UNAP
FINESI

IV Semestre

Modelos Discretos

Inferencia de los Coeficientes del Modelo

Mtr. Alcides Ramos Calcina

Inferencia de los Coeficientes del Modelo

Mtr. Alcides Ramos Calcina

1. Prueba de la significancia de los coeficientes

Suponiendo que se cumple el modelo de regresión logística, estamos interesados en determinar si la variable "EDAD" es significativa para explicar "enfermedad cardiaca coronaria (ECC)". Planteamos entonces el siguiente contraste:

H_0: $\beta_1 = 0$ (la variable "EDAD" no es significativa)
H_1: $\beta_1 \neq 0$ (la variable "EDAD" es significativa)

a) Estadístico de Wald

En este primer caso, utilizamos el estadístico de Wald, el cual se obtiene dividiendo

$$Wald\ z = \frac{\beta_i}{S_{\beta_i}}$$

1. Prueba de la significancia de los coeficientes

- En R se muestra este estadístico de Wald *z*, que es la raíz cuadrada de la prueba de Wald χ^2 visto en la sección anterior.

- Esto se basa en el hecho matemático de que, si a la distribución normal estándar Z la elevamos al cuadrado, se obtiene una distribución chi-cuadrada con $df = 1$.

$$W = \left(\frac{\beta_i}{S_{\beta_i}} \right)^2 \approx \chi_1^2$$

Solución

Continuando con el ejemplo 1.

Observe que el estadístico z dado en R es la estimación dividida por el error estándar y el p-valor se basa en la distribución normal estándar.

$$Wald\ z = \frac{\beta_i}{S_{\beta_i}} = \frac{0.11092}{0.02406} = 4.610$$

Existe una relación significativa entre EDAD y la ECC, con Wald z = 4.6109 y un valor de probabilidad asociado al estadístico de Wald de $p = 0.000004$.

Solución

Dado que esta estadística es chi-cuadrado con *df* = 1, el valor de *p* es:

```
p_value_Chi <- 1 - pchisq(21.25340, df = 1)
p_value_Chi
[1] 4.02396e-06
```

Debido a que el p-value = 0.000004 < 0.05, se rechaza la hipótesis nula (H_0), por tanto, la variable EDAD si es significativa.

1. Prueba de la significancia de los coeficientes

b) Devianza

La devianza puede interpretarse de manera análoga a la Suma Total de Cuadrados de una regresión lineal.

En el caso de la regresión logística, se reportan dos devianzas, una para un modelo "nulo" (`Null deviance`), que incluye únicamente la intercepción y ninguna variable independiente; otra que es la devianza residual (`Residual deviance`), es decir la que queda luego de haber incluido las variables independientes del modelo.

```
    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35
```

Solución

Continuando con el ejemplo 1.

Cuanto mayor sea la diferencia entre ambas devianzas, mejor será el ajuste del modelo.

Diferencia: $\chi^2 = 2LL(nulo) - 2LL(residual) = 136.66 - 107.35 = 29.31$

Es posible usar una prueba de Chi Cuadrado para evaluar si la diferencia entre las devianzas es estadísticamente significativa.

Solución

```
anova(modelo, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: ECC

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    99     136.66
EDAD  1    29.31        98     107.35 6.168e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La variable EDAD (edad) permite reducir la *deviance residual* en 29.31, con 1 grado de libertad, lo que supone una significativa reducción, con *p*-valor igual a $0.0000 < 0.05$.

Se concluye que el término es importante en el modelo

2. Intervalo de confianza de la estimación

Los extremos de un intervalo de confianza de (1 - α)% para el coeficiente de pendiente es:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\,\widehat{SE}\left(\hat{\beta}_1\right)$$

y para el intercepto es

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\,\widehat{SE}\left(\hat{\beta}_0\right)$$

donde

$z_{1-\alpha/2}$: es el punto crítico de la distribución normal estándar.

$\widehat{SE}(\ )$: denota un estimador basado en el error estándar del modelo del respectivo parámetro.

Solución

Continuando con el ejemplo 1.

Para el ejemplo, el intervalo de confianza de Wald para β_1 se calcula de la siguiente manera:

$$IC: P\left[\hat{\beta}_1 - z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + z_{1-\alpha/2}\widehat{SE}\left(\hat{\beta}_1\right)\right] = 1-\alpha$$

Con: $z_{1-\alpha/2} = \pm 1.96$

Se tiene:

$$P\left[0.11092 - (1.96)(0.02406) \leq \beta_1 \leq 0.11092 + (1.96)(0.02406)\right] = 0.95$$

$$IC: P\left[0.064 \leq \beta_1 \leq 0.158\right] = 0.95$$

Solución

En R.

```
confint(modelo)
                    2.5 %      97.5 %
(Intercept) -7.72587162 -3.2461547
EDAD         0.06693158  0.1620067
```

Exponencia este intervalo para obtener un intervalo de confianza para la razón de probabilidades.

