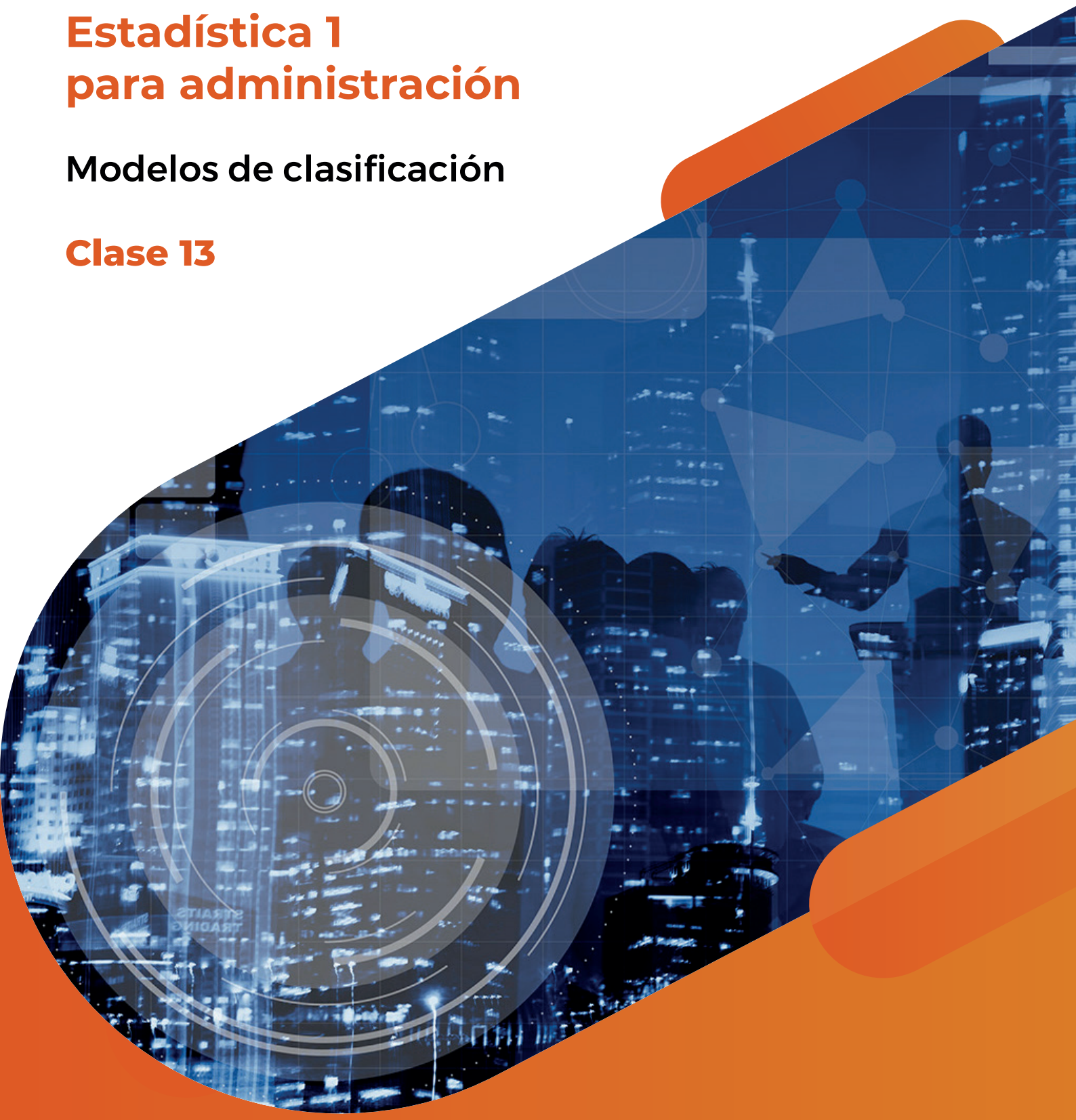
PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Modelos de clasificación

Clase 13



La excelencia no se improvisa

Introducción de la clase

En esta clase trataremos sobre modelos de clasificación, hasta el momento todos los modelos fueron desarrollados con variables de respuesta continuas, sin embargo, más comunes son los modelos de clasificación, existen una gran cantidad de modelos de clasificación en la ciencia de datos, pero en esta clase nos centraremos específicamente en modelos de regresión logística.

Dentro de estos modelos será muy importante el tratar las distintas técnicas de codificación de variables categóricas. En modelos de clasificación de la ciencia de datos no se admiten trabajar con variables categóricas directamente por lo que requieren algún tipo de codificación, por lo tanto, en esta segunda parte vamos a mencionar algunos mecanismos para codificar variables categóricas.

Resultado o resultados de aprendizaje que será abordado con el contenido de la clase
Presentar informes estadísticos interpretando los resultados dentro del contexto del problema.

Clase 13: Modelos de clasificación

Cuando hablamos de modelos cuya variable de respuesta es discreta o categórica nos referimos como modelos de clasificación. Si la variable de respuesta solo puede tomar 2 valores (verdadero/falso o 0/1) nos referimos como modelos binomiales, por el contrario, si puede tomar conjunto de valores nos referimos como modelos multinomiales. En la presente clase veremos modelos logísticos binomiales.

Modelo logístico

La regresión logística es un modelo que se utiliza para determinar la probabilidad de que ocurra un evento. Estos modelos muestran la relación entre variables y luego calcula la probabilidad de un resultado determinado.
Para la creación de este tipo de modelos recurrimos a la función glm().

Hasta ahora hemos utilizado la función lm() para generar modelos que tiene determinados supuestos, por ejemplo, distribución debe ser gaussiana o normal y que su varianza debe ser homocedástica. Pero no todos los problemas cumplen con estos supuestos, por ejemplo, ¿Qué pasa si la data refleja una distribución binomial o Poisson? Para esto introduciremos la función glm.

Modelos lineales generalizados

Explica el uso de glm

glm().

Los modelos GLM son una generalización de los modelos lineales vistos, que permiten crear modelos en los cuales algunas condiciones puedan ser especificadas. Una de estas condiciones, supone que cada resultado Y de las variables de respuesta se genera a partir de una distribución particular de la familia exponencial, esta incluye la distribución normal, binomial, Poisson y gamma, entre otras.

Los modelos glm tiene 3 componentes:

Componente aleatorio: Especifica la distribución de probabilidad de la variable de respuesta, específicamente, para una distribución dada (normal, binomial, etc.) estima su media y varianza. No hay un término de error separado (ε) como en la regresión lineal.

Componente Sistemático: Especifica las variables explicativas del modelo y su combinación lineal, esto es similar a la regresión lineal.

Función de enlace: Indica cómo el valor esperado de la respuesta se relaciona con la combinación lineal de variables predictoras. Glm NO asume una relación lineal entre la variable de respuesta y las variables predictoras (como lm), pero sí asume una relación lineal entre la respuesta transformada en términos de la función de enlace y las variables explicativas. Específicamente para regresión logística utilizaremos la familia binomial (family=binomial(link = "logit"))

Note que la distribución está dada por distintas “familias” y entre paréntesis aparece la función “link”, ésta indica justamente la función de enlace que se usará en el proceso.

La familia de binomial dispone de varias funciones de enlace que permiten el comportamiento requerido. La siguiente tabla presenta el nombre de la función y su fórmula de cálculo.

