Análisis de Regresión con información categórica

http://www.academia.utp.ac.pa/humberto-alvarez/diseno-de-experimentos-y-regresion

Información cualitativa en los modelos de regresión

- Hasta ahora las variables regresoras que se han utilizado para explicar la variación de una variable tenían un carácter cuantitativo.
- Sin embargo, hay otras variables de carácter cualitativo que pueden ser importantes para explicar el comportamiento de la variable endógena, como el sexo, raza, religión, nacionalidad, región geográfica, etc.
- Las variables cualitativas pueden ser dicotómicas o binarias y variables de categorías.
- Las variables binarias o dicotómicas tienen valores de 1 o 0 dependiendo si cierta característica está presente o no.
- En el caso de varias categorías, se definen variables que, como se mencionó anteriormente, pueden ser sexo, raza, etc.

Modelos con variables dicotómicas

- Los factores cualitativos se pueden presentar en forma de información binaria.
- Cuando los factores cualitativos se presentan en forma dicotómica la información relevante puede mostrarse como una variable binaria o una variable de cero-uno.
- Las variables binarias que se utilizan como regresores son comúnmente llamadas variables ficticias.
- En la definición de una variable dicotómica, se debe decidir a qué caso se le asigna el valor 1 y a cual se le asigna el valor 0.

Representando las variables binarias

- En este caso es necesario transformar una variable categórica *X* con *k* categorías en *k-1* variables binarias (con valor 0 ó 1).
- Una variable binaria de este tipo es conocida con el nombre de variable auxiliar, ficticia o variable dummy.
- El conjunto de *k-1* variables auxiliares es utilizado para identificar cada una de las categorías de la variable original *X*.
- Por ejemplo, si *X* tiene *k*=3 categorías, dos variables auxiliares D1 y D2 serán suficientes para representar cada una de las categorías de *X*.
- Por ejemplo, la combinación D1=1 y D2=0
- puede identificar la primera categoría, D1=0 y D2=1 la segunda categoría y D1=0 y D2=0 la tercera. A esta última se la suele llamar categoría de referencia.

Ejemplo

- La tabla del archivo Polímero, presenta los valores de turbidez del medio (Y) y su pH para tres tipos de polímeros A, B, C.
- El interés se centra en la dependencia de la turbidez del medio respecto al pH del mismo.
- Se quiere probar el siguiente modelo de regresión:

$$Y = \beta_0+\beta_1 pH+\beta_2 D1+\beta_3 D2+\beta_4 D1*pH+\beta_5 D2*pH+\varepsilon$$

Y	pH	Polímero
292	6.5	A
329	6.9	A
352	7.8	A
378	8.4	A
392	8.8	A
410	9.2	A
198	6.7	B
227	6.9	B
277	7.5	B
297	7.9	B
364	8.7	B
375	9.2	B
167	6.5	C
225	7	C
247	7.2	C
268	7.6	C
288	8.7	C
342	9.2	C

Caso	Y	pH	Polímero
1	292.00	6.50	A
2	329.00	6.90	A
3	352.00	7.80	A
4	378.00	8.40	A
5	392.00	8.80	A
6	410.00	9.20	A
7	198.00	6.70	B
8	227.00	6.90	B
9	277.00	7.50	B
10	297.00	7.90	B
11	364.00	8.70	B
12	375.00	9.20	B
13	167.00	6.50	C
14	225.00	7.00	C
15	247.00	7.20	C
16	268.00	7.60	C
17	288.00	8.70	C
18	342.00	9.20	C

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	18	0.97	0.96	556.03	154.26	160.50

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-158.27	48.52	-263.98	-52.57	-3.26	0.0068		
pH	53.82	6.25	40.20	67.45	8.61	<0.0001	78.08	2.98
Polímero_A	197.69	68.79	47.80	347.58	2.87	0.0140	12.26	88.99
Polímero_B	-108.74	71.05	-263.55	46.07	-1.53	0.1518	6.34	94.93
Polímero_A_pH	-13.56	8.74	-32.60	5.48	-1.55	0.1466	6.41	92.39
Polímero_B_pH	17.39	9.09	-2.41	37.20	1.91	0.0798	7.66	96.82

