Capítulo 4

Regresión ponderada y falta de ajuste

4.1. Introducción

En este capítulo se presentan la regresión ponderada y la prueba de falta de ajuste como un conjunto adicional de herramientas usadas para la estimación de los modelos de regresión múltiple.

4.2. Mínimos cuadrados ponderados

En muchos casos el supuesto de función variancia constante, $\mathrm{Var}\,(Y|\mathbf{X}) = \sigma^2$, podría no ser razonable. Suponga que se tiene una función media para el i-ésimo caso:

$$\mathrm{E}\,(Y|\mathbf{X} = \mathbf{x}_i) = \mathbf{x}_i^{'}\boldsymbol{\beta}$$

pero que en lugar de asumir que los errores son constantes se tiene:

$$\mathrm{Var}\,(Y|\mathbf{X} = \mathbf{x}_i) = \mathrm{Var}(e_i) = \frac{\sigma^2}{w_i}$$

donde $w_1, \cdots, w_n$ son números *positivos* conocidos. La función variancia se encuentra definida en términos de σ^2, sin embargo las variancias pueden ser diferentes para cada caso. Esto lleva al uso de los *mínimos cuadrados ponderados* para obtener los estimados. En términos matriciales, sea $\mathbf{W}$ una matriz diagonal $n \times n$ cuyos elementos son los w_i.

El modelo a usar es:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{Var}\,(\mathbf{e}) = \sigma^2 W \tag{4.2.1}$$

El estimador $\hat{\boldsymbol{\beta}}$ se escoge de tal forma que minimice la *suma de cuadrados residual ponderada*:

$$\begin{aligned} SCRes\,(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{'} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} w_i \left(y_i - \mathbf{x}_i^{'}\boldsymbol{\beta}\right)^2 \end{aligned} \tag{4.2.2}$$

La suma de cuadrados residual ponderada considera que alguno de los errores son más variables que otros ya que las observaciones con valores grandes de w_i tendrán menor variancia y mayor peso en la *SCRes* ponderada. El estimador por mínimos cuadrados ponderados es:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y} \tag{4.2.3}$$

Para poder obtener la ecuación anterior de manera directa, es conveniente transformar el problema 4.2.1 en uno que pueda ser resuelto por mínimos cuadrados ordinarios.

Sea $\mathbf{W}^{1/2}$ una matriz diagonal $n \times n$ cuyos elementos son los $\sqrt{w_i}$, entonces $\mathbf{W}^{-1/2}$ es también una matriz diagonal con elementos $1/\sqrt{w_i}$ tal que $\mathbf{W}^{1/2}\mathbf{W}^{-1/2} = \mathbf{I}$. La matriz de variancia covariancia de $\mathbf{W}^{1/2}\mathbf{e}$ es:

$$\begin{aligned} \text{Var}\left(\mathbf{W}^{1/2}\mathbf{e}\right) &= \mathbf{W}^{1/2}\text{Var}\,(\mathbf{e})\,\mathbf{W}^{1/2} \\ &= \mathbf{W}^{1/2}\left(\sigma^2\mathbf{W}^{-1}\right)\mathbf{W}^{1/2} \\ &= \mathbf{W}^{1/2}\left(\sigma^2\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\right)\mathbf{W}^{1/2} \\ &= \sigma^2\left(\mathbf{W}^{1/2}\mathbf{W}^{-1/2}\right)\left(\mathbf{W}^{-1/2}\mathbf{W}^{1/2}\right) \\ &= \sigma^2\mathbf{I} \end{aligned} \tag{4.2.4}$$

Es decir que $\mathbf{W}^{1/2}\mathbf{e}$ es un vector aleatorio con matriz de variancia covariancia igual a σ^2 veces la matriz identidad. Premultiplicando ambos lados de la ecuación 4.2.1 por $\mathbf{W}^{1/2}$ se obtiene:

$$\mathbf{W}^{1/2}\mathbf{Y} = \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{1/2}\mathbf{e} \tag{4.2.5}$$

Sean $\mathbf{Z} = \mathbf{W}^{1/2}\mathbf{Y}$, $\mathbf{M} = \mathbf{W}^{1/2}\mathbf{X}$ y $\mathbf{d} = \mathbf{W}^{1/2}\mathbf{e}$, entonces la ecuación anterior se convierte en:

$$\mathbf{Z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{d} \tag{4.2.6}$$