Tabla 1

Funciones de enlace para regresión logística

Describe la función y su cálculo de la probabilidad de la ocurrencia del evento. Prado (2025)

Adaptado de Rindang B (2019)

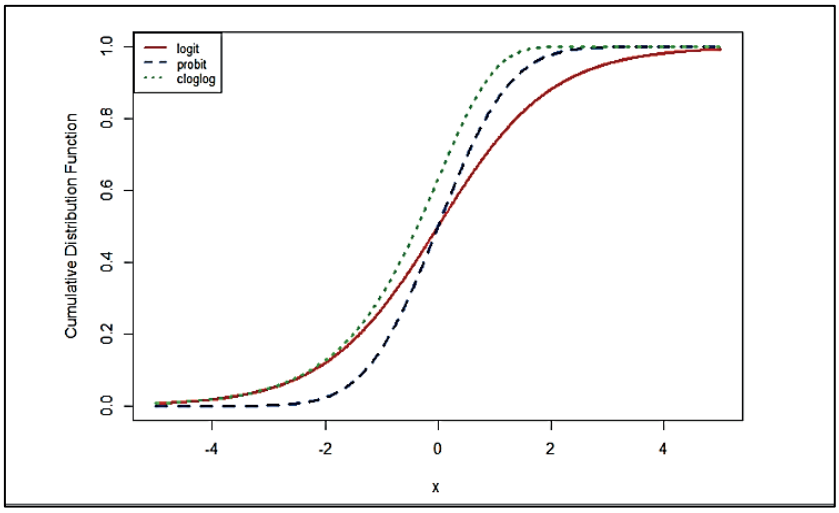
Nombre	Función	Probabilidad
Logit	$z = \log(\frac{p}{1-p})$	$p = \frac{e^z}{1+e^z}$
Probit	$z = \Phi^{-1}(p)$	$p = \Phi(z)$
Cauchit	Na	$p = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$
Cloglog	Na	$p = 1 - e^{-e^z}$

Las funciones de enlace tienen distintas características como se evidencia en la siguiente figura.

Figura 1

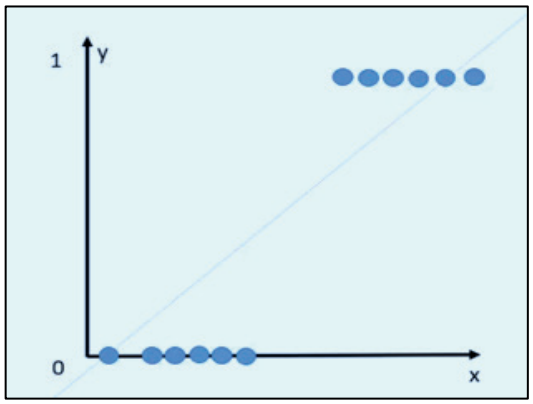
Características de funciones de enlace

Tomado de Rindang B. (2019)



La siguiente figura explica el concepto de la función de enlace. Para cierto rango de valores de la variable predictora X se desea que la respuesta Y sea 0 (o esté debajo de cierto umbral). Sin embargo, cuando X sube de cierto valor se espera que Y sea 1.
Por defecto en el caso de distribución binomial, la función es "logit", que como vemos en la tabla 1, se basa en el concepto de Odds.

Figura 2
Relación entre X y Y en un modelo logístico
Creación del autor Alfonso Prado



Funciones de enlace logit

ODDS y ODDS RATIO

El análisis de datos categóricos se basa típicamente en tablas de contingencia de dos o más dimensiones, tabulando la frecuencia de ocurrencia de niveles de datos nominales y/o ordinales.
Una tabla de contingencia es una herramienta utilizada en la cual se crean al menos dos filas y dos columnas para representar datos categóricos en términos de conteos de frecuencia.
La tabla permite medir la interacción entre dos variables para conocer una serie de información "oculta" de gran utilidad y comprender con mayor claridad los resultados de una investigación.
Ejemplo: En la siguiente figura se presenta la ocurrencia de una enfermedad (mf), pero esta segmentada por otras variables como área, rango-edad, sexo y queremos entender cómo se distribuye esta enfermedad de acuerdo al área (o cualquier otra variable categórica)

Figura 3

Dataset Oncho (a) y tabla de contingencia de la relación mf vs area (b)

Creación del autor Alfonso Prado

(a)

	id ID	mf Infection	area Residence	agegrp Age group (years)	sex Sex	mfload Microfiliariae load	lesions Severe eye lesions?
1	1	Infected	Savannah	20-39	Female	1	No
2	2	Infected	Rainforest	40+	Male	3	No
3	3	Infected	Savannah	40+	Female	1	No
4	4	Not-infected	Rainforest	20-39	Female	0	No

(b)

	area	
mf	Savannah	Rainforest
Not-infected	267	213
Infected	281	541

Note que las filas corresponden a los niveles de la variable Y, sin embargo, en algunos campos de investigación se estila crear la tabla con nombres genéricos como Exposure y Outcome, donde las filas (Exp+ y Exp -) indican haber estado sujeto a un "tratamiento" y las columnas (Out +, Out -) haber desarrollado un efecto o no. Debemos notar que la cantidad de observaciones de un valor nominal no va a ser necesariamente la misma de otro valor. Por ejemplo, la cantidad de encuestados hombres no es igual a las mujeres, por lo tanto, la tabla debe ser procesada a fin de encontrar el valor del odds. El Odds nos indicará la probabilidad entre Exposure y Outcome y se calcula mediante la relación entre outcome+ y outcome – para cada uno de los exposed. Ejemplo: En la figura 12 odds de los expuestos=140/84, odds de no expuestos=139/92.

Figura 4

Obtención del Odds y Odds Ratio mediante la función epi.2by2

Creación del autor Alfonso Prado

```
           Outcome +    Outcome -      Total        Inc risk *        Odds
Exposed +        140           84        224              62.5        1.67
Exposed -        139           92        231              60.2        1.51
Total            279          176        455              61.3        1.59

Point estimates and 95% CIs:
-------------------------------------------------------------------
Inc risk ratio                                 1.04 (0.90, 1.20)
Odds ratio                                     1.10 (0.76, 1.61)
```

Y para comparar los dos grupos (expuestos y no expuestos) dividiremos los valores de cada grupo y esto se denomina el odds ratio u OR = 1.67/1.51 =1.1

¿Cómo interpretar el odds-ratio? La razón de probabilidades nos dice cuánto más altas son, las probabilidades de exposición entre los casos de una salida positiva que entre los casos de una salida negativa.

Si la razón de posibilidades (odds-ratio) es mayor que 1, los eventos se asocian positivamente, si la razón de probabilidades es menor que 1 los eventos se asocian negativamente. Si es cercano a 1 no hay influencia. Podemos entender esto como una correlación entre variables nominales. Note que divide la probabilidad de que se presente el evento para la probabilidad de que no se presente, cuyo logaritmo es utilizado por la función logit. (Ver tabla 1)

Figura 5

Generalización del Odds ratio como relación de probabilidad

Creación el autor Alfonso Prado

$$\pi_1/(1-\pi_1)$$

Donde

Π representa la probabilidad de que se presente el evento

Prueba Chi-cuadrado de independencia

El proceso de tablas de contingencia visto anteriormente funciona bien para variables dicotómicas, pero que pasa si tenemos más de dos niveles de la variable, por ejemplo, del mismo dataset mencionado anteriormente existen variables como agegrp (grupo etario), que tiene 4 niveles. En este caso la función 2by2 no funciona, por lo que tenemos que recurrir a otra prueba de independencia conocida como Chi-cuadrado o Ji-cuadrado.

Tabla 2

Tabla de contingencia mf vs aegrp

Creación del autor Alfonso Prado

```
> Oncho %>%
+     dplyr::select(mf, agegrp ) %>%  table()
              agegrp
mf             5-9 10-19 20-39 40+
  Not-infected 156   119   125  80
  Infected      46    99   299 378
```

En estos casos necesitamos realizar los siguientes pasos:

a) Mostrar que existe una relación para lo cual usaremos la prueba de independencia Xi^2 (se pronuncia chi-cuadrado).
b) La prueba Xi^2 (al igual que otras pruebas) consiste en 2 partes, obtener un estadístico y compararlo con su distribución, en este caso la distribución es Chi cuadrado, y los grados de libertad se calcula como (filas -1)*(columnas -1) de la tabla de contingencia
c) Validar en forma gráfica mediante diagramas de mosaicos
d) Digitalizar o codificar la variable

De acuerdo a Ramirez-Alan ,(2016) "El test Xi^2 considera la hipótesis nula (H0) de que las variables son independientes, si esto es verdad, la frecuencia de ocurrencia debería estar dada por la cantidad de casos totales multiplicada por la probabilidad esperada". En el ejemplo anterior, si la probabilidad de contagiarse de la enfermedad es igual para los 4 niveles (π_{ij} = 25%) entonces si se han detectado 1000 casos estas deberían estar distribuidas en las 4 categorías ($n*\pi_{ij}$ =1000*.25). Este valor conocido como μ_{ij} .

