Primer taller

jueves, 12 de junio de 2025 11:09 a. m.

Ejercicio con datos reales

Considere el siguiente conjunto de datos que agrupa una serie de características enfocadas en *clasificar* la personalidad de múltiples individuos. **Se incluyen variables cuantitativas y cualitativas.** La información puede profundizarse en: https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data

Table 1: Información en análisis

Personality	Time_spent_Alone	Stage_fear	Social_event_attendance	Going_outside
Extrovert	4	No	4	6
Introvert	9	Yes	0	0
Introvert	9	Yes	1	2
Extrovert	0	No	6	7
Extrovert	3	No	9	4

Considere a *'Personality'* como la variable respuesta. *Algunas de las covariables en análisis se especifican en la tabla mostrada con anterioridad.* **De respuesta a los siguientes planteamientos:**

X_1: Tiempo solo
X_2: Sociales
X_3: Salir
X_4: Amigos
X_5: Publicaciones

W: Patrón asistencia. "Yes" ; "No"

$W = \begin{cases} 1, & \text{"No"} \\ 0, & \text{"Yes"} \end{cases}$

Y_i: Personalidad

$Y_i = \begin{cases} 1, & \text{"introvertido"} \\ 0, & \text{"extrovertido"} \end{cases}$

- (1). $Y_i \sim Ber(\theta)$: $f(y_i | X_{i1}, \dots, X_{i6}) = \theta^{y_i}(1-\theta)^{1-y_i}$; θ = Probabilidad de éxito evento

Y_i: Personalidad de la i-ésima persona ; $i = 1, \dots, n$

- Logístico: Clasificaciones

(1). Regresor lineal: $\ln(\psi(x_i)) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W$

(2). Probabilidad de éxito i: θ_i

(3). Función de enlace: $\text{logit}(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right)$ → odds (chances)

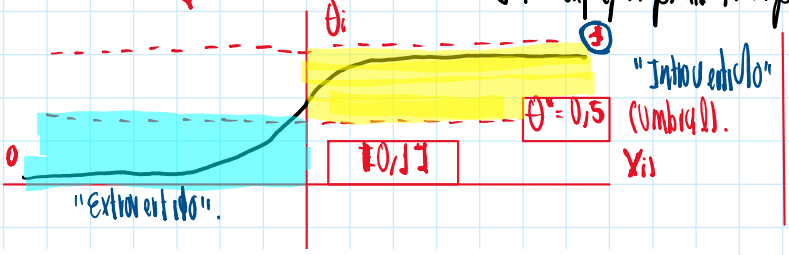
odds

- $\frac{\theta_i}{1-\theta_i} = \frac{P(\text{éxito})}{P(\text{fracaso})}$ (1);

Función log-odds:

Construcción modelo: $\text{logit}(\theta_i) = \ln(\psi(x_i)) = \ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W_i$

- $\theta_i = P(y_i = 1 | X_{i1}, \dots, X_{i5}, W) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W_i)}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W_i)} = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W_i)]}$



$\frac{\psi(X_{i1}+1)}{\psi(X_{i1})} = \exp(\beta_1)$ $j = 1, \dots, k$ $j = 1, \dots 6$.

Interpretar el cambio en las "odds"

1. Determine el valor de verdad de las siguientes afirmaciones.

- (a) En un modelo de regresión logística, la variable respuesta es de naturaleza binaria, tal que $Y_i \sim$ Ber(θ_i); donde θ es la probabilidad de éxito en el i- ésimo ensayo. Verdadero
- (b) Los coeficientes β de un modelo de regresión logística se estiman numéricamente el método de mínimos cuadrados ordinarios, ya que la función de verosimilitud es no- lineal. Máxima verosimilitud Falso
- (c) Al incluir variables categóricas al modelo, es necesario definir una categoría de referencia, y la interpretación de tales coeficientes se realiza en comparación con tal referencia. Verdadero
- (d) El valor $\exp(\beta_j)$ se interpreta como el cambio en el logaritmo de los *'odds'* por cada incremento unitario en la variable predictora X_j. Falso Falso
- (e) El estadístico de prueba $\chi_c = D_0^2 - D^2$, con una región de rechazo asociada $R_c = \{\chi_c > \chi_{p,\alpha}\}$ se emplea para probar la significancia global del modelo, mientras que se prefiere la prueba *t- student* para la significancia individual.