Como $\text{Var}(\mathbf{d}) = \sigma^2\mathbf{I}$ entonces la ecuación 4.2.6 puede resolverse usando los minimos cuadrados ordinarios:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{Z} \\ &= \left(\left(\mathbf{W^{1/2}X}\right)'\mathbf{W^{1/2}X}\right)^{-1}\left(\mathbf{W^{1/2}X}\right)'\mathbf{W}^{1/2}\mathbf{Y} \\ &= \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}\end{aligned}$$

que es el estimador dado en 4.2.3.

Resumiendo, la regresión por mínimos cuadrados ponderados de $\mathbf{Y}$ sobre $\mathbf{X}$ con pesos dados en la matriz diagonal $\mathbf{W}$ es la misma que la regresión por mínimos cuadrados ordinarios de $\mathbf{Z}$ sobre $\mathbf{M}$, donde:

$$\mathbf{M} = \begin{pmatrix} \sqrt{w_1} & \sqrt{w_1}x_{11} & \cdots & \sqrt{w_1}x_{1p} \\ \sqrt{w_2} & \sqrt{w_2}x_{21} & \cdots & \sqrt{w_2}x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_n} & \sqrt{w_n}x_{n1} & \cdots & \sqrt{w_n}x_{np} \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} \sqrt{w_1}y_1 \\ \sqrt{w_2}y_2 \\ \vdots \\ \sqrt{w_n}y_n \end{pmatrix}$$

Inclusive la columna de unos es multiplicada por los $\sqrt{w_i}$. El problema de regresión se resuelve usando $\mathbf{M}$ y $\mathbf{Z}$ en lugar de $\mathbf{X}$ y $\mathbf{Y}$.

4.2.1. Aplicaciones de los mínimos cuadrados ponderados

Los pesos pueden determinarse de varias formas. Si la i-ésima respuesta es un promedio de n_i observaciones entonces $\text{Var}(y_i) = \sigma^2/n_i$, es decir $w_i = n_i$. Si y_i es un total de n_i observaciones entonces $\text{Var}(y_i) = n_i\sigma^2$, es decir $w_i = 1/n_i$. Si la variancia es proporcional a algún predictor x_i entonces $\text{Var}(y_i) = x_i\sigma^2$, es decir $w_i = 1/x_i$.

Capullos de manzana

Suponga que cierta especie de árboles produce dos tipos morfológicos de capullos. Los capullos grandes pueden crecer de 15 a 20 cm en una estación de crecimiento mientras que los capullos pequeños rara vez exceden 1 cm.

Bland (1978) realizó un estudio descriptivo de las diferencias entre los capullos grandes y pequeños en árboles de manzana McIntosh. Usando árboles saludables, Bland tomo muestras de capullos grandes y pequeños de estos árboles durante la temporada de crecimiento de 1971, que duro aproximadamente 106 días. Los capullos muestreados fueron removidos del árbol, marcados y llevados al laboratorio para el análisis. El tamaño de los capullos podría presentar diferencias dependiendo del número de tallos, el tamaño promedio de estos tallos o ambos.

La data `Bland` contiene información para los capullos grandes y pequeños. Solo se considera la información correspondiente a los capullos grandes. El objetivo es encontrar una ecuación que permita describir adecuadamente la relación entre los $x =$ días de inactividad y $y =$ el número promedio de tallos. Al no tener una relación teórica para esta ecuación se examina el gráfico de dispersión de la Figura 4.1 que sugiere utilizar la función media lineal:

$$\mathrm{E}\left(Y|X=x\right)=\beta_0+\beta_1 x \tag{4.2.7}$$

Para cada día muestreado, se tiene $n =$ número de capullos muestreados, $y =$ número promedio de capullos en ese día y SD = desviación estándar dentro de cada día. Como $\mathrm{Var}\left(y|x\right)=\sigma^2/n$ se puede calcular la regresión por mínimos cuadrados ponderados para el número promedio de tallos en función a los días de inactividad cuyos pesos estan dados por los valores de n. Los resultados obtenidos se muestran a continuación.

