Práctica en R

Taller 2

Jorge M. Orozco

Métodos Cuantitativos

Facultad de Ciencias Sociales
Universidad Alberto Hurtado

August 9, 2023

uah/
Facultad de Ciencias Sociales
Universidad Alberto Hurtado

1 Análisis de datos con R

August 9, 2023

Problem 1

Ocupe el modelo logit en los datos pruebas_alumnos.xlsx.

Debe predecir si los alumnos necesitan la clase de repaso en función de su sexo y el resultado del examen de lectura.

Para no complicarse, una vez cargado los datos deben aplicar las siguientes funciones:

```
datos$sexo <- as.factor(datos$sexo)
datos$clases_repaso <- as.factor(datos$clases_repaso)
```

1. Ejecute el modelo usando las funciones que vimos en clases para el logit.
2. ¿Cuáles variables salieron significativas?
3. Obtenga los intervalos de confianza de los ponderadores, usando confint.
4. Gráfique la variable clases_repaso versus la probabilidad predicha.
5. En función de la anterior pregunta, ¿Cómo aproximará las variables si usan un punto de corte en la probabilidad 0.5?
6. Usando la información de la pregunta 5, establezca el punto de corte óptimo según su criterio, jueguesela como analista, y pruebe como predice las variables, usando el criterio que debe tener clases_repaso si la probabilidad mayor o igual que el umbral que Usted defina.
7. Muestre la matriz de confusión que resultan con el punto de corte óptimo. Use la función confusionMatrix de la librería caret. Así obtendrá los indicadores de la tabla. https://www.digitalocean.com/community/tutorials/confusion-matrix-in-r
8. Gráfique la matriz de confusión usando:

```
llibrary(ggplot2)
mosaic(matriz_confusion, shade = T, colorize = T,
  gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2,
2)))
```

9. En función de la pregunta 6, ¿cuántos falsos negativos predice? ¿Cuántos porcentualmente?
10. ¿Cuál es el Accuracy del punto de corte que Usted estableció?

Solution. 1. La especificación del modelo es:

```
modelo_logit<-glm(clases_repaso~sexo,examen_lectura, data=pruebas_alumnos,
  family="binomial")
summary(modelo_logit)
```

2. Los resultados del modelo fueron los siguientes:

Table 1: Resumen del Modelo Logit

	Dependent variable:
	clases_repaso
sexomujer	−0.647**
	(0.325)
examen_lectura	−0.026**
	(0.012)
Constant	1.184
	(0.786)
Observations	189
Log Likelihood	-112.319
Akaike Inf. Crit.	230.639
Note:	*p<0.1; **p<0.05; ***p<0.01

Siendo variables significativas la variable de sexo femenino y la puntuación en el examen de lectura.

3. Los intervalos de confianza de los ponderadores son:

Table 2: Intervalos de confianza

	2.5 %	97.5 %
(Intercept)	−0.297	2.799
sexomujer	−1.293	−0.016
examen_lectura	−0.052	−0.004

Para el coeficiente de (Intercept), el intervalo de confianza del 95% es de -0.297 a 2.799. Esto significa que estamos 95% seguros de que el verdadero valor del coeficiente en la población estará dentro de este rango.

Para sexomujer, el intervalo de confianza del 95% es de -1.293 a -0.016, lo que indica que el género femenino tiene un efecto significativo en la probabilidad de necesitar clases de repaso. Similarmente, para examen_lectura, el intervalo de confianza del 95% es de -0.052 a -0.004, lo que sugiere que a medida que el resultado del examen de lectura disminuye, la probabilidad de necesitar clases de repaso aumenta.

4. El gráfico solicitado es:

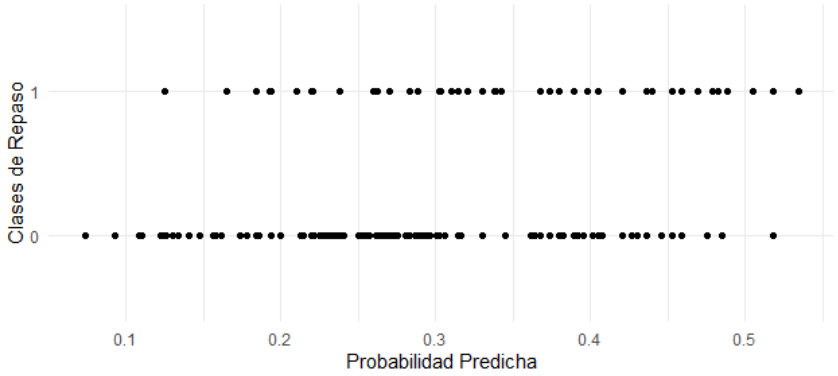

Figure 1: Relación entre la probabilidad predicha y las clases de repaso

5. La aproximación de las variables con un punto de corte 0.5 da el siguiente resultado:

	0	1
0	129	56
1	1	3

Verdaderos Negativos (VN): 129. Esto significa que 129 observaciones fueron correctamente clasificadas como ”0” (no necesitan repaso) por el modelo.

Falsos Positivos (FP): 56. Esto significa que el modelo predijo que 56 observación necesitaba repaso (clase ”1”), pero en realidad no lo necesitaba.

Falsos Negativos (FN): 1. Esto significa que el modelo predijo incorrectamente que 1 observaciones no necesitaban repaso (clase ”0”), pero en realidad sí lo necesitaban.

Verdaderos Positivos (VP): 3. Esto significa que 3 observaciones fueron correctamente clasificadas como necesitando repaso (clase ”1”) por el modelo.

6. Punto de corte óptimo personal: 0.45

	0	1
0	122	45
1	8	14

Verdaderos Negativos (VN): 122. Esto significa que 122 observaciones fueron correctamente clasificadas como ”0” (no necesitan repaso) por el modelo.

Falsos Positivos (FP): 45. Esto significa que el modelo predijo que 45 observación necesitaba repaso (clase ”1”), pero en realidad no lo necesitaba.

Falsos Negativos (FN): 8. Esto significa que el modelo predijo incorrectamente que 8 observaciones no necesitaban repaso (clase ”0”), pero en realidad sí lo necesitaban.

Verdaderos Positivos (VP): 14. Esto significa que 14 observaciones fueron correctamente clasificadas como necesitando repaso (clase ”1”) por el modelo.

7. Matriz de confusión:

	0	1
0	122	45
1	8	14

Accuracy	0.7196
95% CI	(0.6498, 0.7824)
No Information Rate	0.6878
P-Value [Acc > NIR]	0.1947
Kappa	0.2121
Mcnemar's Test P-Value	7.615e-07
Sensitivity	0.9385
Specificity	0.2373
Pos Pred Value	0.7305
Neg Pred Value	0.6364
Prevalence	0.6878
Detection Rate	0.6455
Detection Prevalence	0.8836
Balanced Accuracy	0.5879
'Positive' Class	0

La matriz muestra cómo las predicciones del modelo se comparan con las clases reales. Los valores en la diagonal principal (122 y 14) son las predicciones correctas (verdaderos positivos y verdaderos negativos). Los valores fuera de la diagonal son las predicciones incorrectas.

Accuracy: La proporción total de predicciones correctas en relación con todas las predicciones. En este caso, la precisión es 0.7196, lo que significa que el 71.96% de las predicciones son correctas.

95% CI (Confidence Interval): Un intervalo de confianza al 95% para la precisión. Está en el rango de (0.6498, 0.7824), lo que significa que estamos 95% seguros de que la precisión real está en ese intervalo.

Kappa: La estadística kappa mide la concordancia entre las predicciones del modelo y las clases reales, teniendo en cuenta la concordancia que se podría esperar solo por azar. Un valor de kappa cercano a 1 indica una alta concordancia.

Sensitivity: La proporción de verdaderos positivos en relación con todos los casos positivos reales (clase 1).

Specificity: La proporción de verdaderos negativos en relación con todos los casos negativos reales (clase 0).

Pos Pred Value (Positive Predictive Value): La proporción de verdaderos positivos en relación con todas las predicciones positivas del modelo.

Neg Pred Value (Negative Predictive Value): La proporción de verdaderos negativos en relación con todas las predicciones negativas del modelo.

Prevalence: La proporción de casos positivos en el conjunto de datos.

Detection Rate: La proporción de verdaderos positivos en relación con todos los casos positivos reales.

Detection Prevalence: La proporción de casos positivos predichos por el modelo en relación con el total de casos.

Balanced Accuracy: El promedio aritmético de la sensibilidad y la especificidad.

'Positive' Class: La clase considerada como positiva en la evaluación (clase 1 en este caso).

el modelo tiene una sensibilidad alta (0.9385), lo que indica que es bueno para detectar casos positivos, pero tiene una especificidad baja (0.2373), lo que significa que no es tan bueno para identificar casos negativos.

8. Matriz de confusión gráficamente:

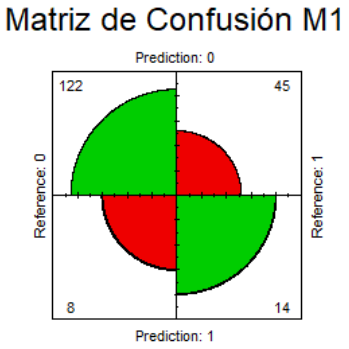

Figure 2: Matrix Confusion

9. En función de la pregunta 6, ¿cuántos falsos negativos predice? ¿Cuántos porcentualmente?

Hay 8 falsos negativos. Y porcentualmente hablando corresponden al 36,36%.

10. ¿Cuál es el Accuracy del punto de corte que Usted estableció?

La Accuracy es 0,72.