

Problema 1

La siguiente tabla muestra los resultados parciales de dos encuestas que forman parte de un estudio para evaluar el desempeño del Primer Ministro del Canadá. Se tomó una muestra aleatoria de 1600 ciudadanos canadienses mayores de edad y en los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después de la primera encuesta, sólo 880 aprueban su desempeño:

Primera encuesta	Segunda encuesta		Total
	$Y = 1$, Aprueba	$Y = 0$, Desaprueba	
$x = 1$, Aprueba	794	150	944
$x = 0$, Desaprueba	86	570	656
Total	880	720	1600



CIMAT



Unidad Monterrey

Centro de Investigación en Matemáticas, A.C.

Problema 1

a) Considere el modelo de regresión logística

$$\log \frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} = \beta_0 + \beta_1 x_i$$

Escriba la logverosimilitud correspondiente. Muestre explícitamente (i.e. maximizando la logverosimilitud), que el estimador máximo verosimilitud para β_1 es el logaritmo de la tasa de momios de la tabla dada (En general, en regresión logística los estimadores de máxima verosimilitud no tienen una forma explícita, sin embargo, en el presente caso si).

b) Sea p_1 la proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial y sea p_2 la proporción correspondiente seis meses después. Considere la hipótesis $H_0: p_1 = p_2$, ¿Cómo puede hacerse esta prueba?



CIMAT



Unidad Monterrey

Centro de Investigación en Matemáticas, A.C.

Problema 2

Se tiene la siguiente tabla donde se eligen varios niveles de ronquidos y se ponen en relación con una enfermedad cardíaca. Se toman como puntuaciones relativas de ronquidos los valores $\{0, 2, 4, 5\}$.

Ronquido	Enfermedad Cardíaca		
	SI	NO	Proporción de SI
<i>Nunca</i>	24	1355	0.017
<i>Ocasional</i>	35	603	0.055
<i>Casi cada noche</i>	21	192	0.099
<i>Cada noche</i>	30	224	0.118


Ajuste un modelo lineal generalizado **logit y probit (investigar sobre el link probit)** para analizar si existe una relación entre los ronquidos y la posibilidad de tener una enfermedad cardíaca.



Problema 3

Entre los cangrejos cacerola se sabe que cada hembra tiene un macho en su nido, pero puede tener más machos concubinos.

Se considera que la variable respuesta es el número de concubinos y las variables explicativas son: color, estado de la espina central, peso y anchura del caparazón.

 Datos_Crabs.txt: Bloc de notas

Archivo	Edición	Formato	Ver	Ayuda
Color	Spine	Width	Satellite	Weight
3	3	28.3	8	3050
4	3	22.5	0	1550
2	1	26.0	9	2300
4	3	24.8	0	2100
4	3	26.0	4	2600
3	3	23.8	0	2100
2	1	26.5	0	2350

Realizar e interpretar los resultados de ajustar un modelo lineal generalizado tipo poisson.

Problema 4

Suponga $(x_1, y_1), \dots, (x_n, y_n)$ observaciones independientes de variables aleatorias definidas como sigue:

$$Y_i \sim \text{Bernoulli}(p), \quad i = 1, \dots, n$$

$$X_i \mid \{Y_i = 1\} \sim N(\mu_1, \sigma^2)$$

$$X_i \mid \{Y_i = 0\} \sim N(\mu_0, \sigma^2)$$

Usando el Teorema de Bayes, muestre que $P(Y_i = 1|X_i)$ satisface el modelo de regresión logística, esto es

$$\text{logit}(P(Y_i = 1|X_i)) = \alpha + \beta X_i$$

con $\beta = (\mu_1 - \mu_0)/\sigma^2$.



Problema 5

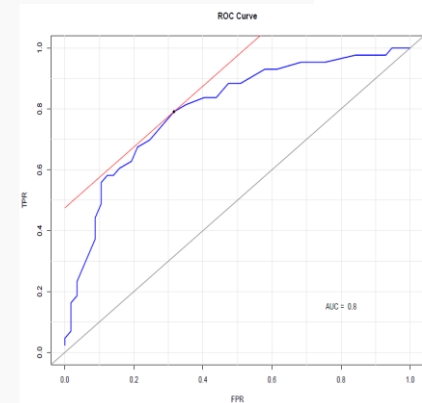
Cuando usamos un modelo de regresión logística para clasificación, tenemos que definir el umbral, p , a partir del cual declaramos un “positivo”.

Las curvas ROC grafican las tasas TPR vs FPR para diferentes umbrales p .

$$TPR = \text{True Positive Rate} = \frac{TP}{P} = \text{“sensitividad”}$$

$$FPR = \text{False Positive Rate} = \frac{FP}{N} = 1 - \text{“especificidad”}$$

	Observados	
	1	0
Decisión = 1	TP	FP
Decisión = 0	FN	TN
	P	N



Problema 5

- La gráfica de TPR vs FPR puede interpretarse como una gráfica de “poder” vs “error tipo I”.
- Idealmente, una regla de decisión estaría en el punto $(0, 1)$
- El área bajo la curva, AUC, puede verse, es la probabilidad de que un individuo de los positivos, tomado al azar, tenga un riesgo estimado mayor que un individuo de los negativos, tomado al azar.
- El estadístico J de Youden, es una medida que, con un sólo número, trata de capturar el desempeño de una prueba de diagnóstico. Es la máxima distancia vertical, entre la diagonal y la curva ROC, o equivalentemente

$$J = \text{sensitividad} - (1 - \text{especificidad})$$



Problema 5

- ❑ Construyan la curva ROC para el problema de daño coronario y su relación con la edad visto en la clase 3 del curso.

```
# Hosmer, D.W. & Lemeshow, S.(1989) Applied logistic regression. Wiley
# Edad y Coronaria (daño significativo en coronaria)

edad <- c(
  20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32, 32, 33, 33,
  34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39, 40, 40, 41,
  41, 42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48,
  48, 48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57,
  57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64, 65, 69)

coro <- c(
  0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,1,0,
  0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,0,1,1,0,0,1,0,1,0,0,1,1,1,1,0,1,1,1,1,1,0,
  0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1)
```



