.995

Tabla 5-27.: Índices de condición y descomposición de varianza para parámetros estimados en linealización de modelo no lineal para respuesta ultrasónica, distancia del metal y distancia del metal doble.

Se observan grandes proporciones de varianza de cada uno de los parámetros estimados asociadas a un solo índice de condición. No obstante, según los límites propuestos por Belsley, Kuh and Welsch (1982), este índice de condición no es lo suficientemente grande para considerar que haya un problema de multicolinealidad en los datos. De esta manera, para este caso particular, la presencia de una relación de dependencia lineal en la matriz de diseño ori ginal no se traduce en problemas de multicolinealidad en el modelo lineal. En otras palabras, si bien la matriz de diseño utilizada en el modelo no lineal presenta columnas linealmente

dependientes entre sí, esto no conlleva necesariamente a problemas de precisión en la estimación de los coeficientes de dicho modelo, ya que el modelo lineal mediante el cual se estiman éstos por mínimos cuadrados ordinarios no se contagia de dichas dependencias.

A manera de conclusión, la metodología, en general, no detecta problemas de multicolinealidad en los datos originales que se modelan de forma no lineal, lo cual no es grave, puesto que dichas dependencias no se transmiten al modelo lineal sobre el cual se estiman por mínimos cuadrados ordinarios los parámetros del modelo no lineal, lo que significa que estas dependencias iniciales no provocan una pérdida de precisión en la estimación de los parámetros de interés.

6. Conclusiones

Los indicadores utilizados por Belsley, Kuh and Welsch (1982) para diagnosticar presencia de datos influyentes o atípicos en un modelo lineal están basados tanto en la variable respuesta como en el set de variables regresoras. Al utilizar estos mismos indicadores sobre la variable respuesta transformada y las variables pseudo regresoras utilizadas en el modelo linealizado obtenido, mediante expansión de Taylor, a partir del modelo no lineal, estos indicadores pasan a depender, adicionalmente, de la forma no lineal del modelo, al ser $\mathbf{W}$, la matriz de pseudo diseño, función de las derivadas parciales de ésta. Adicional a esto, al ser la fila i de la matriz $\mathbf{W}$ función de la fila i de $\mathbf{X}$, una observación (o fila) extrema en el conjunto de datos originales, puede también serlo en el conjunto de datos transformados. Cabe recordar que los procedimientos mostrados en este trabajo están basados de manera exclusiva en la eliminación de una observación a la vez, siendo por este motivo incapaces de detectar observaciones extremas en presencia, por ejemplo, de enmascaramiento. Este tipo de problemas deben ser abordados utilizando procedimientos basados en técnicas de eliminación de más de una sola observación a la vez.

La metodología descrita por Belsley, Kuh and Welsch (1982) para diagnosticar problemas de multicolinealidad en la estimación de un modelo lineal, se basa únicamente en las propiedades numéricas de la matriz de diseño $\mathbf{X}$. Análogamente, al aplicar esta metodología en el caso del modelo no lineal, luego de realizar la respectiva aproximación lineal, el análisis debe centrarse en las propiedades de la matriz pseudo regresora $\mathbf{W}$. De esta forma, al estar $\mathbf{W}$ estrechamente ligada con la forma funcional del modelo no lineal a través de sus derivadas parciales, la metodología finalmente podría terminar basándose también en la forma como se relacionan las variables regresoras originales con la variable respuesta de interés. Por otra parte, al detectar problemas asociados a multicolinealidad en el modelo no lineal, se recomienda hacer especial énfasis en analizar si estos son de tipo nocivo o degradante, tal como se explica en la sección 3.10 de este trabajo, a fin de tener un panorama más claro respecto a cómo deben ser tratados.

Además de las posibles causas, mencionadas en el Capítulo 3, que pueden llevar a problemas de multicolinealidad en un modelo lineal, en el caso no lineal tener un modelo inadecuadamente especificado puede ser por sí misma otra causa adicional del problema. Un modelo sobre especificado, aunque puede estar bien justificado teóricamente, tal como se observa en una de las aplicaciones, puede llevar a serios problemas de multicolinealidad al crear rela-

ciones lineales matemáticamente perfectas entre las variables pseudo regresoras utilizadas en la linealización del mismo. Algunas veces, como se muestra en la misma aplicación, este tipo de problemas pueden ser resueltos realizando una adecuada re especificación del modelo sin que esto repercuta en una pérdida sustancial de su interpretación.

