Tarea 4: Modelos Estadísticos I.

Rojas Gutiérrez Rodolfo Emmanuel

11 de marzo de 2021

Ejercicio 1:

Los datos de la Figura 1 se han ordenado de manera conveniente para poder ver la relación entre la varianza de la respuesta y la variable explicativa. La muestra consta de 33 observaciones (X_i, Y_i) con un número de conjuntos que tienen repeticiones exactas de la X o repeticiones aproximadas. Estos conjuntos están indicados por las agrupaciones.

(a) Ajuste un modelo de regresión lineal simple, por mínimos cuadrados no ponderados

(b) Haga una gráfica de residuos contra valores ajustados. Comente la gráfica.

(c) Haga una gráfica de residuos contra valores de la variable independiente. Comente la gráfica.

(d) ¿Qué relación tiene la varianza (de las Y_s) con los valores de la variable explicativa? ¿Lineal, cuadrática o cubica?

(e) Asigne los pesos $\hat{w}_i = \hat{\sigma}^{-2}$ que les corresponden a las observaciones (llenar la tercera columna de la Tabla en la Figura 1).

(f) Ajuste el modelo de regresión lineal simple por mínimos cuadrados ponderados.

(g) Compare y comente los ajustes de los incisos (a) y (f).

X	Y	$\hat{w}$	X	Y	$\hat{w}$	X	Y	$\hat{w}$	X	Y	$\hat{w}$	X	Y	$\hat{w}$
3.00	2.60		5.34	5.92		7.70	7.68		9.03	9.47		10.17	9.78	
3.00	2.67		5.38	5.35		7.80	9.81		9.07	11.45		10.18	12.39	
3.00	2.66		5.40	4.33		7.81	6.52		9.11	12.14		10.22	11.03	
3.00	2.78		5.40	4.89		7.85	9.71		9.14	11.50		10.22	8.00	
3.00	2.80		5.45	5.21		7.87	9.82		9.16	10.65		10.22	11.90	
						7.91	9.81		9.37	10.64		10.18	8.68	
						7.94	8.50					10.50	7.25	
												10.23	13.46	
												10.03	10.19	
												10.23	9.93	

Figura 1: Datos para el problema de mínimos cuadrados ponderados.

Solución. **(a)** Se ajustará un modelo de regresión lineal simple a los datos presentados en la Figura 1, en el que se consideraran términos de error $i.i.d$ con media 0 y varianza constante $\sigma^2 > 0$, es decir que

$$Y = \mathbf{X}\beta + \varepsilon,$$

donde:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{X} = \begin{array}{cc} 1 & X \\ \multicolumn{2}{c}{\begin{pmatrix} 1.0 & 3.0 \\ 1.0 & 3.0 \\ 1.0 & 3.0 \\ 1.0 & 3.0 \\ 1.0 & 3.0 \\ 1.0 & 5.34 \\ 1.0 & 5.38 \\ 1.0 & 5.4 \\ 1.0 & 5.4 \\ 1.0 & 5.45 \\ 1.0 & 7.7 \\ 1.0 & 7.8 \\ 1.0 & 7.81 \\ \vdots & \vdots \\ 1.0 & 9.16 \\ 1.0 & 9.37 \\ 1.0 & 10.17 \\ 1.0 & 10.18 \\ 1.0 & 10.22 \\ 1.0 & 10.22 \\ 1.0 & 10.22 \\ 1.0 & 10.18 \\ 1.0 & 10.5 \\ 1.0 & 10.23 \\ 1.0 & 10.03 \\ 1.0 & 10.23 \end{pmatrix}} \end{array}, \quad Y = \begin{pmatrix} 2.6 \\ 2.67 \\ 2.66 \\ 2.78 \\ 2.8 \\ 5.92 \\ 5.35 \\ 4.33 \\ 4.89 \\ 5.21 \\ 7.68 \\ 9.81 \\ 6.52 \\ \vdots \\ 10.65 \\ 10.64 \\ 9.78 \\ 12.39 \\ 11.03 \\ 8.0 \\ 11.9 \\ 8.68 \\ 7.25 \\ 13.46 \\ 10.19 \\ 9.93 \end{pmatrix}, \tag{1}$$

y ε es el vector de términos de error, a partir de este punto se hará referencia a la columna X de la matriz $\mathbf{X}$ como los valores de la variable independiente, o incluso se le denominará variable independiente. De este modo las estimaciones por mínimos cuadrados para $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ están dados por:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = \begin{pmatrix} -0.506 \\ 1.127 \end{pmatrix}. \tag{2}$$

Por lo que, la ecuación de regresión para la estimación de la media de la variable dependiente Y dado un valor de la variable independiente X esta dada por:

$$\hat{E}[Y_i|X_i] = -0.506 + 1.127X_i, \ i \in \{1, \ldots, 33\}. \tag{3}$$

Por otro lado, el coeficiente de determinación para este modelo esta dado por

$$R^2_{OLS} = \frac{\beta' \mathbf{X}' Y - n\overline{Y}^2}{Y'Y - n\overline{Y}^2} = 0.796. \tag{4}$$

Para concluir este inciso se muestra en la Figura 2 una gráfica con los datos de la Figura 1, la correspondiente recta de regresión lineal (3) y bandas de confianza y predicción al 95 % de confianza. Esto con la finalidad de comparar está gráfica con el futuro modelo de mínimos cuadrados ponderados que se ajustará posteriormente.