Tabla 4.1: Regresión para la data `Bland`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.97	0.31	31.74	0.00
Dia	0.22	0.01	40.71	0.00

Tabla 4.2: ANVA para la data `Bland`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dia	1	6164.28	6164.28	1657.24	0.0000
Residuals	20	74.39	3.72		

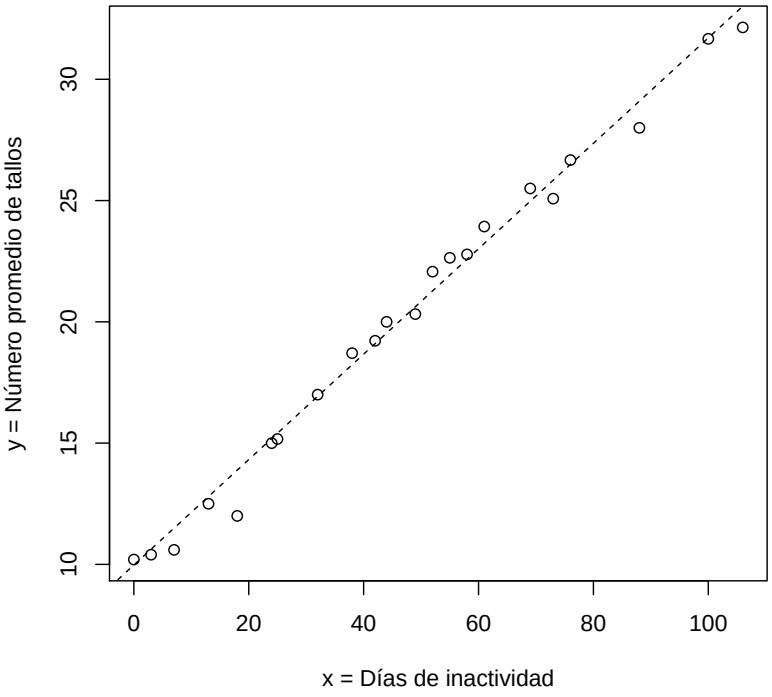

Figura 4.1: Diagrama de dispersión para los capullos grandes

4.3. Prueba de falta de ajuste

En esta sección se presenta una prueba de *falta de ajuste* para determinar si la función media elegida es adecuada para la data. La prueba de falta de ajuste requiere tener observaciones repetidas para algunos de los valores de los predictores, lo cual podria ocurrir de manera poco frecuente sobre todo si se trabaja con data observacional.

Si σ^2 es desconocida la prueba requiere un modelo libre que estime la variancia. El más común de estos modelos libres usa la dispersión que existe entre casos que tienen los mismos valores para todos los predictores. Por ejemplo, considere la data artificial con $n = 10$ en la Tabla 4.3. Si se tienen los valores correspondientes a y_i se puede calcular la desviación estándar

SD. Si se asume que la variancia es la misma para todos los valores de x, se debe calcular una variancia común ponderando las desviaciones estándar en un solo estimado. Si n es el número de casos para un valor x y SD es la desviación estándar correspondiente, entonces la *suma de cuadrados del error puro*, denotado por *SCEp*, es:

$$SCEp = \sum (n-1)\,\mathrm{SD}^2 \tag{4.3.1}$$

donde la suma se realiza sobre todos los grupos. Por ejemplo, *SCEp* es la suma de los valores en la cuarta columna de la Tabla 4.3:

$$SCEp = 0{,}0243 + 0 + 0{,}1301 + 2{,}2041 = 2{,}3585$$

Tabla 4.3: Cálculo del error puro

x	y	SD	$(n-1)\,\mathrm{SD}^2$	gl
1	2.55			
1	2.75	0.1102	0.0243	2
1	2.57			
2	2.40	0	0	0
3	4.19	0.3606	0.1301	1
3	4.70			
4	3.81			
4	4.87	0.8571	2.2041	3
4	2.93			
4	4.52			
			2.3585	6

Asociados con la *SCEp* están sus grados de libertad $glE_p = 6$. El estimado del *error puro* de la variancia ponderada es $\hat{\sigma}_{Ep} = SCE_p/glE_p = 0{,}3931$ y no hace referencia a la función media de regresión lineal. Solo usa el supuesto que la variancia residual es la misma para cada x.