La implementación de una interfaz gráfica fácil de manejar y con la versatilidad suficiente para recibir como insumos una tabla de datos y una forma funcional no lineal es de valor agregado para los usuarios de regresión no lineal ya que les permite hacer un completo diagnóstico enfocado en detectar la presencia de observaciones extremas o relaciones de multicolinealidad en sus datos, siendo estos dos problemas causantes directos de efectos nocivos en el proceso de estimación de cualquier modelo estadístico.

A. Anexo: Interfaz Gráfica en R

Los procedimientos descritos en este trabajo para diagnosticar presencia de datos extremos y/o problemas de multicolinealidad en modelos de regresión no lineal son implementados en una interfaz gráfica creada con la librería shiny de R.

En este anexo se explica detalladamente cómo debe utilizarse esta aplicación ya sea para detectar datos extremos, para descubrir problemas de multicolinealidad o para realizar ambos procedimientos en un modelo no lineal específico y con una tabla de datos predeterminada.

Inicialmente es importante resaltar que la librería shiny está disponible para ser descargada en una versión del R igual o superior a la 2.14.1. Al abrir la pantalla del R, disponiendo de conexión a la web, esta librería puede instalarse con el siguiente comando:

install.packages(shiny)

Una vez instalada la librería, se recomienda remover los objetos que R pueda tener almacenados en su memoria con el fin de evitar posibles interferencias entre estos y los nuevos que se crearán durante la ejecución de la aplicación. A continuación, se debe invocar la librería recientemente instalada. Estos dos pasos se ejecutan, respectivamente, con los siguientes comandos:

rm(list=ls(all=TRUE))
library(shiny)

Ahora bien, cumpliendo en este momento con todos los requerimientos iniciales mínimos para ejecutar una interfaz gráfica diseñada mediante shiny, se deben tener en cuenta algunos aspectos fundamentales: shiny funciona haciendo un llamado al directorio de trabajo del R, en el cual deben estar almacenados, dentro de una carpeta titulada con el mismo nombre de la aplicación (en este caso 'ExtrMult-app'), ciertos archivos que incluyen, entre otros, las tablas de datos, los programas con que se calculan los resultados y se construyen los gráficos mostrados en las salidas de la aplicación y otros archivos que permiten personalizar el diseño gráfico de la aplicación y las tablas e imágenes mostradas en ella. Es de vital importancia tener en cuenta que esta carpeta junto con un instalador del R 3.1.0 se entregan en un disco anexo a una copia impresa de este trabajo.

Teniendo en cuenta estos aspectos, la aplicación debe ser llamada desde la consola del R a través del siguiente comando:

runApp("ExtrMult-app")

Tenga en cuenta que el argumento de la función es, encerrado entre comillas, el nombre de la carpeta en la cual se encuentran almacenados los archivos que requiere la aplicación.

Una vez ejecutada esta línea, se abre una página de navegación en la cual se visualiza la aplicación en su estado inicial, tal como se muestra en la Figura **A-1**.