$Z_{\alpha/2}$

1. Ajuste un modelo de regresión logístico. Analice los componentes $\theta_i, g(\theta_i), \psi(\mathbf{x}_i)$. ¿Por qué se emplea la función *'logit'* como función de enlace?

```
# Ajuste del modelo de regresión
datos ← read.csv(file.choose())
modelo ← glm(Personality ~ ., data = datos, family = binomial(link = "logit"))
```

$\text{logit}(\theta_i) = \ln(\psi(x_i)) = \ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 W_i$

2. Realice una prueba de significancia general del modelo de regresión logístico propuesto. Interprete los resultados obtenidos.

$\begin{cases} H_0: \beta_1 = \dots = \beta_6 = 0 \\ H_1: \text{Algún } \beta_j \neq 0 ; j = 1, \dots, 6 \end{cases}$

$\chi_c = D_0^2 - D^2$; $R_c = \{\chi_c > \chi_{p,\alpha}\}$

$\text{logit}(\theta_i) = \beta_0$

```
# ——————————————
#        SEGUNDO PUNTO
# ——————————————
# Modelo reducido
modelo_reducido ← glm(Personality ~ 1, data = datos, family = binomial(link = "logit"))
# Considerar la prueba con anova
anova(modelo_reducido, modelo, test = "Chisq")
```

```
> anova(modelo_reducido, modelo, test = "Chisq")
```

- Conclusión: el modelo es significativo.

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

Model 1: Personality ~ 1
Model 2: Personality ~ Time_spent_Alone + Stage_fear + Social_event_attendance +
    Going_outside + Friends_circle_size + Post_frequency
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2522     3496.2
2      2516     1227.0  6   2269.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Realice una prueba de significancia de los parámetros individuales. Provea una *interpretación directa* de los parámetros del modelo de regresión logística, así como una *interpretación en términos de la razón de probabilidades.*

$\begin{cases} H_0: \beta_j = 0 \; ; \; j = 0,1,\dots,6 \\ H_1: \beta_j \neq 0 \end{cases}$ Considerar $N(0,1)$: $Z = \dfrac{|\hat{\beta}_j|}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}$

```
> #         TERCER PUNTO
> # ------------------------------------------
> summary(modelo)

Call:
glm(formula = Personality ~ ., family = binomial(link = "logit"),
    data = datos)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -8.05656    0.79539 -10.129  < 2e-16 ***
Time_spent_Alone       -0.23651    0.04504  -5.251 1.51e-07 ***
Stage_fearYes          11.17932    0.75636  14.780  < 2e-16 ***
Social_event_attendance  0.19931    0.05664   3.519 0.000434 ***
Going_outside            0.32057    0.07851   4.083 4.44e-05 ***
Friends_circle_size      0.18036    0.03714   4.856 1.19e-06 ***
Post_frequency           0.11784    0.05587   2.109 0.034931 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3496.2  on 2522  degrees of freedom
Residual deviance: 1227.0  on 2516  degrees of freedom
AIC: 1241

Number of Fisher Scoring iterations: 6
```

- A un nivel de significancia del 5%, se puede determinar que cada variable de forma individual es significativa (manteniendo las demás efectos constantes).

(1). Interpretación clásica: (log-odds)

- $\hat{\beta}_2$: Por un cambio en la covariable X_2 (Tiempo solo), el cambio en el log-odds es de $-0,2365$ Unidades. (manteniendo las demás efectos constantes.

$\ln\left(\dfrac{\theta_i}{1-\theta_i}\right) = \ln(\theta_i) - \ln(1-\theta_i)$

(2). Interpretación odds: $\dfrac{\theta_i}{1-\theta_i}$

- $\exp(\hat{\beta}_j)$: $\exp(\hat{\beta}_3)$ =

```
> exp(0.32057)
[1] 1.377913
```

- De forma porcentual: $(\exp(\hat{\beta}_3) - 1) = 0,377$: $37,7\%$: Las odds van a multiplicarse en un 37,7%.