Suponga ahora que se estima una función media de regresión lineal para la data. El ANVA se muestra en la Tabla 4.4, proporcionando un estimado de σ^2 que depende de la función media. Luego, se tienen dos estimados de σ^2 y si el segundo es mucho mayor que el primero el modelo es inadecuado.

Se puede obtener una prueba si la suma de cuadrados residual en la Tabla 4.4 se divide en dos partes: la suma de cuadrados del error puro, dado

en la Tabla 4.3, y la suma de cuadrados de falta de ajuste, o $SCFaj = SCRes - SCEp = 1{,}8581$ con $gl = 8 - 6 = 2$. La prueba F es la razón entre el cuadrado medio para la falta de ajuste y el cuadrado medio del error puro. El $p-$valor obtenido sugiere que no existe falta de ajuste para esta data.

A pesar que el ejemplo en esta sección solo considera un predictor, las ideas usadas para obtener un modelo libre de estimación para σ^2 se pueden generalizar sin problemas. El estimado del error puro de la variancia esta basado en la suma de cuadrados entre los valores de la variable respuesta de los casos que tiene los mismos valores en todos los predictores.

Tabla 4.4: ANVA para la data de la Tabla 4.3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	4.57	4.57	11.62	0.0143
Residuals	8	4.22	0.53		
Lack of fit	2	1.86	0.93	2.36	0.1750
Pure Error	6	2.36	0.39		

4.4. Prueba F general

La teoría para las pruebas F es mucho más general. En su estructura básica, se compara una función media sencilla con una función media general en la H_1. La función media sencilla se puede obtener a partir de la función media general cuando algunos de sus parámetros son cero, iguales a otros, o cuando toman valores específicos. Un ejemplo es aquel que permite probar si los últimos q términos en la función media general son importantes. Las hipótesis en términos matriciales son:

$$\begin{aligned} H_0 &: \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e} \\ H_1 &: \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e} \end{aligned}$$

donde $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ y $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. El modelo más sencillo se obtiene cuando $\boldsymbol{\beta}_2 = \mathbf{0}$.

El procedimiento para obtener la prueba F consiste en estimar las dos regresiones. Luego de estimar la función media en H_0, se calcula $SCRes_0$ y gl_0. Similarmente, bajo la función media alternante se calcula $SCRes_1$ y gl_1

con $SCRes_0 > SCRes_1$ ya que el estimado de H_1 debería ser al menos tan bueno como el estimado de H_0. La prueba F muestra evidencia contra H_0 si:

$$F = \frac{(SCRes_0 - SCRes_1) / (gl_0 - gl_1)}{SCRes_1/gl_1} \tag{4.4.1}$$

Grasa

La data Grasa contiene información que sirve para estimar el porcentaje de grasa en el cuerpo humano en función de la edad, peso, altura, longitud del cuello, longitud del pecho, longitud del abdomen, longitud de la cadera, longitud del muslo, longitud de la rodilla, longitud del tobillo, longitud del bíceps, longitud del antebrazo y longitud de la muñeca. Se tomaron las mediciones anteriores en una muestra de 252 sujetos.

El primer paso es estimar el modelo de regresión lineal múltiple que considera todas las variables predictoras. Las pruebas de significación sugieren que las variables Abdomen, Muñeca, Cuello y Antebrazo son importantes. Sin embargo los $p-$valores para las variable Edad y Peso superan ligeramente el 5 %.

Tabla 4.5: Regresión para la data `Grasa`

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-18.1885	17.3486	-1.05	0.2955
Edad	0.0621	0.0323	1.92	0.0562
Peso	-0.0884	0.0535	-1.65	0.0998
Altura	-0.0696	0.0960	-0.72	0.4693
Cuello	-0.4706	0.2325	-2.02	0.0440
Pecho	-0.0239	0.0991	-0.24	0.8100
Abdomen	0.9548	0.0864	11.04	0.0000
Cadera	-0.2075	0.1459	-1.42	0.1562
Muslo	0.2361	0.1444	1.64	0.1033
Rodilla	0.0153	0.2420	0.06	0.9497
Tobillo	0.1740	0.2215	0.79	0.4329
Biceps	0.1816	0.1711	1.06	0.2897
Antebrazo	0.4520	0.1991	2.27	0.0241
Muñeca	-1.6206	0.5349	-3.03	0.0027

El siguiente paso será comparar el modelo de regresión que incluye las variables predictoras significativas al 10 % con el modelo de regresión que además incluye a las variables Edad y Peso.