```
exp(confint(modelo))
                    2.5 %    97.5 %
(Intercept) 0.0004412621 0.0389236
EDAD        1.0692223156 1.1758681
```

Evaluación del Modelo

Mtr. Alcides Ramos Calcina

Evaluación del Modelo

- A la hora de evaluar la validez y calidad de un modelo de regresión logística, se analiza tanto el modelo en su conjunto como los predictores que lo forman.

- Como se sabe en la regresión lineal simple y múltiple, el coeficiente de determinación R^2 es una medida muy intuitiva de lo bien que el modelo predice la variable dependiente, pues es la parte de la varianza total explicada por las variables independientes.

- Desgraciadamente no hay un equivalente tan intuitivo en la regresión logística.

- Veamos algunas propuestas:

1. Método de Máxima Verosimilitud

- Basado en las observaciones de si el evento ocurre o no para un sujeto.
- Así, para ese *i*-ésimo sujeto el suceso Y toma los valores 0 (no ocurre) o 1 (ocurre), y el valor predicho, $P[Y = k \mid X = x]$, variará entre 0 y 1.

$$\log - likelihood = -2\ln\left(Lik\left(H_0\right)\right) - 2\ln\left(Lik\left(H_A\right)\right) \approx \chi_k^2$$

- El estadístico $\log - likelihood$ es análogo a la suma de cuadrados residual en la regresión múltiple en el sentido de que es un indicador cuánta información sin explicar queda en la variable respuesta tras haber ajustado el modelo.
- Grandes valores del $\log - likelihood$ indican un pobre ajuste del modelo, cuanto mayor sea este valor, más variabilidad sin explicar queda en el modelo.

2. Pseudo R^2

Una medida análoga al R^2 en la regresión logística puede ser:

$$R_L^2 = \frac{2LL(nuevo) - 2LL(referencial)}{2LL(referencial)}$$

- Es la reducción proporcional en valor absoluto de $log-likelihood$ y mide cuánto del error del ajuste disminuye al incluir las variables predictoras.
- Proporciona una medición de la significación real del modelo.
- Esta puede variar entre 0 (indicando que los predictores son inútiles prediciendo la variable respuesta) y 1 (indicando que el modelo predice perfectamente la respuesta).

2. Pseudo R^2

R^2 de **Cox y Snell**. (1989)

Está basado en la devianza de el modelo ($-2LL(nuevo)$) y la devianza del modelo base o referencia ($-2LL(referencia)$), y el tamaño de la muestra n.

$$R^2_{CS} = 1 - \exp\left(\frac{\left(-2LL\left(nuevo\right)\right) - \left(-2LL\left(referencial\right)\right)}{n}\right)$$

Nagelkerke (1991)

El estadístico de Cos y Snell nunca alcanza su máximo teórico de 1. Por lo tanto, sugirió la siguiente enmienda:

$$R^2_N = \frac{R^2_{CS}}{1 - \exp\left(-\frac{-2LL\left(referencial\right)}{n}\right)}$$

Solución

Continuando con el ejemplo 1.

El R tiene diversas funciones que permiten calcularlos, pero es necesario instalar la librería **DescTools**, para utilizar la función PseudoR2().

```
# install.packages("DescTools")
# library(DescTools)
PseudoR2(modelo, c("CoxSnell", "Nagel"))
  CoxSnell Nagelkerke
 0.2540516  0.3409928
```

El R^2 más alto es 0.341 ≈ 34.1% (R Nagelkerke x 100) es decir que, solo un 34.1% de la enfermedad cardiaca coronaria es explicada por la variable edad, por tanto, hay un 65.9% que no está explicado por la variable edad.

3. Test de Hosmer - Lemeshow

La prueba de Hosmer-Lemeshow (prueba HL) es una prueba de bondad de ajuste para modelos de clasificación binaria que indica qué tan bien se ajustan los datos a un modelo determinado y se examina mediante la Chi-cuadrada de Pearson.

La idea es agrupar las observaciones en categorías sobre la base de sus probabilidades previstas en una serie de grupos y se calcula con la siguiente fórmula:

$$G_{HL} = \sum_{j=1}^{g} \frac{\left(o_j - e_j\right)^2}{e_j} \sim \chi^2_{(g-2)g.l.}$$

Donde:

o_j : es el número de éxitos y fracasos observados

e_j : es el número de éxitos y fracasos esperados para un grupo dado j.

g : es el número de subgrupos, generalmente se calcula: $g > P + 1$, con P como el número de parámetros en un modelo dado.

3. Test de Hosmer - Lemeshow

Si el modelo de interés se ajusta bien a los datos $G_{HL} \sim \chi^2_{(g-2)g.l.}$

Esto significa que cuando usamos esta prueba, nuestro objetivo es tener un valor *p* grande para aceptar nuestro modelo.

Se plantea la siguiente hipótesis:

H_0: $o_j = e_j$ (El modelo se ajusta a los datos)
H_1: $o_j \neq e_j$ (El modelo no se ajusta a los datos)

Solución

Continuando con el ejemplo 1.

Para solicitar el test en R, necesitamos instalar la librería: **ResourceSelection**.

```
# install.packages("ResourceSelection")
# library(ResourceSelection)
hoslem.test(datos$ECC, fitted(modelo))
```

```
        Hosmer and Lemeshow goodness of fit (GOF)
        test

data:  datos$ECC, fitted(modelo)
X-squared = 2.2243, df = 8, p-value = 0.9734
```

Como p-value = 0.9734 > 0.05, entonces aceptamos la hipótesis nula, es decir el modelo de regresión logística se ajusta a los datos.

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS