UNIVERSIDAD DE JAÉN

Material del curso "Recursos metodológicos y estadísticos para la docencia e investigación"

Manuel Miguel Ramos Álvarez

MÓDULO XI "EXPLICACIÓN DE DATOS CATEGÓRICOS"

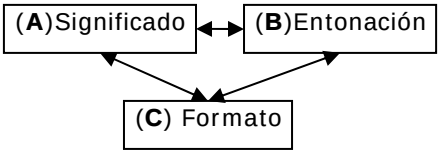
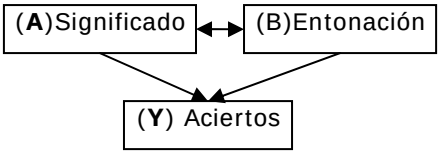
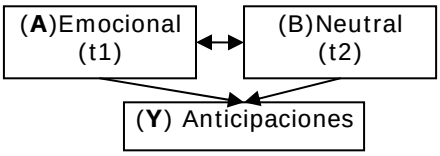
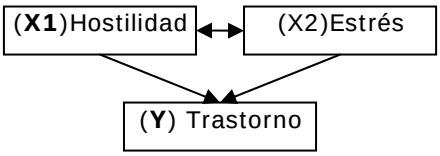
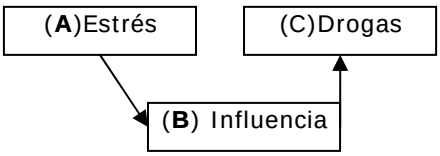
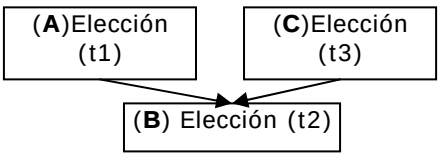
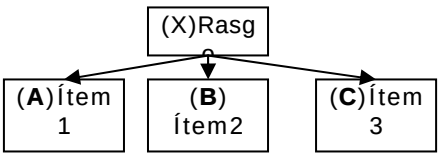
Índice

11. Acercamiento basado en datos categóricos.

- Introducción general para diferenciar las variantes principales.
- Aproximación basada en la regresión logística para investigaciones explicativas.

11.1. La lógica del análisis de datos categóricos

Clasificación de las diferentes variantes:

VARIABLES	EJEMPLO	MODELO	DISEÑO DE APLICACIÓN
Categóricas todas. Sin diferenciar estatus variables.		Log-lineal	Descriptivo
Categóricas todas. Unas son var.ind. y otra es var.dep. (mm. independientes)		Logit Probit	Explicativo(i.e. ≈Experimental)
Categóricas todas. Unas son var.ind. y otra es var.dep. (mm. relacionadas)		Logit-GSK	Explicativo(i.e. ≈Experimental medidas repetidas)
No categóricos los Predictores y Categórico el criterio		Regresión Logística	Explicativa(i.e. Correlacional)
Cadena causal de variables categóricas		Logit-causal	Explicativo(i.e. ≈Cuasiexperi-mental)
Categóricas a través del tiempo (t1, t2,...).		Logit-Markov	Descriptivo(i.e. longitudinal, observacional)
Categóricas en diversas var.criterio		Logit-Latente	Descriptivo (i.e. ≈Análisis Factorial en Tests)

11.2. Análisis de Regresión Logística para investigaciones de tipo explicativo

11.2.1. Regresión logística: regresión con una variable dependiente no métrica

- Sirve para la predicción de variables no métricas, donde no es adecuada regresión múltiple y donde el análisis discriminante es muy exigente en cuanto a los supuestos que impone.
- El objetivo es predecir la pertenencia a una categoría (o grupo) a partir de variables predictoras que pueden ser de tipo no métrico o no.
- Como en represión lineal, la significación de los parámetros asociados a los diferentes predictores nos permitirá construir un modelo estadístico basado únicamente en los predictores que aportan significación estadística.
- Permite incluir la interacción, a diferencia de la técnica de análisis discriminante.
- También sirve para clasificar a individuos en categorías, como en el tipo discriminante. Para ello se fija un punto de corte en la probabilidad (por ejemplo, 0.50) y se asignan a una categoría aquellos cuya probabilidad sea superior y a la otra categoría a los que estén por debajo.

11.2.2. Bases. Las ecuaciones básicas de la regresión logística

- Para predecir una variable dicotómica, que adopta valores 0 o 1, su relación con los predictores es no lineal. Lo que se predice no es directamente la variable sino la **probabilidad de que la variable adopte** un cierto valor, por ejemplo, la probabilidad de que se produzca éxito escolar (p).
- Para predecir una probabilidad pueden utilizarse diferentes funciones, entre las que destaca la logística.

$p = \frac{e^u}{1+e^u}$; donde el modelo lineal aparece realmente en el exponente

$$u = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Supuesto M4
Analizar -> Regresión -> Logística binaria -> Dependiente: RE -> Covariables: CE, TC, Horas -> Covariables categóricas: CE, TC -> Continuar -> Método: Introducir -> Aceptar (análisis)

Statistics→

- expresión es conocida como *logit*, o logaritmo de las verosimilitudes. Ya que alternativamente se puede expresar como:

$$\ln\left(\frac{p}{1-p}\right) = u = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

- Que viene expresado en función del cociente de probabilidades: "*odds ratio*"

$$\frac{p}{1-p} = e^u$$

- El procedimiento de estimación que se emplea es el de máxima verosimilitud, un método de carácter iterativo que tiende a proporcionar la solución tras varios pasos.
- Es importante precisar que las variables independientes son llamadas covariables por SPSS.
- La interpretación de los parámetros depende del sistema de codificación empleado. El más usual es el tipo "dummy". Crear tantas variables dummy como categorías tenga la variable menos 1. En cada una de las variables ficticias se codifica una categoría (asignándole 1 a los sujetos que la poseen y 0 a los que no), y es en el conjunto donde quedan codificadas todas.

- Para evaluar la contribución de las variables predictoras hay varias alternativas:
- Mediante la prueba de Wald: cociente entre el coeficiente y su error típico.
- Mediante la bondad de ajuste basada en el logaritmo de la verosimilitud (log-likelihood) comparando modelos que difieren en uno parámetro cada vez:

$$\log - likelihood = \sum_{i=1}^{N}\left[Y_i \ln(p_i) + (1-Y_i)\ln(1-p_i)\right]$$

$$\chi^2 = 2\left[(\log - likelihood(MOD1)) - (\log - likelihood(MOD2))\right]$$

Ejemplo de codificación:

CE	TC	CE(1)	CE(2)	TC(1)	TC(2)
1	1	1	0	1	0
2	2	0	1	0	1
2	3	0	1	0	0
3	3	0	0	0	0
...	---	---	---	---	---
1	2	1	0	0	1
1	2	1	0	0	1
2	2	0	1	0	1
1	1	1	0	1	0

11.2.3. Fase de resumen del modelo

- En este contexto, los estadísticos de Cox-Snell y de Nagelkerke sirven para la estimación del efecto, tipo RPE ajustada (ver detalles en Catena, Ramos y Trujillo, 2003).
- Respecto al análisis de los residuos se utilizan métodos análogos a los de regresión, usualmente basados en los residuales tipificados o estandarizados (i.e. 2 desviaciones) y la distancia de Cook (valor de 1).

Supuesto M4
Analizar -> Regresión -> Logística binaria ->...
Guardar -> Residuos: Tipificados -> Continuar ->

...-> Guardar -> Influencia: De Cook -> Continuar ->

Statistics→

11.2.4. Variantes

A) Regresión logística simultánea

- Los predictores son introducidos todos a la vez en la ecuación de regresión. La significación de los mismos se evalúa por turnos, uno cada vez sobre la base de lo que no comparte con los otros.
- La exclusión del modelo vendrá dada cuando su tolerancia sea demasiado baja.
- Siempre que no se tenga claro qué variables pueden ser más importantes

Supuesto M4
Analizar -> Regresión -> Logística binaria -> Dependiente: RE -> Covariables: Horas –> Siguiente (Bloque) -> Covariables: CE -> Siguiente (Bloque) -> Covariables: TC -> Categórica -> Covariables categóricas: CE, TC -> Contraste: Indicador -> Continuar -> Aceptar (análisis)

Statistics→

B) Regresión logística secuencial

- Se introduce o elimina un predictor en cada paso. Hay dos subtipos que se diferencian por las razones por las cuales son incluidos/excluidos de la ecuación.
    - **Serial -jerárquica-.** Son introducidos según un criterio teórico y en cada paso puede introducirse más de un predictor, pero una vez que el predictor está dentro de la ecuación no puede ser eliminado de la misma.
    - **Por etapas –stepwise-.** Cada predictor es incluido o eliminado de la ecuación según cumpla los criterios estadísticos de inclusión/exclusión prefijados por el investigador. En ausencia de hipótesis específicas sobre la importancia de los predictores. El problema más importante es que puede excluirse alguna variable que tenga una alta relación con la dependiente debido a su correlación con otras predictoras –multicolinealidad-. La variante concreta depende de la dirección y del algoritmo estadístico.

C) La clasificación de individuos

- Con el modelo final, es importante determinar si el modelo permite asignar los individuos a las categorías de forma significativa.

$$p \geq 0,5 \Rightarrow s \in G = 1$$

$$p < 0,5 \Rightarrow s \in G = 0$$

- Por defecto suele emplearse la probabilidad 0.5 como criterio, de modo que:
- Realizar una prueba de bondad de ajuste sobre las frecuencias así obtenidas.
- Sería conveniente que el punto de corte seleccionado trate de equilibrar la relación entre errores de clasificación de un tipo o de otro, hacia puntos intermedios normalmente.

Supuesto M4
Analizar -> Regresión -> Logística binaria ... -> Opciones -> Punto de corte para la clasificación: 0.6 -> Continuar ...

-> Guardar -> Valores pronosticados: Grupo de pertenencia -> Continuar

Statistics→

D) El análisis de regresión politómico o multinomial

Analizar -> Regresión -> Logística multinomial
Statistics→

- Cuando la variable de agrupamiento tiene más de dos niveles, se obtienen tantas ecuaciones de predicción como grados de libertad tiene la variable de agrupamiento.
- Hay dos posibilidades:

La variable de agrupamiento es nominal, cada ecuación predice la probabilidad de que el individuo sea miembro de un grupo diferente. Esto es,

$$P(Y = j) = \frac{e^u}{1+e^u}$$, como en el caso básico.

Si La variable es ordinal (i.e. rendimiento escolar: bajo, medio, alto), las ecuaciones predicen la probabilidad de que el sujeto se sitúe en el grupo superior al índice de la ecuación:

$$P(Y > j) = \frac{e^u}{1+e^u}$$

11.2.5. 5. Limitaciones y supuestos del análisis de regresión logística

Linealidad de la función logit

- Se puede probar el supuesto incluyendo la interacción en el modelo, de manera que si ésta es significativa entonces incumplimos el supuesto. En este caso bastará con mantener dicha interacción en el modelo.

Independencia de los errores

Multicolinealidad. Expresa el grado de interrelación entre los predictores y lo que la técnica de regresión asume es que ésta es de baja magnitud. Su incumplimiento tiene graves consecuencias.

- Hay dos alternativas cuando la multicolinealidad es alta:
    - Regresión sesgada (“ridge regresión”), intenta estabilizar los parámetros manipulando las varianzas.
    - Regresión por componentes principales, que se basa en la alta correlación entre predictores para definir variados que son combinaciones lineales de los predictores y emplear los variados como nuevos predictores del criterio.

Número de variables y número de sujetos

- No es recomendable con bajo número de participantes ya que se la estimación no se hace adecuadamente y además se distorsiona la interpretación.

Puntos extremos

- La presencia de puntos extremos puede traducirse en una baja capacidad predictiva del modelo.

11.3. Realización de los supuestos de prácticas

11.3.1. Ejemplificación del análisis tipo Regresión Logística mediante el Supuesto 3

Analizar -> Regresión -> Logística binaria -> Dependiente: RE -> Covariables: CE, TC, Horas -> Covariables categóricas: CE, TC -> Continuar -> Método: Introducir -> Aceptar (análisis)

...-> Guardar -> Residuos: Tipificados -> Continuar -> ...

...-> Guardar -> Influencia: De Cook -> Continuar -> ...

Analizar -> Regresión -> Logística binaria -> Dependiente: RE -> Covariables: Horas –> Siguiente (Bloque) -> Covariables: CE -> Siguiente (Bloque) -> Covariables: TC -> Categórica -> Covariables categóricas: CE, TC -> Contraste: Indicador -> Continuar -> Aceptar (análisis)

En SPSS podemos seleccionar la probabilidad de corte (por ejemplo, 0.6) con la siguiente secuencia de comandos:

...-> Opciones -> Punto de corte para la clasificación: 0.6 -> Continuar ...

grupos predichos en una nueva variable. La secuencia de instrucciones es la siguiente

... -> Guardar -> Valores pronosticados: Grupo de pertenencia -> Continuar ->

cuyo resultado es la creación de una nueva variable, pgr_1, que podremos utilizar

Volver Principio