XLVIII Coloquio Argentino de Estadística

VI JORNADA DE EDUCACIÓN ESTADÍSTICA "MARTHA DE ALIAGA"

27 al 30 oct 2020

Poster:

Una aplicación de regresión con métodos robustos en una red altimétrica topográfica

Claudio Eduardo Justo, María Valeria Calandra

FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

SAE

XLVIII Coloquio Argentino de Estadística

UNA APLICACIÓN DE REGRESIÓN CON MÉTODOS ROBUSTOS EN UNA RED ALTIMÉTRICA TOPOGRÁFICA

Claudio Eduardo Justo, María Valeria Calandra (claudio.justo@ing.unlp.edu.ar /mava@mate.unlp.edu.ar)

UIDET Grupo de Aplicaciones Matemáticas y Estadísticas de la Facultad de Ingeniería (UIDET GAMEFI) Departamento de Ciencias Básicas, Facultad de Ingeniería, UNLP.

RESUMEN

Se ajustaron las observaciones de una red altimétrica topográfica mediante Mínimos Cuadrados Ponderados y dos métodos robustos. El ajuste de redes altimétricas topográficas es realizado en forma extendida mediante Mínimos Cuadrados Ponderados. Para la ponderación a cada desnivel se le asigna un peso de 1/L donde L es la longitud del recorrido necesario para obtener cada desnivel medido. En este caso se estudió el comportamiento del ajuste de la red por el método tradicional y por dos métodos robustos. Se encontró que el primer método representa un buen ajuste en el rango los valores atípicos presentes. Dado que las observaciones y los parámetros se encuentran funcionalmente relacionados por condiciones geométricas la presencia de un valor atípico que desvirtúe un ajuste es muy poco probable. Pero siempre que se hace un análisis por Mínimos Cuadrados, y se encuentran valores atípicos, sería conveniente también hacer un ajuste robusto y si los resultados concuerdan, o la diferencia es exigua, se podrían usar los resultados de los estimadores de Mínimos Cuadrados. Sin embargo, si difieren, se deben identificar las razones de tales diferencias .De acuerdo a la comparación de los resultados obtenidos en las condiciones mencionadas puede considerarse a Mínimos Cuadrados Ponderado como un buen ajuste.

INTRODUCCIÓN

El ajuste de redes altimétricas topográficas es realizado en forma extendida mediante el método de Mínimos Cuadrados Ponderados. Para la ponderación a cada desnivel de una red altimétrica topográfica se le asigna para el ajuste un peso de 1/L donde L es la longitud del recorrido necesario para obtener cada medición. Los desniveles son las variables de respuestas y las variables predictoras son coeficientes -1,0,1 por tratarse de una red de Grafos. El modelo altimétrico topográfico, es válido para determinar las diferencias de alturas entre puntos y poder resolver la dirección del escurrimiento de fluidos en el entorno de obras a nivel municipal. En Geodesia al modelo topográfico se lo denomina Sistema de Alturas Geométricas. Para poder dejar establecidas las alturas de marcas físicas colocadas expresamente con ese propósito, y conocidas como ménsulas, es que se realizan mediciones de desniveles entre ellas. El desnivel entre dos ménsulas A y B se obtiene mediante la sumatoria de n desniveles individuales medidos a lo largo del itinerario que se requiere para llegar de una a otra. Estos desniveles obtenidos serán las observaciones que se someterán a un ajuste por Mínimos Cuadrados Ponderados para salvar la inconsistencia debida a los factores aleatorios que están siempre presentes en las mediciones. La estimación por MCP permite obtener indicadores de la calidad del trabajo realizado. Se presentan mediciones realizadas en el campus de la Facultad de Ingeniería de la Universidad Nacional de La Plata (UNLP) donde existe una red altimétrica con ménsulas distribuidas en casi todos los edificios de las distintos Departamentos. (Figura 1). Las redes mencionadas comparten el mismo origen o datum en el Cero del Mareógrafo de la ciudad de Mar del Plata, Argentina.

Figura 1. Esquema de la Red Altimétrica (Campus Facultad de Ingeniería, UNLP)

El análisis de regresión es una de las técnicas estadísticas más empleadas, y dentro del mismo, el método de Mínimos Cuadrados clásico es considerado poco robusto cuando las observaciones no provienen de una distribución normal o hay observaciones atípicas. Los valores atípicos pueden ser causados por sucesos excepcionales o podrían ser resultado de un factor aún no considerado en un estudio. Incluso podría estar sucediendo algo sistemáticamente. Hay circunstancias en el que los datos pueden eliminarse de forma justificada, pero en general, dado que hay observaciones no necesariamente "malas", es razonable concluir que no deben descartarse. Para evaluar la respuesta ante esta situación se introduce una comparación con estimaciones hechas por dos métodos robustos. Para que un estimador de regresión robusto sea de utilidad práctica debe tener punto de ruptura y eficiencia relativa altos. El punto de ruptura es la mínima fracción de datos atípicos que puede causar que el estimador no se útil. Este valor se puede usar como una medida de la robustez del estimador. El punto de ruptura finito de los estimadores mínimos cuadráticos es 1/n, para una muestra de tamaño n, equivale a decir que una sola observación puede distorsionar el estimador. Esto tiene un impacto potencialmente grave sobre su uso práctico. Cuando las observaciones provienen de una Distribución Normal y no hay observaciones atípicas es correcto, además de seguro, utilizar Estimadores Mínimo Cuadráticos (EMC). Además del punto de ruptura, para caracterizar a los estimadores robustos, se define la eficiencia de estos, como el cociente entre el cuadrado medio residual obtenido con los EMC y el cuadrado medio residual obtenido con el procedimiento robusto. Es esperable que esa medida de eficiencia se debe aproximar a 1.

METODOLOGÍA

El modelo lineal clásico relaciona las variables independientes, o respuestas y_i , con las variables dependientes o explicativas $x_{i1}, x_{i2} \ldots . x_{ip}$ para $i = 1,2 \ldots . n$, tal que:

$y_i = x_i^t \beta + \varepsilon_i \quad i = 1,2 \ldots . n \quad$ con $\quad x_i^t = (x_{i1}, x_{i2} \ldots . x_{ip}),\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$ y ε_i es el término del error, una variable aleatoria perteneciente a una Distribución Normal con media 0 y varianza σ^2. Es necesario fijar $x_{i1} = 1$ para todo i para que el primer elemento de β corresponda al término independiente. En este caso trabajaremos sin el término independiente. La información del Datum u origen de las referencias de alturas, que permite eliminar el déficit de rango que ocurriría de solo tener información relativa de los desniveles puros, se resuelve por sustitución. El conjunto de todas las observaciones junto con los β lleva al modelo de Ecuaciones de Observación $y = X.\beta + \varepsilon$ Donde X es la matriz de *n x p* con elementos x_{ij}, el vector ε es un vector con elementos ε_i el vector y con elementos y_i

Un valor atípico en el caso de una regresión es un valor y_i que se aleja de la relación lineal seguida de la mayoría de los datos (valores atípicos verticales). Otro tipo de dato atípico, en las regresiones, es aquel que se aleja del conjunto de la mayoría de las variables explicativas del modelo (valores atípicos verticales). Cabe recordar que las variables explicativas consisten de -1, 0 y 1 por tratarse del modelado de una red de Grafos (Strang y Borre, 1997) lo que nos previene de encontrarnos ante este tipo de valores atípicos salvando el caso de una equivocación. El concepto de un M-estimador para un modelo de regresión lineal fue introducido por Huber (1973)