Figura 6

Frecuencia esperada

Creación del autor Alfonso Prado

Frecuencia esperada $\mu_{ij} = n\ \pi_{ij}$

Pero si Ha está en lo correcto va a existir una diferencia entre frecuencia observada (ηij) y la esperada (μij) indicando que existe algún fenómeno por detrás que influencia la frecuencia observada. La prueba Xi^2 (chisq.test) calcula esta diferencia de la siguiente forma:

Figura 7

Fórmula de Xi^2

Creación del autor Alfonso Prado

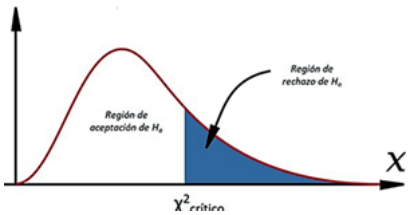
$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

Mientras mayor sea este indicador mayor las probabilidades de que las variables estén relacionadas, por el contrario, el numerador tiende a 0 . X^2 debe ser contrastado contra un Xi-crítico, obtenido en base a la función qchisq, la siguiente figura explica las zonas de aceptación y rechazo de la asociación entre variables para esta distribución

Figura 8

Zonas de aceptación y rechazo de la prueba Xi^2

Creación del autor Alfonso Prado



Continuando con el ejemplo

```
#Obtenemos la tabla de contingencia de agegrp
tab <- table(data$mf, data$agegrp)
```

```
              agegrp
mf             5-9 10-19 20-39 40+
  Not-infected 156   119   125  80
  Infected      46    99   299 378
```

```
#Ejecutamos la prueba Chi2
chisq.test(tab)
```

```
        Pearson's Chi-squared test

data:  tab
X-squared = 254.78, df = 3, p-value < 2.2e-16
```

```
#calculando chi-crítico
```

```
> chi_critico <- qchisq(.05, df=3, lower.tail=FALSE)
> chi_critico
[1] 7.814728
```

Al obtener un estadístico chi^2 de 254 con un chi-critico de 7.81 concluimos que existen diferencias substanciales lo que indica que la variable es significativa, en otras palabras, el indicador se halla en la zona de rechazo de H0.

Función chisq.test

Explica el uso de la función

chisq.test().

Validación Gráfica

Una forma alternativa de validar la significancia de las variables categóricas (factores) con múltiples niveles es mediante un diagrama de mosaico. En estos diagramas el área es directamente proporcional con la frecuencia de ocurrencia de la tabla, y debemos

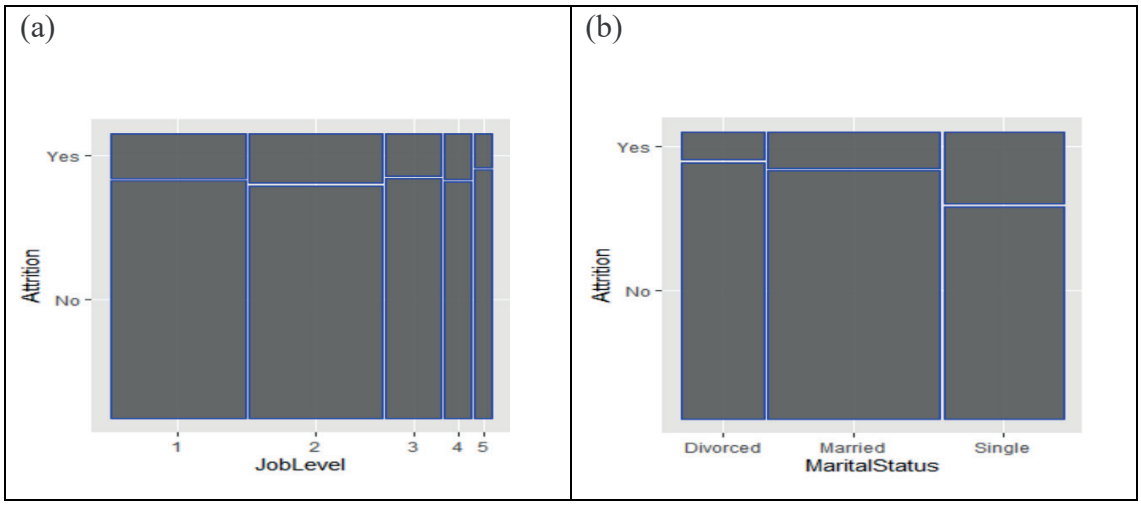
interpretarla de la siguiente forma: Si obtenemos un diagrama muy "regular" indicaría que las dos variables no están relacionadas. La palabra "regular" debemos entenderlo como que las proporciones se mantienen. Por otro lado, si el diagrama es irregular indicaría que algún factor por detrás está afectando a la relación con la variable predictora y, por lo tanto, la variable puede ser significativa para un modelo logístico.

Los siguientes diagramas de mosaico visualizan la relación entre dos variables, figura 7a muestra independencia cuando el diagrama es regular, figura 7b el diagrama es irregular que demuestra que hay una relación.

Figura 9

Diagramas de mosaico para mostrar relacionamiento, variable no significativa (a) y variable significativa (b)

Creación del autor: Alfonso Prado



Interpretación de los Coeficientes en Regresión Logística

Un tópico que llama la atención en regresiones logísticas es interpretar los coeficientes del modelo de regresión debido a que están en la escala log-odds. Debemos tener cuidado de convertirlos antes de interpretar los términos de las variables originales.

Recordando la definición de logit.

Figura 10

Fórmula de logit

Tomado de Agresti (2007)

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

Entonces la función exp(β0) representa las probabilidades de que la característica de éxito esté presente para un individuo x=0, es decir en la línea de base. Si hay múltiples predictores involucrados, todos deberían establecerse en 0 para esta interpretación.

La función exp(βi) representa el aumento multiplicativo en las probabilidades de éxito por cada aumento de 1 unidad en x. Si βi > 0 indica que es una relación positiva, si es negativo entonces el incremento de X decrementa la posibilidad de que se presente el evento.

Por ejemplo, de acuerdo con la siguiente figura, el coeficiente de la variable "MaritalStatus_Divorced" es -1.109 , y su exp(-1.109) es 0.32 , esto debe interpretarse de la siguiente forma: Por cada unidad que aumenta la variable MaritalStatus_divorced el **odds** de que se presente el evento (no el OR) disminuye en 0.32.

Figura 11

Ejemplo regresión logística del laboratorio

Creación del autor Alfonso Prado

	Estimate	Std. Err	z value	Pr(>|z|)
(Intercept)	-0.985727	0.249089	-3.957	7.58E-05
MaritalStatus_Divorced	-1.109196	0.129183	-8.586	2.00E-16
MaritalStatus_Married	-0.875288	0.095759	-9.141	2.00E-16
YearsAtCompany	-0.075929	0.009404	-8.074	6.80E-16
BusinessTravelTravel_Frequently	1.390731	0.199132	6.984	2.87E-12
BusinessTravelTravel_Rarely	0.665648	0.186944	3.561	0.00037
DepartmentHuman_Resources	0.754718	0.177866	4.243	2.20E-05
JobSatisfaction_2_3	0.534664	0.108733	4.917	8.78E-07
JobSatisfaction_1	0.967785	0.124902	7.748	9.31E-15
WorkLifeBalance2	-0.84261	0.173749	-4.85	1.24E-06
WorkLifeBalance3	-1.106309	0.160361	-6.899	5.24E-12
WorkLifeBalance4	-0.758334	0.197157	-3.846	0.00012

La regresión logística, al igual que la regresión ordinaria, puede tener múltiples variables explicativas.

Algunos o todos estos predictores pueden ser categóricos, en lugar de cuantitativos. Esta sección muestra cómo incluir predictores categóricos, comúnmente llamados factores.

Supongamos que una respuesta binaria Y tiene dos predictores binarios, X y Z. Los datos se representarían en una tabla de contingencia de 2 × 2 × 2.

Si x y z tomaran valores 0 y 1 para representar las dos categorías de cada variable explicativa, el modelo para

$$P(Y = 1) = \text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$$

Para esta codificación, la siguiente tabla muestra los valores logit en las cuatro combinaciones de valores de los dos predictores.

Tabla 3

Modelos logit para 2 variables predictores binarias

Creación del autor Alfonso Prado

x	z	Logit
0	0	α
1	0	$\alpha + \beta_1$
0	1	$\alpha + \beta_2$
1	1	$\alpha + \beta_1 + \beta_2$

Codificación por One Hot para variables categóricas nominales multinivel

Esta codificación, se implementa mediante variables "dummy" usando el paquete R caret. En primer lugar, a partir de una variable categórica X (que debe ser del tipo factor) con p niveles, se crean p variables nuevas y cada una de estas nuevas se asocia de forma individual a un elemento del conjunto Xi.

En segundo lugar, todas las variables nuevas toman el valor de 0, excepto aquella que representa el valor original del registro sin modificar que tomará el valor 1 como se mostraba en los laboratorios de regresión.

Vector de codificación	Categorías				
	Azul	Rojo	Verde	Blanco	Negro
$_{Azul}\Phi_{Color}$	1	0	0	0	0
$_{Rojo}\Phi_{Color}$	0	1	0	0	0
$_{Verde}\Phi_{Color}$	0	0	1	0	0
$_{Blanco}\Phi_{Color}$	0	0	0	1	0
$_{Negro}\Phi_{Color}$	0	0	0	0	1

Figura 12

Codificación one-hot para variable categórica color

Fuente Rocha A.(2020)

Mejores prácticas en la creación de modelos logísticos

En la selección de variables elimine primero variables que no demuestran correlación o aquellas numéricas que tengan varianza 0 o cercana a 0.

Si existen variables con NAs proceda con las imputaciones que considere necesarias, sin embargo, tome en cuenta que las imputaciones podrían cambiar completamente la forma de los datos

Si utiliza la transformaciones log o BoxCox no centre los datos ni ninguna operación que pueda generar valores negativos, esto le llenará de NAs el dataset. En este caso deberá utilizar YeoJohnson

Aplique one-hot primero. One-Hot da como resultado que los datos estén más dispersos, lo que muchos algoritmos pueden usar de manera eficiente. Si estandariza los datos primero, creará datos densos con los cuales los algoritmos se ejecutarán de manera menos eficiente.

Lo mismo ocurre con la agrupación de categorías, primero haga los agregados y luego codifique con one-hot.

Referencias citadas en la Clase 13

Agresti, A., (2007), An introduction to categorical data analysis, John Wiley and Sons publication

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional Publishing Schmuller J.(2022), Statistical Analysis with Excel For Dummies, Wiley

Teck K, (2022) , Practical t-test Power Analysis with R, Practical Assessment, Research & Evaluation , Volumen 27, No 18

Berman H.G., "*F Distribution* , descargado de https://stattrek.com/probability-distributions/f-distribution

Rodrigo J. (2016), *Análisis de varianza para comparar múltiples medias*, https://cienciadedatos.net/documentos/19_anova

Szretter, M. (2017), Apunte de Regresión Lineal, https://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

Rindang B. y et al ,(2019) A comparison of some link functions for binomial regression models. https://doi.org/10.1063/1.5139815

PUCE Virtual

La excelencia no se improvisa

síguenos

CV_ICYS_M