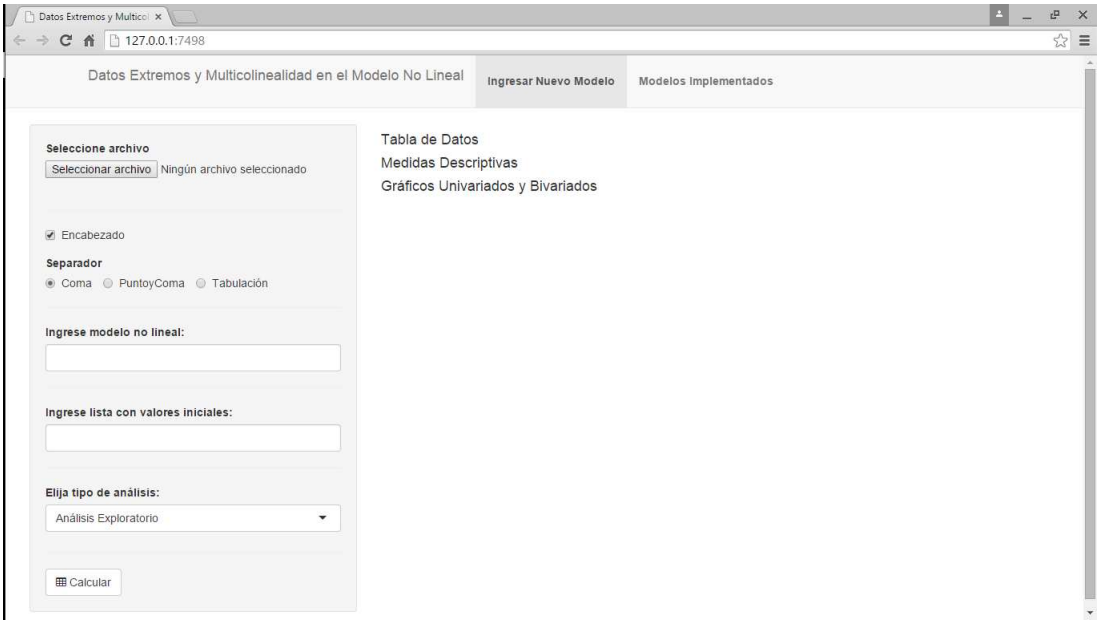

Figura A-1.: Aplicación en blanco

Como se observa en la parte superior de la Figura **A-1**, esta aplicación consta de dos pestañas, llamadas "Ingresar Modelo Nuevo" y "Modelos Implementados", que cumplen funciones diferentes.

En la pestaña "Ingresar Modelo Nuevo", que está seleccionada por defecto, se permite ingresar una nueva tabla de datos y especificarle un modelo no lineal junto con valores iniciales para cada uno de los parámetros a estimar.

La tabla de datos se lee con el botón "Seleccionar archivo", el cual está en la parte superior de la pantalla gris que aparece en el costado izquierdo de la ventana. La aplicación permite

leer tablas de datos cuyas columnas estén separadas por comas, punto y comas o tabulaciones. Al leer la tabla de datos, en este caso llamada "DatosPrueba.txt", y especificar el separador de columnas adecuado, en este caso "Tabulación", en la pantalla aparecerá un encabezado con las primeras seis filas de la tabla, algunas medidas descriptivas para cada variable y algunos gráficos que resumen, individualmente y por pares, el comportamientos de las variables. La forma como se presentan estos resultados se muestra en la Figura **A-2**.