Polimero

Caso	Y	pH	Polímero	Polímero_A	Polímero_B	Polímero_A_pH	Polímero_B_pH
1	292.00	6.50	A	1	0	6.50	0.00
2	329.00	6.90	A	1	0	6.90	0.00
3	352.00	7.80	A	1	0	7.80	0.00
4	378.00	8.40	A	1	0	8.40	0.00
5	392.00	8.80	A	1	0	8.80	0.00
6	410.00	9.20	A	1	0	9.20	0.00
7	198.00	6.70	B	0	1	0.00	6.70
8	227.00	6.90	B	0	1	0.00	6.90
9	277.00	7.50	B	0	1	0.00	7.50
10	297.00	7.90	B	0	1	0.00	7.90
11	364.00	8.70	B	0	1	0.00	8.70
12	375.00	9.20	B	0	1	0.00	9.20
13	167.00	6.50	C	0	0	0.00	0.00
14	225.00	7.00	C	0	0	0.00	0.00
15	247.00	7.20	C	0	0	0.00	0.00
16	268.00	7.60	C	0	0	0.00	0.00
17	288.00	8.70	C	0	0	0.00	0.00
18	342.00	9.20	C	0	0	0.00	0.00

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	82707.78	5	16541.56	77.76	<0.0001
pH	15759.53	1	15759.53	74.08	<0.0001
Polímero_A	1756.64	1	1756.64	8.26	0.0140
Polímero_B	498.26	1	498.26	2.34	0.1518
Polímero_A_pH	512.47	1	512.47	2.41	0.1466
Polímero_B_pH	778.95	1	778.95	3.66	0.0798
Error	2552.67	12	212.72		
Total	85260.44	17			

Regresión logística

- Permite modelar la relación entre una variable respuesta de naturaleza *dicotómica* en relación a una o más variables independientes o regresoras.
- Puede ser usado para predecir la probabilidad (pi) de que la variable respuesta asuma un valor determinado, por ejemplo, probabilidad de éxito (y=1) en una variable dicotómica que asume los valores 0 y 1.
- Para una respuesta binaria, el modelo de regresión logística simple, es decir con una regresora, tiene la siguiente forma:

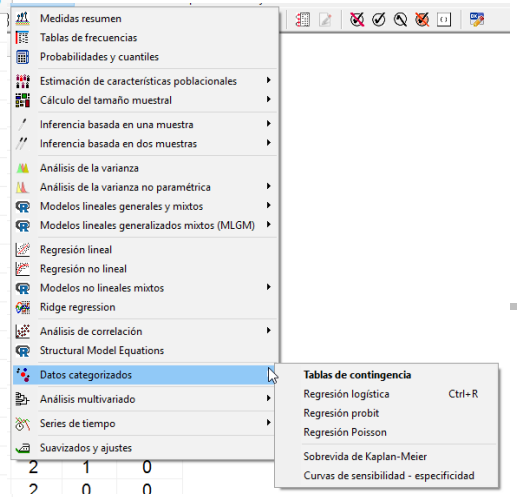
$$Logit(p_i) = \log(p_i/(1-p_i)) = \alpha + \beta X_i$$

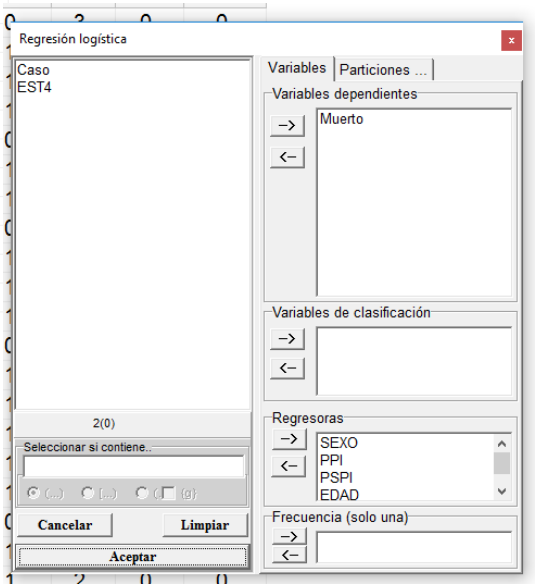
- Donde p_i es la probabilidad de éxito dado X_i, α es la ordenada al origen (constante), β es la pendiente o coeficiente de regresión asociado a X y X es la variable regresoras.
- Se modela la transformación *Logit* de la probabilidad de éxito como una función lineal de una o más variables regresoras.