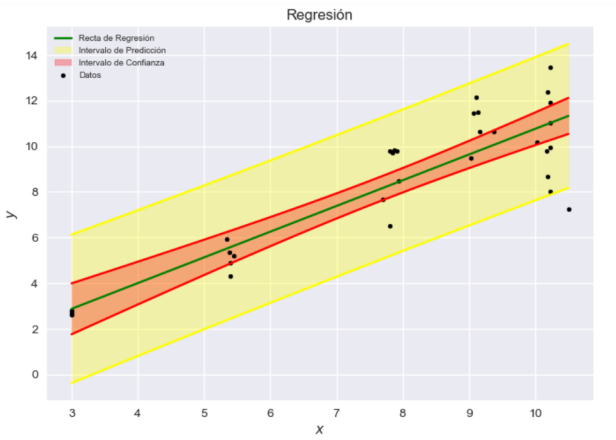

Figura 2: Modelo de Regresión Lineal Simple Ajustado por MCO, con Bandas de Predicción y Confianza al 95 %.

(b) La gráfica solicitada se presenta en la Figura 3, los datos se encuentran clasificados por colores uno por cada uno de los 5 grupos dados en un inicio, donde los grupos se han etiquetado como $G_1, G_2, \ldots, G_5$ y representan a cada una de las 5 columnas de la tabla 1 respectivamente. Ahora, dado que los residuales y los valores ajustados son elementos aleatorios no correlacionados, si es que se cumplen los supuestos del modelo de regresión lineal simple, entonces no se debería de notar una relación entre dichas cantidades observadas y se esperaría que los residuales se vieran dispersos como variables aleatorias de media 0 y varianza constante[1], sin embargo en la gráfica construida es fácil observar que conforme crece el valor ajustado $\hat{Y}$ también lo hace la dispersión que existe entre los residuales. Como observación importante tome en cuenta que la varianza parece cambiar a través de los grupos.

(c) La gráfica solicitada se presenta en la Figura 4, nuevamente los datos se encuentran clasificados por colores de manera análoga a como se hizo en el inciso **(b)**. Dado que el valor de la variable independiente no debería influir en el comportamiento de los residuales, bajo los supuestos del modelo de regresión lineal simple, entonces la dispersión de los residuales a través de los diferentes

[1]Pese a que no son de varianza constante, pero al ser las estimaciones de los errores deberían reflejar en cierto modo su comportamiento.

Figura 3: Gráfica de residuales vs Valores Ajustados.

valores de la variable independiente (X) debería ser la misma, sin embargo se observa un patrón similar al que se tenía en la gráfica del inciso **(b)**, es decir conforme aumenta el valor de la variable independiente también parece incrementar la variabilidad de los residuales respecto al 0, este último hecho era de esperarse dado que $\beta_1 X_i = \hat{Y}_i$ con $\beta_1 = 1.127$, es decir los valores ajustados y los valores de la variable independiente están relacionados linealmente mediante un coeficiente de pendiente positivo. De este modo, los incisos **(b)** y **(c)** parecen indicar la existencia de cierta heterocedasticidad en la variabilidad de los errores[2], relacionada con la variable independiente X, por lo que un modelo de regresión lineal simple ajustado por mínimos cuadrados ponderados podría ser mas sensato para estos datos, que uno de regresión lineal simple ajustado por MCO.

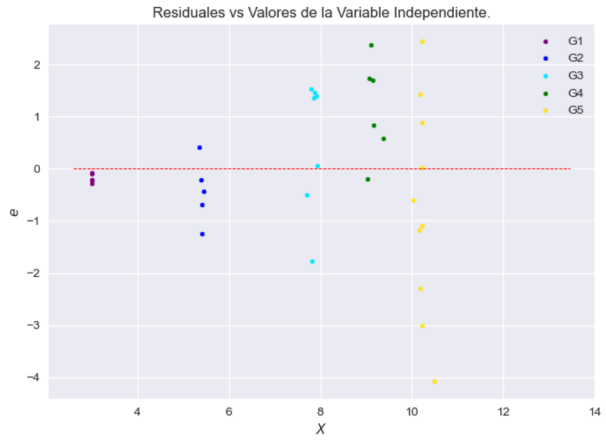

Figura 4: Gráfica de residuales vs Valores de la Variable Independiente.

[2]finalmente igual a la variabilidad de las observaciones.

(d) Dado que se esta pensando en ajustar un modelo por mínimos cuadrados ponderados, y dado que pareciera que el valor de la variable independiente afecta de manera directa a la variabilidad de las observaciones, este inciso tendrá como finalidad encontrar de que manera es que se da esta dependencia, ya que la misma ayudará a establecer los pesos adecuados para el ajuste del modelo. Para ello se realizó el siguiente análisis, se obtuvo la varianza muestral de las observaciones para cada uno de los 5 grupos,[3] dichos valores se representaran de manera vectorial como σ^2 y a las componentes de este vector como σ_i^2, con estos datos se realizó una gráfica de los valores de la variable independiente (X) contra la varianza muestral de las observaciones de la variable dependiente (Y) por grupo, dando como resultado el gráfico mostrado en la Figura 5.[4]

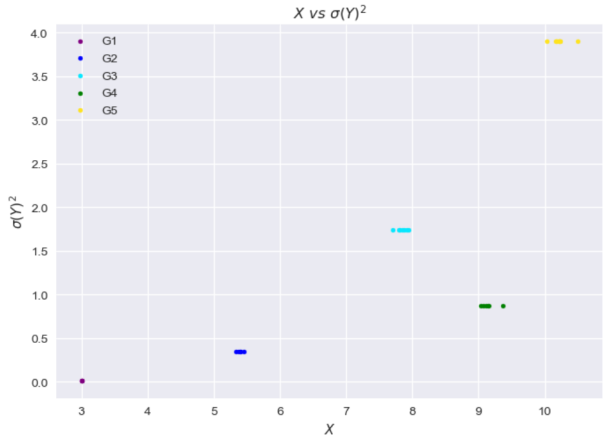

Figura 5: Gráfica Variable Independiente vs Varianzas por Grupo.