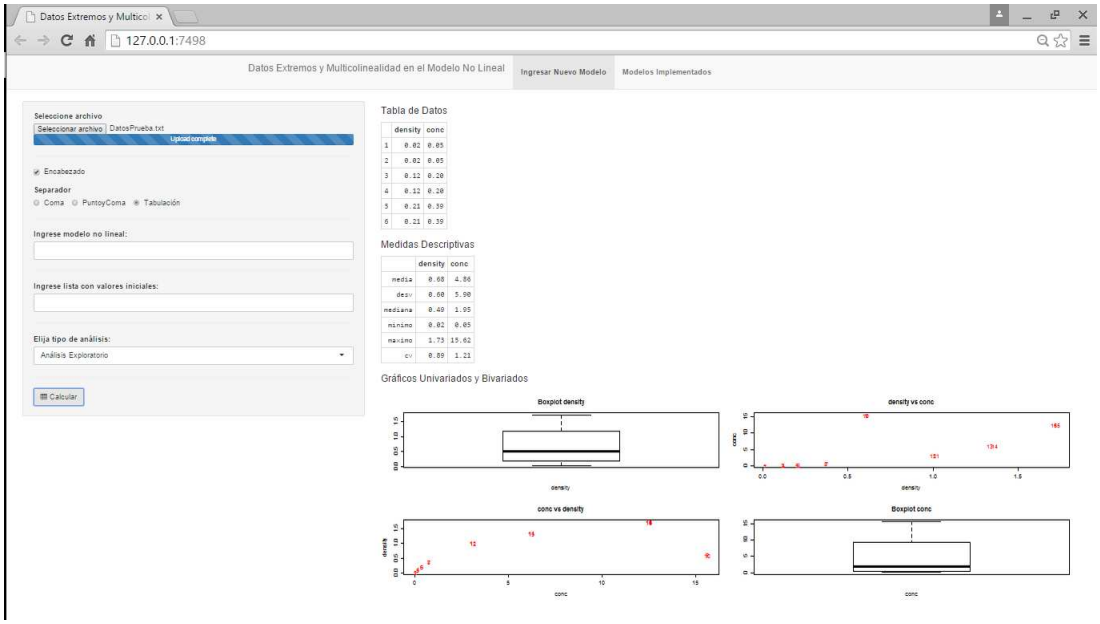

Figura A-2.: Tabla de datos y resumen descriptivo

En el mismo recuadro donde se selecciona la tabla de datos, aparecen dos espacios en blanco donde se deben especificar, respectivamente, el modelo no lineal y una lista con valores iniciales para los parámetros que se deben estimar.

Una vez leídos los datos, se deben especificar en los renglones disponibles en blanco tanto el modelo no lineal como los valores iniciales para los parámetros considerados en éste. El modelo no lineal se debe declarar de la misma forma que se hace en R cuando se utiliza la función nls, es decir, la variable respuesta debe ir unida a la función no lineal mediante una virgulilla. Para el ejemplo mostrado aquí, siendo *density* la variable respuesta y $Asym/(1+exp(xmid–log(conc))/scal)$ la función no lineal con *conc* como variable regresora, el modelo no lineal debe ser especificado como $density \tilde{} Asym/(1+exp((xmid-log(conc))/scal))$. Los valores iniciales deben ir dentro de una lista, cada uno asociado al nombre de un parámetro a estimar, el cual debe coincidir plenamente con los nombres utilizados en la especificación del modelo. Para el ejemplo mostrado aquí, siendo los valores iniciales para $Asym$, $xmid$ y $scal$, respectivamente, 3, 0 y 1, éstos deben ser especificados como list($Asym = 3$, $xmid =$

0, $scal = 1$). La Figura **A-3** muestra lo explicado anteriormente.

Figura A-3.: Modelo no lineal y valores iniciales de parámetros a estimar

Una vez especificada adecuadamente la tabla de datos y el modelo, se oprime el botón "Calcular" que aparece en la parte inferior del recuadro.

Al oprimir este botón, se calculan todas las medidas necesarias para diagnosticar presencia de datos extremos en la información o problemas de multicolinealidad en el modelo.

Ahora bien, en el menú desplegable que tiene como título "Elija tipo de análisis", se puede seleccionar alguna de las siguientes opciones: Análisis Exploratorio, Estim. No Lineal, Detección de Datos Extremos o Detección de Multicolinealidad.

La opción "Análisis Exploratorio", que viene seleccionada por defecto, muestra el encabezado de la tabla de datos, un breve análisis descriptivo univariado y los diagramas de dispersión para cada par de variables estudiadas. La forma como aparecen estos resultados en la aplicación para la tabla de datos "DatosPrueba.txt" y el modelo no lineal $density \tilde{} Asym/(1 + exp(xmid – log(conc))/scal))$ se muestra en la Figura **A-2**.

El tipo de análisis "Estim. No Lineal" presenta un resumen detallado de las estimaciones por mínimos cuadrados no lineales obtenidas para modelo no lineal especificado. También presenta el encabezado de una tabla que tiene como columnas la variable respuesta transformada y las variables pseudo regresoras obtenidas a través del proceso de linealización del modelo no lineal. Finalmente muestra el resumen del modelo lineal estimado con esta tabla, que debe coincidir plenamente con los resultados del modelo no lineal. La forma como se

muestran estos resultados se presenta en la Figura **A-4**.

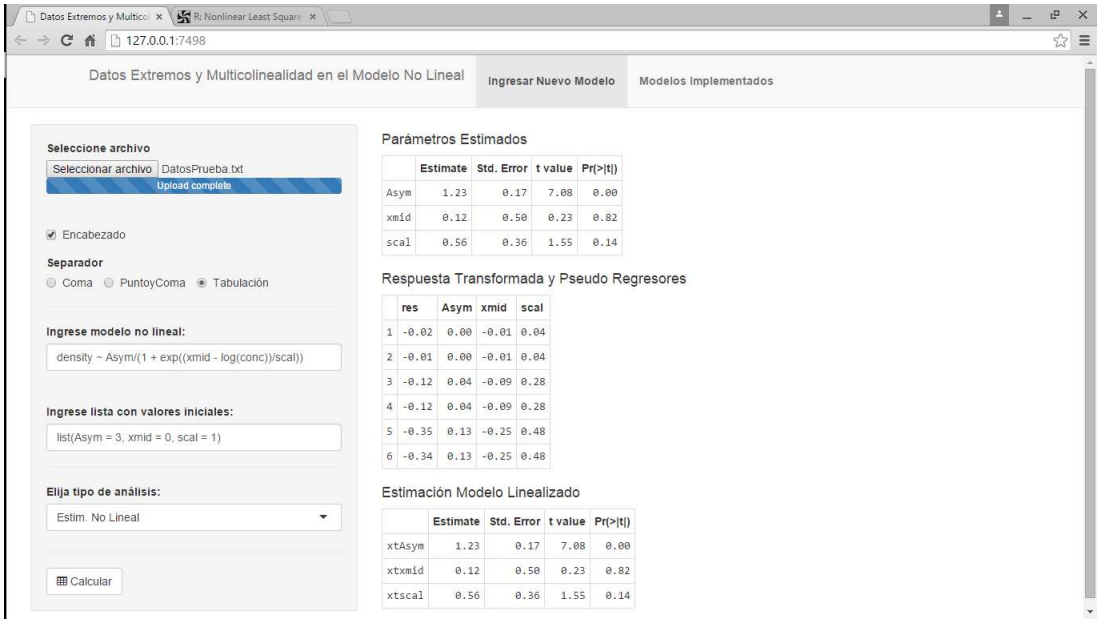

Figura A-4.: Estimación no lineal, variable respuesta transformada y pseudoregresores

Al seleccionar como tipo de análisis la opción "Detección de Datos Extremos", se presenta una tabla con aquellas observaciones cuyo DFBETAs, DFFITs, Radio de Covarianza, Distancia de Cook o hii supere el límite establecido por Belsley, Kuh and Welsch (1982) para ser considerada extrema. Adicionalmente se presentan algunos gráficos de diagnóstico entre los que se incluyen los gráficos de regresión parcial. Las Figuras **A-5**, **A-6** y **A-7** muestran la forma en que se presentan estos resultados en la aplicación para el ejemplo estudiado.

Finalmente, al seleccionar como tipo de análisis la opción "Detección de Multicolinealidad", se presenta la matriz de correlación entre las variables pseudo regresoras y una tabla donde se muestran los índices de condición y la proporción de varianza de cada coeficiente asociada a cada uno de los índices. La forma como se presenta esto se muestra en la Figura **A-8**.

En la pestaña "Modelos Implementados", ubicada justo a la derecha de "Seleccionar Modelo Nuevo", se encuentran disponibles los ejemplos estudiados en el Capítulo 7: Stromberg, Chwirut2 y RaízConc.