Ejemplo:

- Se estudia el efecto de la edad, la pérdida de peso inicial, el % del peso normal (PPI), el sexo (1:Masculinos, 0:Femeninos) y una medida del estado general del paciente (PSPI) sobre la sobrevida en pacientes con cáncer de pulmón evaluada a los tres meses de iniciado el tratamiento.
- Se propone realizar el ajuste de un modelo de regresión logística múltiple de la variable "muerto" (vale 0 si el paciente está vivo y 1 si el paciente falleció) en relación a la edad, sexo, PPI y PSPI.

Caso	EDAD	PPI	SEXO	PSPI	EST4	Muerto
1	68	19	0	3	0	0
2	69	9	1	2	0	0
3	45	0	1	1	0	0
4	63	0	1	1	0	0
5	39	0	0	2	1	0
6	70	0	1	2	0	0
7	53	5	1	2	0	0
8	53	20	0	1	0	0
9	53	25	1	3	0	1
10	44	6	1	1	1	0
11	59	22	1	3	1	0
12	50	0	0	1	0	0
13	63	27	1	2	0	0
14	58	15	1	3	0	1
15	55	3	1	2	0	0
16	57	0	1	1	0	0
17	65	28	1	3	0	0
18	65	21	0	2	1	0
19	46	20	1	2	0	0
20	53	16	1	2	0	0
21	54	10	1	3	0	1
22	68	22	1	3	1	1
23	70	12	1	2	0	0
24	62	11	1	3	1	0
25	70	10	1	2	0	0
26	56	0	1	1	1	0
27	53	10	1	3	1	0
28	66	12	1	3	1	0
29	52	17	1	2	0	0





Regresión logística

Distribución: Binomial
Función de enlace: Logit

Variable dependiente: Muerto
Codificar como éxito a valores mayores que la media
Número de observaciones: 150
Observaciones faltantes: 0
Iteraciones: 8 (max=20)
Tolerancia: 1E-9 (0.000000000)

Parámetros	Est.	E.E.	O.R.	Wald LI(95%)	Wald LS(95%)	Wald Chi²	p-valor
Constante	-7.60	2.72	5.0E-04	2.4E-06	0.10	7.83	0.0051
SEXO	-0.37	0.86	0.69	0.13	3.74	0.18	0.6691
PPI	0.12	0.05	1.12	1.02	1.24	5.45	0.0195
PSPI	0.88	0.52	2.42	0.88	6.68	2.91	0.0883
EDAD	0.03	0.04	1.03	0.96	1.11	0.69	0.4049

	Valor	gl
Log Likelihood	-36.58	145
Deviance	73.16	145
Escala (fijada)	1.00	

Pruebas de hipótesis marginales

F.V.	gl	-2[L0-L1]	p-valor
SEXO	1	0.17	0.6772
PPI	1	5.48	0.0193
PSPI	1	3.14	0.0762
EDAD	1	0.74	0.3906

Conclusiones

- Para aplicar el modelo de regresión lineal han de respetarse los supuestos del modelo.
- Estos supuestos son: linealidad, normalidad y homocedasticidad.
- En el caso de una variable X dicotómica, la regresión simple equivale a un contraste de medias.
- El supuesto de normalidad en las distribuciones ligadas es equivalente al supuesto de normalidad en las poblaciones orígenes de las dos muestras en el contraste de medias.
- El supuesto de homocedasticidad es el equivalente al de igualdad de varianza en las poblaciones orígenes.
- Por último, el de linealidad, se cumple por cuanto entre dos puntos (las medias de ambas muestras) siempre se puede definir una recta.