- $\dfrac{\theta_i}{1-\theta_i}$; $1,377 > 1$: Las probabilidades de éxito θ_i aumentan en relación a las probabilidades de fracaso $1-\theta_i$

Cambio en θ_i se reduce $<$ $\begin{cases} \beta_j > 0 \\ \beta_j < 0 \end{cases}$

- Intervalo confianza:

```
> confint(modelo)
Waiting for profiling to be done...
                            2.5 %      97.5 %
(Intercept)            -9.650801154 -6.5303979
Time_spent_Alone       -0.326373316 -0.1495924
Stage_fearYes           9.732570149 12.6096052
Social_event_attendance 0.089082648  0.3114096
Going_outside           0.168113749  0.4762027
Friends_circle_size     0.108570987  0.2543500
Post_frequency          0.009004766  0.2283594
```

- Ninguno intervalo contiene al cero (todos los parámetros son significativos).

$\hat{\beta}_j \pm Z_{\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{\beta}_j)}$

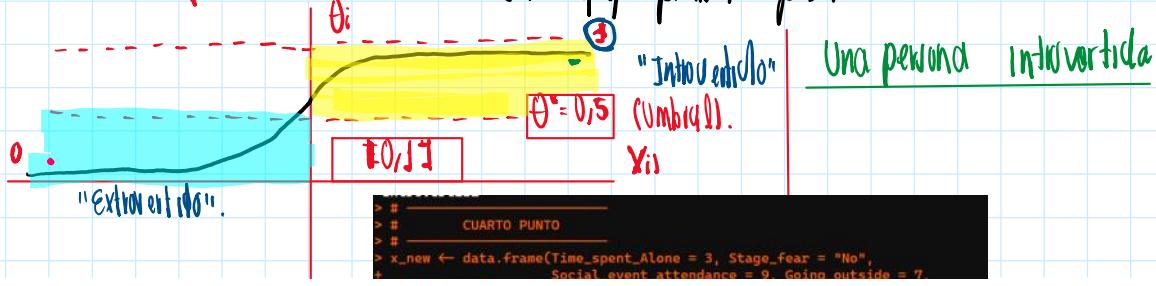
4. Realice una predicción de futuras observaciones. Utilice las probabilidades obtenidas para la construcción un clasificador logístico con etiquetas.

```
> #         CUARTO PUNTO
> # ------------------------------------------
> x_new <- data.frame(Time_spent_Alone = 3, Stage_fear = "Yes",
+                     Social_event_attendance = 8, Going_outside = 7,
+                     Friends_circle_size = 4, Post_frequency = 8)
> predict.glm(modelo, newdata = x_new, type = c("response"))
        1
0.9996352
```

θ_i : θ_1

```
> probability <- predict.glm(modelo, newdata = x_new, type = c("response"))
> resultado <- ifelse(probability > 0.5, "Introvertida", "Extrovertida")
> resultado
          1
"Introvertida"
```

- $\theta_i = P(y_i = 1 \mid X_{i1}, \dots, X_{i5}, w) = \dfrac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_5 X_{i5} + \beta_6 w_i)}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_6 w_i)}$



Una persona introvertida

```
> #         CUARTO PUNTO
> # ------------------------------------------
> x_new <- data.frame(Time_spent_Alone = 3, Stage_fear = "No",
+                     Social_event_attendance = 9, Going_outside = 7,
```

"Extrovertido".

```
> # ------------------------
> #         CUARTO PUNTO
> # ------------------------
> x_new ← data.frame(Time_spent_Alone = 3, Stage_fear = "No",
+                    Social_event_attendance = 9, Going_outside = 7,
+                    Friends_circle_size = 10, Post_frequency = 10)
> probability ← predict.glm(modelo, newdata = x_new, type = c("response"))
> resultado ← ifelse(probability ≥ 0.5, "Introvertida", "Extrovertida")
> resultado
          1
"Extrovertida"
```