Para cada uno de estos ejemplos se pueden obtener todas las medidas necesarias para hacer un análisis sobre detección de datos extremos o diagnóstico de problemas de multicolinealidad. Cada uno de estos análisis se obtiene seleccionando adecuadamente en la pestaña "Elija tipo de análisis" alguna de las opciones disponibles: "Análisis Descriptivo", "Estimación No

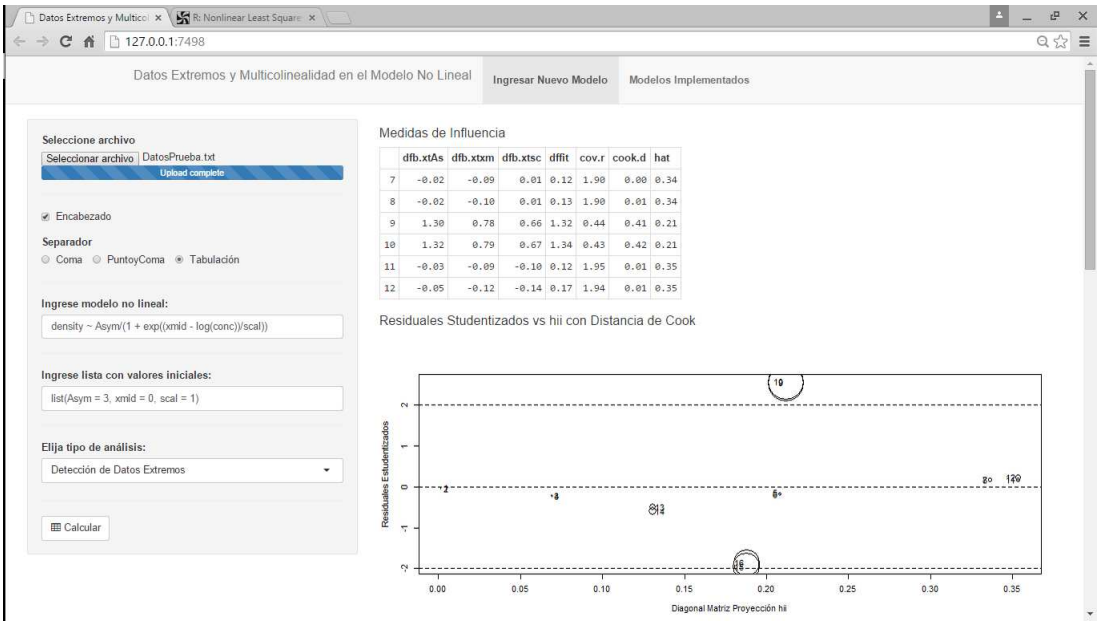

Figura A-5.: Detección Datos Extremos

Lineal", "Detección Datos Extremos" y "Detección Multicolinealidad".

Para que la aplicación se ejecute adecuadamente es indispensable tener en cuenta que al seleccionar, por ejemplo, la tabla de Stromberg, se deben seleccionar únicamente el modelo no lineal y los valores iniciales correspondientes a dicha tabla. Lo mismo debe tenerse en cuenta para los datos Chwirut2 y RaízConc.

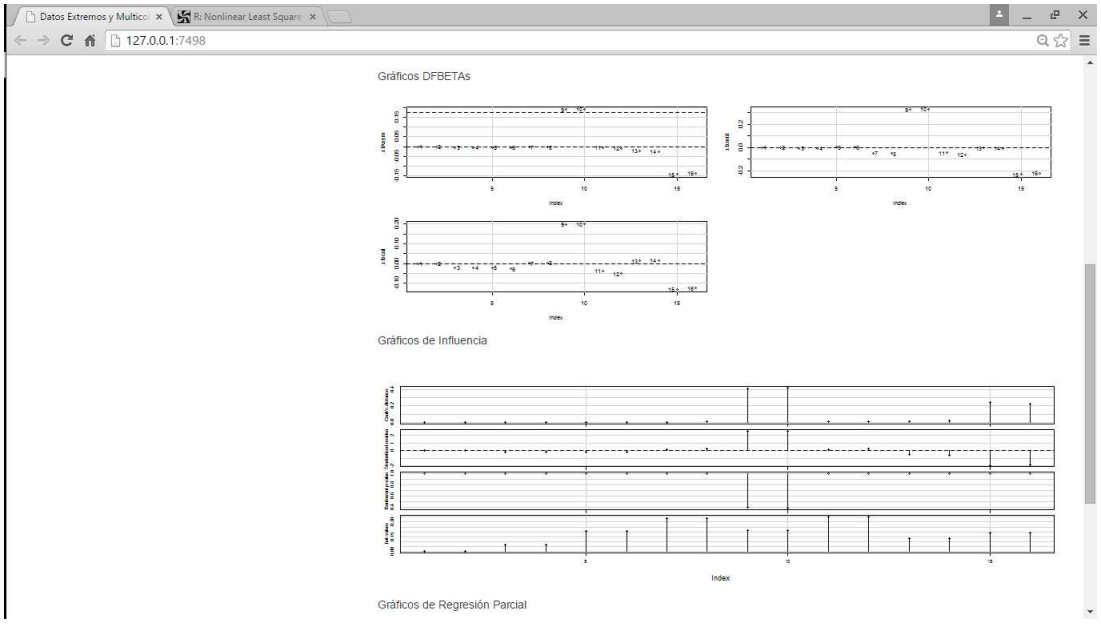

Figura A-6.: Detección Datos Extremos

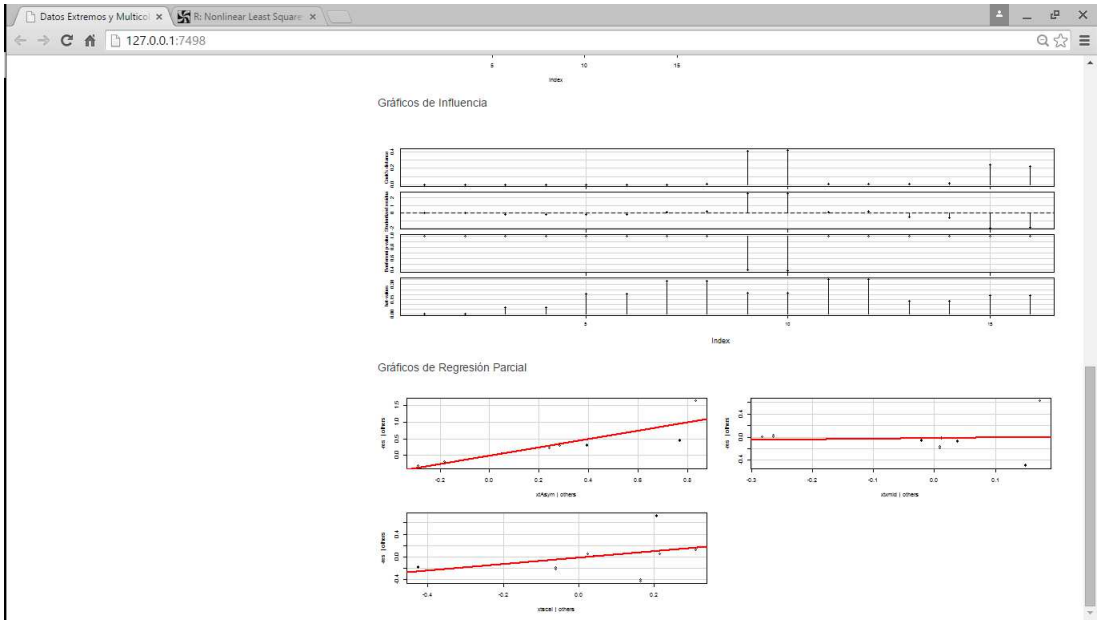

Figura A-7.: Detección Datos Extremos

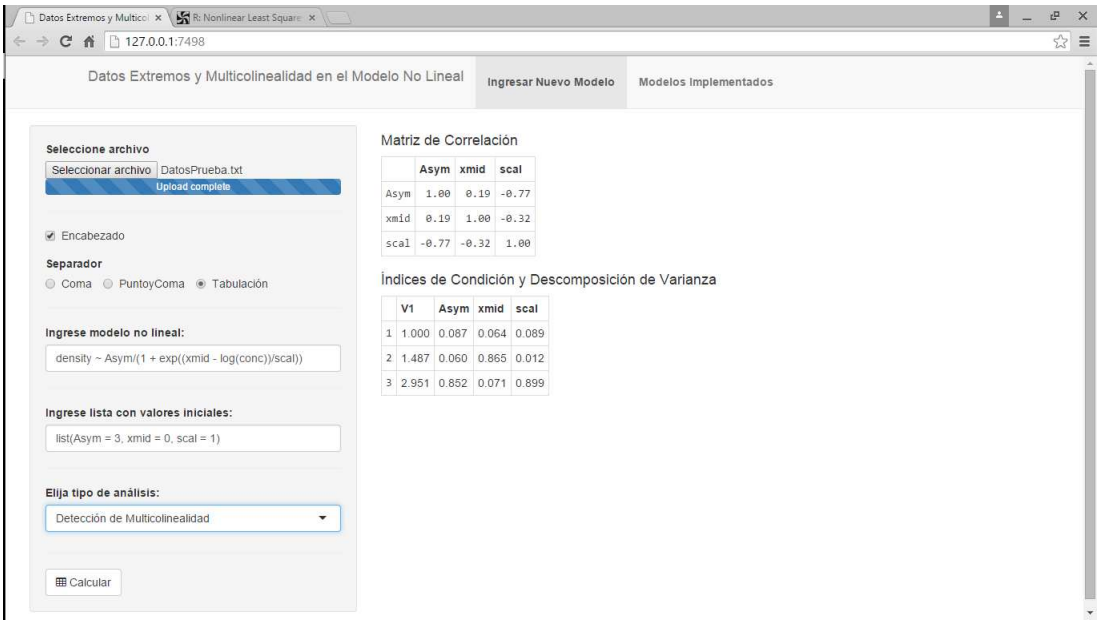

Figura A-8.: Detección Multicolinealidad

Bibliografía

- Bates and Watts (1982), *Nonlinear Regression Analysis and Its Applications*, Wiley Inter-Science, New York.
- Belsley, Kuh and Welsch (1982), *Regression Diagnostics*, Wiley Inter-Science, New Jersey.
- Canavos (1988), *Probabilidad y Estadística: Aplicaciones y Métodos*, McGraw Hill, Madrid.
- Chaterjee and Price (1977), *Regression Analysis by Example*, John Wiley and Sons, New York.
- Chwirut, D., NIST (1979), Ultrasonic Reference Block Study.
- Cook, R. D., (1977), 'Detection of influential observations in linear regression', Technometrics, No. 19, pp. 15-18.
- Cook and Weisberg (1982), *Residuals and influence in regression*, Chapman and Hall, New York.
- Harville, D (1997), *Matrix Algebra from a Statistician's Perspective*, Springer-Werlag, New York.
- Hawkins, D. M. (1973), 'On the Investigation of Alternative Regressions by Principal Component Analysis', Applied Statistics, 22, pp. 275-286.