Un M-estimador de regresión de β se define como el $\hat{\beta}_M$, tal que minimiza: $\sum_{i=1}^{n} \rho(\frac{y_i - x_i^t \beta}{s})$ Siendo $\rho(u)$ una función continua y simétrica, llamada función objetivo con un único mínimo en 0 (Andersen, 2008; Rousseeuw y Leroy, 1987). S es un estimador de escala de los residuos que puede ser estimado antes o en simultáneo. S podría ser la mediana del valor absoluto de los residuos de algún estimador de residuos inicial. No usar estimador de escala *S* en (5), es lo mismo que haciendo abuso de notación sustituir *S* por 1. El hecho de usar el estimador de escala es importante ya que el M-estimador, no es necesariamente invariante con respecto a cambios de escala (es decir, si se multiplicaran los errores $y_i - x_i^t \beta$ por una constante, la nueva solución de la ecuación podría no ser igual que la anterior). En realidad, los estimadores MC de regresión son un caso poco robusto de M-estimadores con función objetivo $\rho(u) = u^2$. La vulnerabilidad de Mínimos Cuadrados (MC) proviene del mayor peso que se otorga a los valores extremos o atípicos por elevar al cuadrado los residuos a ser minimizados. En el caso de los M-estimadores de regresión propuestos por Huber (1973), la función objetivo se define de la siguiente manera: $\rho(u) = \begin{cases} \frac{1}{2}u^2, & |u| < a \\ \frac{1}{2}|u| - \frac{1}{2}a^2, & |u| \geq a \end{cases}$ La constante a es conocida como constante de ajuste, en el caso del presente trabajo $a = 1{,}345$ que corresponde a una propuesta de Huber (1973) con alta eficiencia. El estimador M de Huber es robusto frente a valores extremos en la dirección de *y* (valores atípicos verticales) pero no es robusto frente a valores extremos en la dirección X (valores atípicos horizontales) Cuando la varianza de los errores ε_i no es la misma para todo *i*, los estimadores M son más eficientes que los de mínimos cuadrados. El estimador MM es un tipo especial de estimador M y fue propuesto por Yohai (1987). Los estimadores MM son considerados como una generalización de los estimadores M. Están basados en una función ρ_1 que determina las propiedades robustas del estimador (Stuart, 2011). En este caso ρ_1 , es una función acotada, no decreciente y simétrica alrededor del cero.

Un MM-estimador de regresión de β se define como el $\hat{\beta}_{MM}$, tal que minimiza: $\sum_{i=1}^{n} \rho_1(\frac{y_i - x_i^t \beta}{\hat{\sigma}})$. Donde $\hat{\sigma}$ es un S- estimador de escala robusto introducido por Rousseauw y Yohai (1984) (Maronna y otros, 2019; Montgomery y otros, 2006). En este caso ρ_1 , es la función *bicuadrada* que se define como: $\rho_1(u) = \begin{cases} 3u^2 - 3u^4 + u^6, & |u| \leq 1 \\ 1, & |u| > 1 \end{cases}$. Que de acuerdo con la terminología de Maronna y otros (2019) es una *ρ-función acotada*, lo que permitiría lidiar tanto con valores atípicos verticales como horizontales.

Ajuste por MCP (Mínimos Cuadrados Ponderados)

El ajuste convencional de una red altimétrica es realizado por MCP. Con este ajuste buscaremos explicar las observaciones y_i con los valores de las cotas ajustadas $\hat{\beta}_{MCP}$ mediante la expresión $y_i = x_i^t \beta + \varepsilon_i \quad i = 1,2 \ldots . n$. El valor de la varianza de y_i denotada por $Var(y_i)$ dependerá de la distancia que fue necesaria recorrer para su determinación y puede expresarse como la propagación de una varianza kilométrica y la cantidad de kilómetros recorridos para obtener y_i denotada por Q_i. $Var(y_i) = \sigma_{Km^2}^2 \ Q_i$. El conjunto de todas las observaciones lleva al modelo de Ecuaciones de Observación $y = X.\beta + \varepsilon$. Que se resuelve por MCP mediante la resolución del sistema de Ecuaciones Normales $\hat{\beta}_{MCP} = inv(X^t.W.X).X^t.W.y$ Siendo W una matriz diagonal con los pesos w_i en la diagonal, la ponderación $w_i$se obtiene de la inversa de los cofactores $Q_i$de $Var(y_i)$.

MEDICIÓN Y AJUSTES Levantamiento de observaciones

El levantamiento de las variables de respuesta, los desniveles, se realizó mediante el método de nivelación geométrica desde el medio (Wolf y Ghilani, 2006) con niveles automáticos de 28 aumentos. Los desniveles medidos pueden verse en la Tabla 1 columna 2. La red consta de 8 ménsulas como las de la figura 1 y se les ha otorgado la siguiente nomenclatura: Agrimensura Vieja (AV), Agrimensura Nueva (AV), Partenón (P), Química 1 (Q1) , Química 2 (Q2), Hidráulica(H), Decanato (D), Construcciones (C) Estas ménsulas serán el soporte físico de las cotas de superficies equipotenciales cuyo valor será el resultado del ajuste de las observaciones de desnivel realizadas. Agrimensura Vieja (AV) sirvió para establecer el datum de referencia. En la Tabla 1, columna 1, se muestran los extremos de cada tramo medido. El vector β de 7x1 y que tiene las cotas a ajustar es $\beta = \begin{pmatrix} AN \\ Q_1 \\ D \\ Q_2 \\ H \\ P \\ C \end{pmatrix}$

Las varianzas de cada una las observaciones son directamente proporcionales al recorrido necesario para obtenerlas (Justo, 2018). Si bien estadísticamente son independientes recordemos que funcionalmente no lo son. Esto es porque deben cumplir condiciones de cierre por formar parte de una red de Grafos Orientados (Strang y Borre, 1997). A consecuencia de esto la suma de los residuos será distinta de cero. Por esta circunstancia también, el ingreso de valores atípicos notables puede ser inspeccionado antes de realizar el ajuste.

Tabla 1. Observaciones, residuos brutos y estudentizados externos.

DH medido	DH ajustado MCP	r_i(m)	$t_{(i)}$
2,046	2,046423	-0,0004231	-0,2946033
1,110	1,1100164	-0,0000164	-0,0103331
0,112	0,1105502	0,0014498	1,4412095
0,381	0,3764464	0,0045536	7,0212768
-1,615	-1,6148721	-0,0001279	-0,0699119
-0,920	-0,9212172	0,0012172	1,0040679
0,166	0,1661877	-0,0001877	-0,1433856
-0,435	-0,4342206	-0,0007794	-0,5861182
0,267	0,265363	0,0016367	0,9170405
0,842	0,8419835	0,0000165	0,0103408
-2,046	-2,046423	0,0004231	0,2946033
-1,110	-1,1100164	0,0000164	0,0103331
-0,111	-0,1105502	-0,0004498	-0,4158049
-0,375	-0,3764464	0,0014464	1,0294347
1,616	1,6148721	0,0011279	0,6262164
0,920	0,9212172	-0,0012172	-1,0040679
-0,167	-0,1661877	-0,0008123	-0,6300562
0,436	0,4342206	0,0017794	1,4269477
-0,265	-0,265363	0,0003633	0,1971074
-0,842	-0,8419835	-0,0000165	-0,0103408

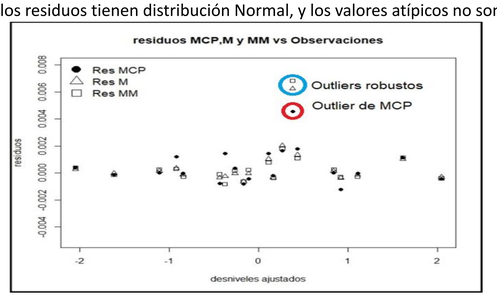
IDENTIFICACIÓN DE VALORES ATÍPICOS

Para identificar posibles valores atípicos en el ajuste por MCP se utilizaron los residuos: $r_i = y_i - \hat{y}_i$. Donde y_i son los desniveles medidos, e $\hat{y}_i$ son los valores ajustados de dichos desniveles por mínimos cuadrados pesados. También se calcularon los residuos estudentizados externos $t_{(i)}$ o residuos por el método Leave-One-Out (Maronna y otros, 2019) donde se mide la influencia de un dato atípico y_i ,en los residuos, excluyendo dicha observación del ajuste. Se define el residuo leave-one–out $r_{(i)}$, calculando: $\boldsymbol{r}_{(i)} = \boldsymbol{y_i} - \hat{\boldsymbol{y}}_{(i)}$. Donde $\hat{y}_{(i)}$ es el desnivel ajustado sin tener en cuenta el valor excluido. $\hat{y}_{(i)} = x_i^t \hat{\beta}_{(i)}$ (Ver Tabla 1)

COMPARACIÓN DE RESULTADOS

En la Figura 2 se puede observar un gráfico de residuos brutos versus los valores de los desniveles ajustados correspondientes a los tres métodos. También se resalta un residuo extremo para MCP en rojo y los residuos extremos de los dos métodos robustos resaltados en color celeste todos correspondientes a un mismo valor ajustado. En el gráfico, el residuo correspondiente a dicho valor es menor para MCP que en los métodos robustos (marcados en color celeste). Esto es debido a que el valor atípico distorsiona o corre el ajuste de regresión por MCP en su dirección. El procedimiento robusto tiende a dejar grandes los residuales asociados con valores atípicos, facilitando así la identificación. El procedimiento de estimación robusto produce, en este caso, casi los mismos valores ajustados de los parámetros obtenidos por el método MCP ya que los residuos tienen distribución Normal, y los valores atípicos no son significativos.

Figura 2. Gráfico de los Residuos vs. Desniveles Ajustados.



Ménsulas	MCP	M	MM
AN	17.960	17.960	17.960
Q1	19.070	19.070	19.071
D	19.181	19.181	19.182
Q2	19.557	19.556	19.556
H	17.794	17.794	17.794
P	16.179	16.179	16.179
C	18.636	18.636	18.636

Tabla 2. Resultados de los coeficientes por los tres métodos de ajuste

CONCLUSIONES

El estudio de outliers o valores atípicos debe incluir estrategias, tanto para su detección como para la medición de su influencia en los parámetros estimados.

Consideramos oportuna la comparación entre los estimadores de MCP, los M y MM dada la posibilidad computacional actual con software como el R. Consideramos que agregar esta información a los ajustes realizados por MCP en las Redes Altimétricas Topográficas otorgará más herramientas de análisis al profesional en cuanto a la pertinencia o no de las observaciones que representen valores atípicos en el ajuste convencional.El análisis robusto se puede utilizar como confirmatorio de mínimos cuadrados. Siempre que se haga un análisis por mínimos cuadrados, sería conveniente también hacer un ajuste robusto. Si concuerdan, o la diferencia es exigua, se deben usar los resultados de los estimadores de mínimos cuadrados, sin embargo, si difieren, se deben identificar las razones de tales diferencias.

La estimación M introducida por Huber (1973) ofrece un enfoque más simple que el ajuste MM. No es robusto a los puntos influyentes en la dirección de *X* pero es apropiado cuando se puede suponer que la contaminación de la muestra está en la dirección de la respuesta que es el caso de estudio.

REFERENCIAS

Andersen, R. (2008). *Modern Methods for Robust Regression*. Thousand Oaks: SAGE Publications.
Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 1, 799-821.
Justo C. (2018) Tratamiento Estadístico de una Red Altimétrica Topográfica. Tesis de Maestría en Ingeniería UNLP. http://sedici.unlp.edu.ar/handle/10915/65539
Maronna R.A., Douglas M.R., Yohai V.J. y Salibián-Barrera M.S. (2019). *Robust Statistics: Theory and Methods (with R)*. 2nd Edition, Wiley.
Montgomery, D.C., Pek, E.A. y Vining, G.G. (2006). *Introducción al Análisis de Regresión Lineal*. 3ª. Edición, Ed. Compañía Editorial Continental, México.
Rousseeuw, P.J. y Leroy A.M. (1987). *Robust Regression and Outlier Detection*. Hoboken, Wiley. DOI:10.1002/0471725382.
Rousseeuw P. y Yohai V. (1984) Robust Regression by Means of S-Estimators. En: Franke J., Härdle W., Martin D. (eds) *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics*, vol 26. Springer, New York, NY. https://doi.org/10.1007/978-1-4615-7821-5_15
Strang G. y Borre K. (1997). *Linear Algebra, Geodesy and GPS*. Wellesley Cambridge Press.
Stuart, C.A. (2011). Robust Regression. Recuperado de: https://www.semanticscholar.org/paper/Robust-Regression-Stuart/f50f6e74b773ba0b89df34744523bd7c6b148125
Wolf, P. y Ghilani, C. (2006). *Adjustment Computations: Spatial Data Analysis*, Fourth Edition. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470121498
Yohai, V. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. The Annals of Statistics, 15(2), 642-656. Recuperado de;: http://www.jstor.org/stable/2241331.