Utilizando los datos anteriores se ajustaron:

1. Un polinomio lineal: $a_0 + a_1 x$.
2. Un polinomio cuadrático: $b_0 + b_1 x + b_2 x^2$,
3. Un polinomio cúbico: $c_0 + c_1 x + c_2 x^2 + c_3 x^3$.

Los ajustes se realizaron por Mínimos Cuadrados Ordinarios (MCO), utilizando como variable explicativa los valores de la variable independiente[5] y como variable explicada a los valores de las varianzas muéstrales de la variable dependiente por grupo, obteniendo los siguientes resultados

$$\begin{aligned}
\sigma_i^2 &= -1.973 + 0.485 X_i,\ i \in \{1, \ldots, 33\}, \quad R^2 = 0.663, \\
\sigma_i^2 &= 1.808 - 0.847 X_i + 0.098 X_i^2,\ i \in \{1, \ldots, 33\}, \quad R^2 = 0.776, \\
\sigma_i^2 &= -6.235 + 3.588 X_i - 0.617 X_i^2 + 0.035 X_i^2,\ i \in \{1, \ldots, 33\}, \quad R^2 = 0.825, \qquad (5)
\end{aligned}$$

[3] Sí se ha pecado de obviedad, por ejemplo se tendrá que la primer varianza corresponde a la varianza muestral de los valores de Y, en la primera sección de la Tabla en la Figura 1 y por ende las primeras 5 entradas de σ^2 corresponderán a este valor, luego se pasa la segunda sección de la tabla y se repite el procedimiento.

[4] Los grupos nuevamente se destacaron por colores de manera análoga a como se hizo en los incisos **(b)** y **(c)**.

[5] Independiente respecto al modelo original.

Al ajustar estos tres polinomios se identificaron dos problemas. El primero es que el coeficiente R^2 siempre se vuelve más grande cuando se introduce una covariable extra, por ende si se ajustará un modelo de grado cuarto su coeficiente de determinación sería aún mayor que el de los modelos en (5), sin embargo ese incremento no premiaría la parsimonia del modelo. Por otra parte, es fácil observar en la gráfica presentada en la Figura 6, que los valores que se generarían con este ajuste tendrían posibilidad de ser negativos en el intervalo formado por el mínimo de las observaciones de X y el máximo de las mismas, lo cual no es deseable debido a que se esta modelando la varianza de la variable dependiente, que como se sabe es una cantidad no negativa.

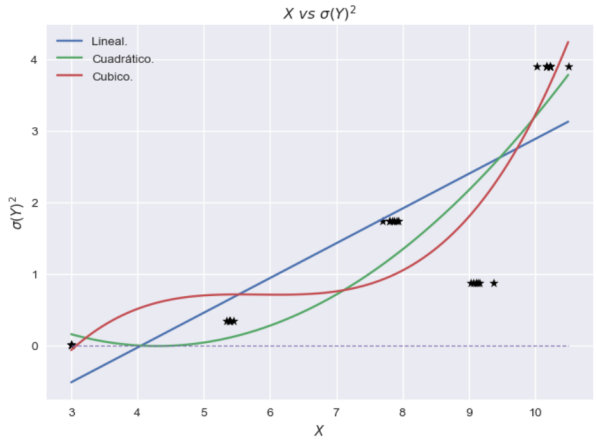

Figura 6: Gráfica Polinomios Ajustados Por MCO I.

A modo de confirmación de este hecho, es posible verificar en el script adjunto a esta tarea que efectivamente el mínimo de cualquiera de estos polinomios en el intervalo mencionado es menor a cero. Para evitar estos problemas, se decidió intentar ajustar alguna de las siguientes dos opciones[6]:

1. Un polinomio cuadrático: b_2x^2,
2. Un polinomio cúbico: c_3x^3,

El modelo lineal se descartó por completo, debido al pobre desempeño mostrado en la Figura 6. Los ajustes nuevamente se realizaron ocupando la técnica de MCO obteniendo los siguientes resultados:

$$\begin{aligned} \sigma_i^2 &= 0.0291X_i^2,\ i \in \{1,\ldots,33\}, \quad R^2 = 0.867, \\ \sigma_i^2 &= 0.0031X_i^3,\ i \in \{1,\ldots,33\}, \quad R^2 = 0.9002, \end{aligned} \tag{6}$$

pudiera sorprender al lector que el coeficiente de determinación en los modelos (6), sea mayor que el de los modelos en (5) dado el comentario de que el agregar covariables siempre incremente su valor, pero dado que los modelos propuestos se han tomado sin intercepto es necesario hacer una modificación al calculo del coeficiente de determinación,[7] la cual además premia la parsimonia de los

[6] Observe que dada la relación creciente que existe entre X y σ^2, se esperaría que las estimaciones de los coeficientes b_2 y c_3 fueran no negativas, lo que evitaría el problema de los pesos negativos.

[7] Dicha modificación se explicará en el anexó a este trabajo.

modelos anteriores. Tomando entonces en cuenta los coeficientes de determinación de los modelos en (6), se concluye para efectos de este inciso que la relación existente entre la varianza de las observaciones y la variable independiente X es cúbica de la forma

$$\sigma_i^2 = \sigma^2 \cdot X_i^3,\ \sigma^2 > 0,\ i \in \{1, \ldots, 33\} \tag{7}$$

.

(e) Sean $\mathbf{X}, Y$ y β como en (1) se intentará ajustar un modelo de regresión lineal simple por Mínimos Cuadrados Ponderados (MCP), es decir se intentará estimar el vector de parámetros β para

$$Y = \mathbf{X}\beta + \varepsilon, \tag{8}$$

pero a diferencia del inciso **(a)** se trabajará bajo el supuesto de que ε es un vector de términos de error, tales que $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2 V$ donde V se tomará por (7) de la siguiente forma:[8]

$$V = \begin{pmatrix} X_1^3 & 0 & \vdots & 0 & 0 \\ 0 & X_2^3 & \vdots & 0 & 0 \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ 0 & 0 & \vdots & X_{32}^3 & 0 \\ 0 & 0 & \vdots & 0 & X_{33}^3 \end{pmatrix}. \tag{9}$$

De este modo, los pesos ω_i quedan dados por los valores en la diagonal de la inversa de la matriz V, dado que esta última es una matriz diagonal bastará con tomar la matriz diagonal cuyas entradas corresponden a los recíprocos de las entradas en la diagonal de V, esto es posible debido a que los 33 valores de la columna X de la matriz $\mathbf{X}$ son positivos, de este modo

$$V^{-1} = \begin{pmatrix} X_1^{-3} & 0 & \vdots & 0 & 0 \\ 0 & X_2^{-3} & \vdots & 0 & 0 \\ \ldots & \ldots & \ddots & \ldots & \ldots \\ 0 & 0 & \vdots & X_{32}^{-3} & 0 \\ 0 & 0 & \vdots & 0 & X_{33}^{-3} \end{pmatrix}. \tag{10}$$

Así pues, la Tabla en la Figura 1 se completa utilizando que $\omega_i = X_i^{-3}$ es el peso asociado a la varianza de Y_i, con $i = 1, \ldots, 33$, de la manera que se presenta en los Cuadros 1 y 2.

(f) Utilizando lo ya comentado en el inciso **(e)** en la parte (8), se estimó el vector de parámetros β para el modelo de regresión lineal simple por MCP como

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}' V \mathbf{X})^{-1} \mathbf{X}' V^{-1} Y = \begin{pmatrix} -0.836 \\ 1.173 \end{pmatrix}. \tag{11}$$

[8] Esto quiere decir que se mantiene el supuesto de de independencia entre los términos de error, más no el supuesto de varianza constante entre los mismos. Por otro lado, observe que una estimación de σ_1^2 esta dada por (6) y es igual a 0.0031.

X	Y	ω	X	Y	ω	X	Y	ω
3.00	2.60	$3.704{\cdot}10^{-2}$	5.34	5.92	$6.567{\cdot}10^{-3}$	7.70	7.68	$2.190{\cdot}10^{-3}$
3.00	2.67	$3.704{\cdot}10^{-2}$	5.38	5.35	$6.422{\cdot}10^{-3}$	7.80	9.81	$2.107{\cdot}10^{-3}$
3.00	2.66	$3.704{\cdot}10^{-2}$	5.40	4.33	$6.351{\cdot}10^{-3}$	7.81	6.52	$2.099{\cdot}10^{-3}$
3.00	2.78	$3.704{\cdot}10^{-2}$	5.40	4.89	$6.351{\cdot}10^{-3}$	7.85	9.71	$2.067{\cdot}10^{-3}$
3.00	2.80	$3.704{\cdot}10^{-2}$	5.45	5.21	$6.177{\cdot}10^{-3}$	7.87	9.82	$2.052{\cdot}10^{-3}$
						7.91	9.81	$2.021{\cdot}10^{-3}$
						7.94	8.50	$1.998{\cdot}10^{-3}$

Cuadro 1: Tabla de Datos con Ponderadores (Redondeados a 3 Dígitos Después del Punto Decimal).

X	Y	ω	X	Y	ω
9.03	9.47	$1.358{\cdot}10^{-3}$	10.17	9.78	$9.507{\cdot}10^{-4}$
9.07	11.45	$1.340{\cdot}10^{-3}$	10.18	12.39	$9.479{\cdot}10^{-4}$
9.11	12.14	$1.323{\cdot}10^{-3}$	10.22	11.03	$9.368{\cdot}10^{-4}$
9.14	11.50	$1.310{\cdot}10^{-3}$	10.22	8.00	$9.368{\cdot}10^{-4}$
9.16	10.65	$1.301{\cdot}10^{-3}$	10.22	11.90	$9.368{\cdot}10^{-4}$
9.37	10.64	$1.216{\cdot}10^{-3}$	10.18	8.68	$9.479{\cdot}10^{-4}$
			10.50	7.25	$8.638{\cdot}10^{-4}$
			10.23	13.46	$9.341{\cdot}10^{-4}$
			10.03	10.19	$9.911{\cdot}10^{-4}$
			10.23	9.93	$9.341{\cdot}10^{-4}$

Cuadro 2: Tabla de Datos con Ponderadores (Redondeados a 3 Dígitos Después del Punto Decimal).