- Inderjit, Streibig and Olofsdotter (2002), 'Joint action of phenolic acid mixtures and its significance in allelopathy research,' Physiologia Plantarum, Vol. 114, No. 1, pp. 422-428.
- Kendall, M. G. (1957), A course in Multivariate Analysis. Griffin, London
- Kuh and Welsch (1977), 'Linear Regression Diagnostics', Massachusetts Institute of Technology and NBER Computer Research Center.
- Mosteller and Tukey (1977), 'Data Analysis and Regression', Addison-Wesley.
- Novales, A. (1993), *Econometría*, McGraw Hill, Madrid.

- Rao (1973), 'Linear Statistical Inference ands Its Applications', 2nd ed., John Wileyand Sons.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990), 'Unmasking Multivariate Outliers and Leverage Points', Journal of the American Statistical Association, 85, 633 - 639.
- Seber, G.A.F. and Wild, C.J. (2003), *Nonlinear Regression*, Wiley Inter-Science, New Jersey.
- Silvey, S. D. (1969), 'Multicollinearity and Imprecise Estimation', Journal of Royal Statistical Society, Series B, 31, pp. 539-552.
- Stromberg, A. (1993), 'Computation of High Breakdown Nonlinear Regression Parameters', American Statistical Asociation, **88**, 237-244.
- Theil, H. (1971), *Principles of Econometrics*, John Wiley and Sons: New York.