Tabla 4.6: Prueba F para la comparación de modelos en la data `Grasa`

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	247	5029.56				
2	245	4559.24	2	470.33	12.64	0.0000

Usando la prueba F se llega a determinar que el modelo que incluye la variable predictora Edad y Peso es mejor.

4.5. Regiones conjuntas de confianza

Las regiones conjuntas de confianza para un grupo de parámetros también requieren del uso de la distribución F y tiene forma elípticas. La región conjunta de $(1-\alpha)\times 100\,\%$ de confianza para $\boldsymbol{\beta}_p$ es el conjunto de vectores tales que:

$$\frac{(\boldsymbol{\beta}_p-\hat{\boldsymbol{\beta}}_p)'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}_p-\hat{\boldsymbol{\beta}}_p)}{(p+1)\hat{\sigma}^2}\leq F_{\alpha,p+1,n-(p+1)} \tag{4.5.1}$$

Por ejemplo, la región de confianza del 95 % para (β_1,β_2) en la regresión de log(Fertilidad) sobre logPBIpc y Purban en la data UN se muestra en la Figura 4.2. La elipse esta centrada en $(-0{,}1255;-0{,}0035)$ que son los respectivos coeficientes estimados en la ecuación de regresión. La orientación de la elipse dada por la dirección del eje mayor es negativa reflejando la correlación negativa que existe entre los estimados de estos coeficientes.

La línea horizontal y vertical en el gráfico determinado por la intersección del eje mayor de la elipse corresponde a los intervalos de confianza al 95 % para cada parámetro por separado. A partir del gráfico, se puede observar que algunos valores de los coeficientes que se encuentran dentro de la región de confianza del 95 % podrían ser considerados como impensables si solo se examinan los intervalos marginales.

Tabla 4.7: Intervalos de confianza individuales para la data UN

	2.5 %	97.5 %
(Intercept)	2.30	2.88
log2(PBIpp)	-0.16	-0.09
Purban	-0.01	0.00

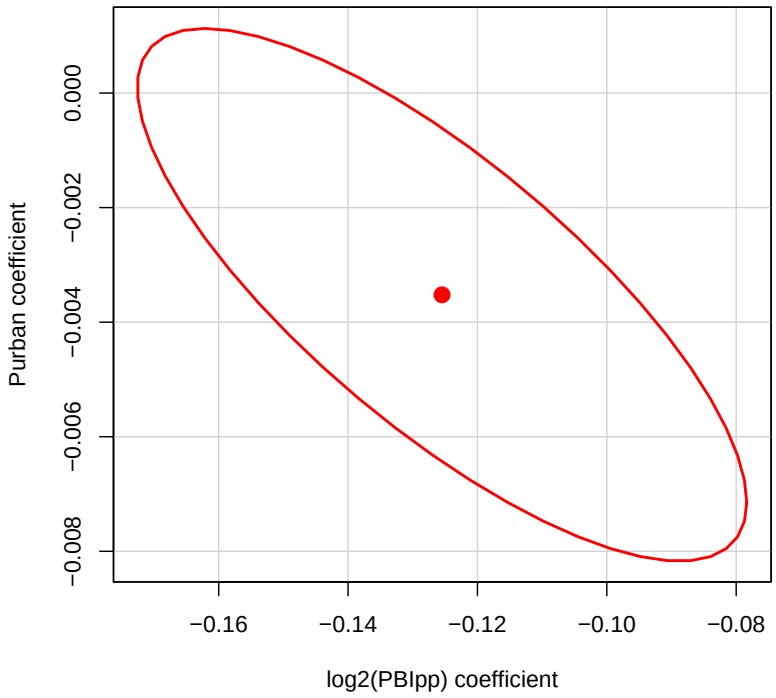

Figura 4.2: Región de confianza del 95 % para la data UN