Por lo que en este caso, la ecuación para la estimación de la media de la variable dependiente Y dado un valor de la variable independiente X, esta dada por:

$$\hat{E}[Y_i|X_i] = -0.836 + 1.173X_i,\ i \in \{1, \ldots, 33\}\,. \tag{12}$$

También se obtuvo el coefiente de determinación de este modelo, de la siguiente manera

$$R^2_{WLS} = 1 - \frac{Y'(V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})Y - Y'V^{-1}Y}{(Y - Y_p)'V^{-1}(Y - Y_p)} = 0.938, \tag{13}$$

donde el la expresión anterior $SS(Res) = SS(Tot)_{NC} - SS(Reg) = Y'(V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})Y - Y'V^{-1}Y$ y $SS(Tot)_C = (Y - Y_p)'V^{-1}(Y - Y_p)$, por último $Y_p = \sum_{i=1}^{3} \frac{\omega_i Y_i}{\sum_{j=1}^{33} \omega_j}$ es decir Y_p es la media ponderada de las Y por los pesos dados en los Cuadros 1 y 2.

(g) Primeramente, observe la gráfica presentada en la Figura 7 esto dará un primer criterio para comparar el ajuste de los modelos en **(a)** y **(f)**, es fácil observar como las bandas de confianza obtenidas bajo el modelo de regresión lineal simple ajustado con MCP, reflejan de mejor manera el comportamiento creciente de la varianza de la variable dependiente respecto a la variable independiente, comentado en los incisos **(c)** y **(d)**, dando una estimación mas precisa y menos conservadora

para valores pequeños de la variable independiente, lo que contrasta con los intervalos obtenidos bajo el modelo de regresión lineal simple ajustado con MCO, ver Figura 2, de igual manera se observa que para valores mas grandes de la variable independiente, se tiene un intervalo un poco mas abierto que el obtenido por MCO, lo que podría sugerir una subestimación de la variabilidad para estos valores bajo el modelo de regresión lineal ajustado por MCO.

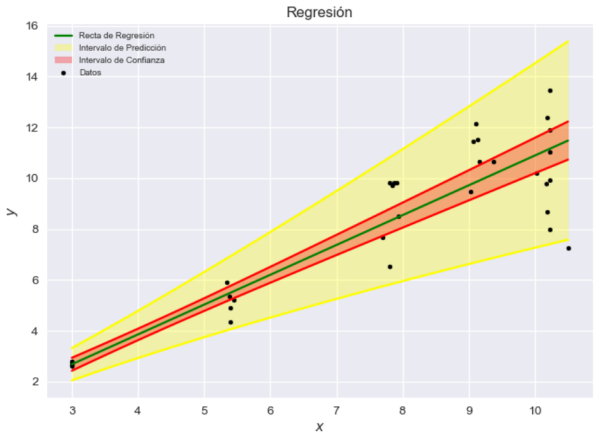

Figura 7: Modelo de Regresión Lineal Simple Ajustado por MCP, con Bandas de Predicción y Confianza al 95 %.

Por otra parte observe la gráfica presentada en la Figura 8, recuerde que para obtener las estimaciones para β (11), el modelo (8) debía ser transformado a un modelo cuyos términos de error fueran de media cero, varianza constante y correlación 0, por lo que se graficaron los residuales del modelo transformado contra los valores ajustados de Y para analizar si existía alguna mejoría en la dispersión de los mismos[9], a través de los distintos valores ajustados, usando que los residuales en el modelo transformado quedan dados por:

$$f = K^{-1}e,$$

con e los residuales del modelo original y

$$K^{-1} = diag(\sqrt{\omega_1}, \ldots, \sqrt{\omega_{33}}),$$

No es difícil observar que la dispersión de los residuales mejoró bastante, respecto a la presentada en la gráfica del inciso **(c)**. No se presenta una gráfica contra los valores de la variable independiente,[10] debido a que nuevamente existe bajo el modelo de regresión lineal por MCP un coeficiente de pendiente positivo relativamente cercano a 1, y por ende, al igual que en los incisos **(b)** y **(c)** se obtendrían resultados similares. Por último, si se comparan los coeficientes de determinación del modelo de regresión lineal simple ajustado con MCP (13) contra el mismo coeficiente para el modelo de regresión lineal simple ajustado con MCO (4), uno se percata que bajo el modelo

[9] Esperando que estos se comportaran como las estimaciones de términos de error de media cero y varianza constante.

[10] Aunque la misma puede consultarse en el script adjunto, para constatar lo que aquí se ha dicho.

ajustado con MCP la cantidad de variabilidad de la variable dependiente Y, explicada por la variable independiente X es mayor. Por lo que, bajo estos tres criterios se concluye que el modelo de regresión lineal simple ajustado por MCP tiene un mejor ajuste a los datos proporcionados.

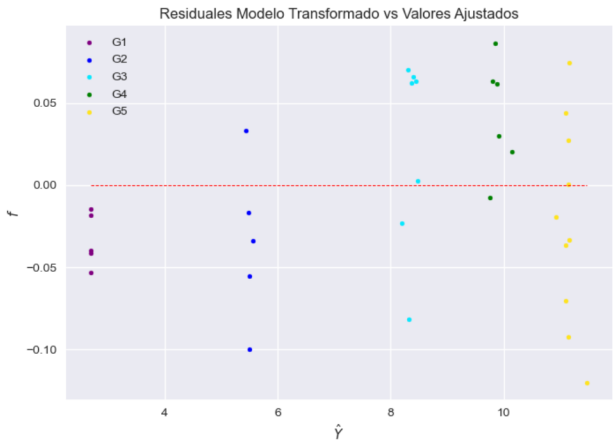

Figura 8: Residuales del Modelo Transformado vs Valores Ajustados para Y.

■

Ejercicio 2:

Considerando los datos que se muestran en la Figura 9, donde se presentan veinte valores de las variables explicativas X_0, X_1, X_2, Z_1 y Z_2, asi como de la variable respuesta Y, se pide reproducir las expresiones y hacer los cálculos que se indican en los cinco incisos del Teorema visto en la clase del miércoles.

(a) Con base en los datos de la Tabla 2, construya la matriz de diseño X formada por las columnas X_0, X_1 y X_2, así como también la matriz Z de regresores adicionales formada por las columnas Z_1 y Z_2. Además construya las matrices W, R_G, L yM , que se describen al inicio del Teorema.

(b) Haga las cuentas matriciales y presente los resultados que se indican en los incisos (i), (ii), (iii), (iv) y (v).

Solución. ■

(a) La matriz de diseño X, la matriz de regresores adicionales Z y la matriz con los valores de

	X0	X1	X2	Z1	Z2	Y		X0	X1	X2	Z1	Z2	Y
[1,]	1	8.640	14.029	12.160	15.552	37.395	[11,]	1	8.708	12.228	8.422	23.260	36.052
[2,]	1	8.245	2.011	13.206	4.625	14.322	[12,]	1	4.993	9.287	6.483	6.828	14.576
[3,]	1	4.340	6.099	13.908	6.529	15.726	[13,]	1	2.396	7.561	13.404	22.648	21.319
[4,]	1	2.462	13.535	1.446	19.185	17.473	[14,]	1	2.040	5.571	12.089	16.171	10.902
[5,]	1	1.502	2.363	4.201	4.535	6.638	[15,]	1	5.105	6.286	16.445	7.888	20.041
[6,]	1	9.337	5.400	5.626	15.311	20.617	[16,]	1	1.344	2.087	15.941	22.767	13.359
[7,]	1	7.936	3.660	7.986	17.047	17.589	[17,]	1	2.515	2.351	18.701	2.846	9.596
[8,]	1	1.151	2.696	17.821	8.105	7.669	[18,]	1	5.671	8.490	11.267	19.710	24.117
[9,]	1	2.241	12.216	13.909	6.214	19.537	[19,]	1	3.251	14.989	5.454	13.263	21.446
[10,]	1	2.799	10.753	11.006	9.401	16.973	[20,]	1	8.627	1.126	13.371	10.317	19.034

Figura 9: Valores de las variables explicativas X_0, X_1 y X_2, así como las covariables adicionales Z_1,Z_2 y de la covariable Y.

la variable dependiente Y se presentan a continuación:

$$X = \begin{pmatrix} 1 & 8.640 & 14.029 \\ 1 & 8.245 & 2.011 \\ 1 & 4.340 & 6.099 \\ 1 & 2.462 & 13.535 \\ 1 & 1.502 & 2.363 \\ 1 & 9.337 & 5.400 \\ 1 & 7.936 & 3.660 \\ 1 & 1.151 & 2.696 \\ 1 & 2.241 & 12.216 \\ 1 & 2.799 & 10.753 \\ 1 & 8.708 & 12.228 \\ 1 & 4.993 & 9.287 \\ 1 & 2.396 & 7.561 \\ 1 & 2.040 & 5.571 \\ 1 & 5.105 & 6.286 \\ 1 & 1.344 & 2.087 \\ 1 & 2.515 & 2.351 \\ 1 & 5.671 & 8.490 \\ 1 & 3.251 & 14.989 \\ 1 & 8.627 & 1.126 \end{pmatrix}, \quad Z = \begin{pmatrix} 12.160 & 15.552 \\ 13.206 & 4.625 \\ 13.908 & 6.529 \\ 1.446 & 19.185 \\ 4.201 & 4.535 \\ 5.626 & 15.311 \\ 7.986 & 17.047 \\ 17.821 & 8.105 \\ 13.909 & 6.214 \\ 11.006 & 9.401 \\ 8.422 & 23.260 \\ 6.483 & 6.828 \\ 13.404 & 22.648 \\ 12.089 & 16.171 \\ 16.445 & 7.888 \\ 15.941 & 22.767 \\ 18.701 & 2.846 \\ 11.267 & 19.710 \\ 5.454 & 13.263 \\ 13.371 & 10.317 \end{pmatrix} \quad Y = \begin{pmatrix} 37.395 \\ 14.322 \\ 15.726 \\ 17.473 \\ 6.638 \\ 20.617 \\ 17.589 \\ 7.669 \\ 19.537 \\ 16.973 \\ 36.052 \\ 14.576 \\ 21.319 \\ 10.902 \\ 20.041 \\ 13.359 \\ 9.596 \\ 24.117 \\ 21.446 \\ 19.034 \end{pmatrix}, \tag{14}$$

Ahora, se supondrá que se busca ajustar un modelo de la forma

$$E[Y|X, Z] = \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} \beta_G \\ \gamma_G \end{pmatrix} = W\delta_G, \tag{15}$$

y que uno de la forma

$$E[Y|X] = X\beta, \tag{16}$$

fue ajustado con anterioridad obteniendo como estimación para β al siguiente vector

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y \begin{pmatrix} 2.722 \\ 1.643 \\ 1.097 \end{pmatrix}. \tag{17}$$

Utilizando las matrices en (14), entonces es posible obtener las matrices W, L, M, R_G, R comentadas[11] en el **Teorema 1.1** y quedan dadas de la siguiente manera:

$$W = \big(X, Z\big) = \begin{pmatrix} 1 & 8.640 & 14.029 & 12.160 & 15.552 \\ 1 & 8.245 & 2.011 & 13.206 & 4.625 \\ 1 & 4.340 & 6.099 & 13.908 & 6.529 \\ 1 & 2.462 & 13.535 & 1.446 & 19.185 \\ 1 & 1.502 & 2.363 & 4.201 & 4.535 \\ 1 & 9.337 & 5.400 & 5.626 & 15.311 \\ 1 & 7.936 & 3.660 & 7.986 & 17.047 \\ 1 & 1.151 & 2.696 & 17.821 & 8.105 \\ 1 & 2.241 & 12.216 & 13.909 & 6.214 \\ 1 & 2.799 & 10.753 & 11.006 & 9.401 \\ 1 & 8.708 & 12.228 & 8.422 & 23.260 \\ 1 & 4.993 & 9.287 & 6.483 & 6.828 \\ 1 & 2.396 & 7.561 & 13.404 & 22.648 \\ 1 & 2.040 & 5.571 & 12.089 & 16.171 \\ 1 & 5.105 & 6.286 & 16.445 & 7.888 \\ 1 & 1.344 & 2.087 & 15.941 & 22.767 \\ 1 & 2.515 & 2.351 & 18.701 & 2.846 \\ 1 & 5.671 & 8.490 & 11.267 & 19.710 \\ 1 & 3.251 & 14.989 & 5.454 & 13.263 \\ 1 & 8.627 & 1.126 & 13.371 & 10.317 \end{pmatrix}, \tag{18}$$

$$L = (X^{'}X)^{-1}X^{'}Z = \begin{pmatrix} 15.587 & 8.074 \\ -0.267 & 0.309 \\ -0.448 & 0.433 \end{pmatrix}, \quad M = (Z^{'}RZ)^{-1} = \begin{pmatrix} 3.009 \cdot 10^{-3} & 2.253 \cdot 10^{-4} \\ 2.253 \cdot 10^{-4} & 1.328 \cdot 10^{-3} \end{pmatrix}. \tag{19}$$

Dado que las matrices R_G y R son de tamaño 20×20 pueden ser consultadas en el script adjunto a esta tarea por cuestiones de espacio.

(b) Para corroborar los incisos (i) y (ii) del **Teorema 1.1**, primero se realizo un calculo directo del estimador de MCO para el vector de parámetros δ del modelo (15) utilizando la matriz W (18) lo que dio como resultado

$$\hat{\delta}_G = (W^{'}W)^{-1}W^{'}Y = \begin{pmatrix} -8.495 \\ 1.672 \\ 1.177 \\ 0.531 \\ 0.365 \end{pmatrix}, \tag{20}$$

luego se calculo por separado $\hat{\beta}_G$ y $\hat{\gamma}_G$ de acuerdo a los incisos (i) y (ii)[12] del **Teorema 1.1**, es decir que

$$\hat{\gamma}_G = (Z^{'}RZ)^{-1}Z^{'}RY = MZ^{'}RY = \begin{pmatrix} 0.531 \\ 0.365 \end{pmatrix}. \tag{21}$$

[11] Los cálculos de las mismas fueron hechos con Julia, los resultados aquí se presentan redondeados a tres dígitos.

[12] En el script adjunto se calcularon los rangos de X y Z con el fin de verificar que ambas fueran de rango completo, y el rango de W para confirmar que las columnas de X y Z fueran $l.i.$

Posteriormente se calculo de las dos formas indicadas en el inciso (i) el valor de $\hat{\beta}_G$, es decir

$$\hat{\beta}_G = (X^{'}X)^{-1}X^{'}(Y - Z\hat{\gamma}_G) = \begin{pmatrix} -8.495 \\ 1.672 \\ 1.177 \end{pmatrix}, \tag{22}$$

y usando $\hat{\beta}$ en (17), es decir

$$\hat{\beta}_G = \hat{\beta} - L\hat{\gamma}_G = \begin{pmatrix} -8.495 \\ 1.672 \\ 1.177 \end{pmatrix}, \tag{23}$$

de (22) y (23) se observa que ambos cálculos para $\hat{\beta}_G$, son iguales bajo redondeo a tres dígitos, como era de esperarse. Luego de (21), (22) y (23) se tiene que

$$\begin{pmatrix} \hat{\beta}_G \\ \hat{\gamma}_G \end{pmatrix} = \begin{pmatrix} -8.495 \\ 1.672 \\ 1.177 \\ 0.531 \\ 0.365 \end{pmatrix}. \tag{24}$$

Comparando (20) y (24), se confirman los incisos (i) y (ii) bajo redondeo a 3 dígitos. Para corroborar los incisos (iii) y (iv), bastará observar que

$$Y^{'}R_GY = 84.892, \tag{25}$$

donde como ya se comento con anterioridad la matriz R_G puede consultarse en el script adjunto. Luego

$$(Y - Z\hat{\gamma}_G)^{'}R(Y - Z\hat{\gamma}_G) = 84.892, \tag{26}$$

donde Y, Z están dados en (14), la matriz R puede consultarse en el script adjunto y γ_G como en (21). Por último

$$Y^{'}RY - \hat{\gamma}_G^{'}Z^{'}RY = 84.892. \tag{27}$$

Los resultados (25), (26) y (27) corroboran los inciso (iii) y (iv) bajo redondeo a tres dígitos. Por último, para el inciso (v) se llevo a cabo un cálculo directo de la matriz de covarianzas de $\hat{\delta}_G$ de la siguiente forma

$$cov(\hat{\delta}_G) = \sigma^2(W^{'}W)^{-1} = \sigma^2 \begin{pmatrix} 1.1835 & -0.0370 & -0.0334 & -0.0487 & -0.0142 \\ -0.0370 & 0.0065 & 0.0004 & 0.0007 & -0.0004 \\ -0.0334 & 0.0004 & 0.0033 & 0.0013 & -0.0005 \\ -0.0487 & 0.0007 & 0.0013 & 0.0030 & 0.0002 \\ -0.0142 & -0.0004 & -0.0005 & 0.0002 & 0.0013 \end{pmatrix}, \tag{28}$$

Luego utilizando el **Teorema 1.1** se obtuvo lo siguiente: [13]

$$(X^{'}X)^{-1} + LML^{-1} = \begin{pmatrix} 1.1835 & -0.0370 & -0.0334 \\ -0.0370 & 0.0065 & 0.0004 \\ -0.0334 & 0.0004 & 0.0033 \end{pmatrix} \tag{29}$$

[13]Los redondeos aquí fueron a cuatro decimales porque con menos dígitos algunas cifras se anulaban.

$$-LM = \begin{pmatrix} -0.0487 & -0.0142 \\ 0.0007 & -0.0004 \\ 0.0013 & -0.0005 \end{pmatrix}, \quad -ML^{'} = (-LM)^{'}. \tag{30}$$

Y la matriz M, redondeada a 4 dígitos esta dada por:

$$M = \begin{pmatrix} 0.0030 & 0.0002 \\ 0.0002 & 0.0013 \end{pmatrix}, \tag{31}$$

de este modo se sigue de (29), (30) y de (31) que

$$\begin{aligned} cov(\hat{\delta}_G) &= \sigma^2 \begin{pmatrix} (X^{'}X)^{-1} + LML^{-1} & -LM \\ -ML^{'} & M \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} 1.1835 & -0.0370 & -0.0334 & -0.0487 & -0.0142 \\ -0.0370 & 0.0065 & 0.0004 & 0.0007 & -0.0004 \\ -0.0334 & 0.0004 & 0.0033 & 0.0013 & -0.0005 \\ -0.0487 & 0.0007 & 0.0013 & 0.0030 & 0.0002 \\ -0.0142 & -0.0004 & -0.0005 & 0.0002 & 0.0013 \end{pmatrix}, \end{aligned} \tag{32}$$

Entonces por (28) y (32) se sigue la igualdad establecida, redondeando a 4 dígitos.

1. Anexo

A.Ejercicio 1. El R^2 en un modelo de regresión lineal múltiple con matriz de diseño X y respuesta Y que no considera intercepto esta dado por:

$$R^2 = \frac{\hat{Y}^{'}\hat{Y}}{Y^{'}Y}.$$

A.Ejercicio 2.

Teorema 1.1. *Sea X matriz de nxp tal que $p < n$ de rango p, y sea Y un vector de $n \times 1$, suponga que un modelo*

$$E[Y|X] = X\beta, \tag{33}$$

fue ajustado por mínimos cuadrados ordinarios con anterioridad, dando como estimación para el vector β a $\hat{\beta}$. Y suponga que Z es una matriz de covariables adicionales de $n \times t$ con rango $t < n$ y tal que $t + p < n$, las cuales se quieren añadir al modelo, entonces si las columnas de Z son l.i de las columnas de X se cumplirá para el modelo

$$E[Y|X, Z] = \begin{pmatrix} X, Z \end{pmatrix} \begin{pmatrix} \beta_G, \\ \gamma_G \end{pmatrix} = W\delta_G, \tag{34}$$

que si se definen $L = (X^{'}X)^{-1}X^{'}Z$, $R = I_{n\times n} - X(X^{'}X)X^{'}$, $R_G = I_{n\times n} - W(W^{'}W)W^{'}$, $M = (Z^{'}RZ)^{-1}$, y $\hat{\delta}_G = \begin{pmatrix} \hat{\beta}_G \\ \hat{\gamma}_G \end{pmatrix}$ denota al estimador de MCO para el modelo (34) entonces:

(i) $\hat{\beta}_G = (X^{'}X)^{-1}X^{'}(Y - Z\hat{\gamma}_G) = \hat{\beta} - L\hat{\gamma}_G$.

(ii) $\hat{\gamma}_G = (Z^{'}RZ)^{-1}Z^{'}RY = MZ^{'}RY.$

(iii) $Y^{'}R_GY = (Y - Z\hat{\gamma}_G)^{'}R(Y - Z\hat{\gamma}_G).$

(iv) $Y^{'}R_GY = Y^{'}RY - \hat{\gamma}_G^{'}Z^{'}RY.$

(v) $V(\hat{\delta}_G) = \sigma^2 \begin{pmatrix} (X^{'}X)^{-1} + LML^{-1} & -LM \\ -ML^{'} & M \end{pmatrix}.$