CIMAT



Unidad Monterrey

Centro de Investigación en Matemáticas, A.C.

Problema 6

La siguiente tabla muestra conteos de células T_4 por mm^3 en muestras de sangre de 20 pacientes (en remisión) con enfermedad de Hodgkin, así como conteos en 20 pacientes en remisión de otras enfermedades. Una cuestión de interés es si existen diferencias en las distribuciones de conteos en ambos grupos.

H	396	568	1212	171	554	1104	257	435	295	397
No-H	375	375	752	208	151	116	736	192	315	1252
H	288	1004	431	795	1621	1378	902	958	1283	2415
No-H	675	700	440	771	688	426	410	979	377	503

- Haga una comparación gráfica exploratoria de estos datos.
- Ajuste un modelo de Poisson apropiado.
- Usando la normalidad asintótica de los estimadores de máxima verosimilitud, dé un intervalo del 90% de confianza para la diferencia en medias. ¿Hay evidencia de diferencias en los dos grupos en cuanto a las medias de los conteos?



Problema 7

Los datos de la tabla en la siguiente hoja son números, n , de pólizas de seguros y los correspondientes números, y , de reclamos (esto es, número de accidentes en los que se pidió el amparo de la póliza). La variable CAR es una codificación de varias clases de carros, EDAD es la edad del titular de la póliza y DIST es el distrito donde vive el titular.

- Calcule la tasa de reclamos, y/n , para cada categoría y grafique estas tasas contra las diferentes variables para tener una idea de los efectos principales.
- Use regresión logística para estimar los efectos principales (cada variable tratada como categórica y modelada usando variables indicadoras) así como sus interacciones.
- Basados en los resultados del inciso anterior, los autores del artículo donde aparecieron estos datos, decidieron que ninguna interacción era importante y que podían considerar que CAR y EDAD fuesen tratadas como variables continuas. Ajuste un modelo incorporando estas observaciones y compárelo con el obtenido en (b). ¿Cuáles son las conclusiones?.

CAR	EDAD	DIST= 0		DIST= 1	
		y	n	y	n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114



Problema 8

A lo largo del curso hemos enfatizado el uso del método de Máxima Verosimilitud para todo lo relacionado con estimación. Consideremos ahora una alternativa: El método de la **Mínima Ji-Cuadrada**. Suponga que las celdas de una multinomial están parametrizadas en términos de un vector $\theta = (\theta_1, \dots, \theta_s)^T$. El método de la mínima ji-cuadrada (ver Agresti, pág. 611) consiste en estimar θ mediante aquel valor que minimice el estadístico de Pearson

$$\chi^2 = \sum \frac{(\text{obs} - \text{esp})^2}{\text{esp}} = \sum_{j=1}^K \frac{(y_j - n\pi_j(\theta))^2}{n\pi_j(\theta)}$$

Considere el siguiente problema. Suponga una población muy grande de objetos que pueden clasificarse en tres categorías, A , B y C . Para estimar las proporciones π_1 , π_2 y π_3 correspondientes a cada una de esas categorías, se efectuó un estudio; se obtuvieron tres muestras de tamaños n_1 , n_2 y n_3 tomadas de la población global, sin embargo, en vez de registrar la frecuencia observada de A 's, B 's y C 's de cada muestra, lo que se hizo fue anotar:

Número de A 's en la muestra de tamaño $n_1 = y_1$

Número de B 's en la muestra de tamaño $n_2 = y_2$

Número de C 's en la muestra de tamaño $n_3 = y_3$

Estime π_1 , π_2 y π_3 usando el método de la mínima ji-cuadrada; suponga que $n_1 = 100, y_1 = 22$, $n_2 = 150, y_2 = 52$, $n_3 = 200, y_3 = 77$. Esto es, encuentre π_1 , π_2 y π_3 que minimizen

$$\frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1} + \frac{[(n_1 - y_1) - n_1(1 - \pi_1)]^2}{n_1(1 - \pi_1)} + \dots + \frac{(y_3 - n_3\pi_3)^2}{n_3\pi_3} + \frac{[(n_3 - y_3) - n_3(1 - \pi_3)]^2}{n_3(1 - \pi_3)}$$

con la restricción $\pi_3 = 1 - \pi_1 - \pi_2$ (sugerimos usar directamente `nlminb` de R).



Problema 9

Se toman los datos relacionados con el hundimiento del Titanic en abril de 1912. El resultado se puede expresar en una tabla de dimensión 4.

Las variables son **Class** de los pasajeros (1, 2, 3, Tripulación), **Sex** de los pasajero (Male, Female), **Age** de los pasajeros (Child, Adult), y **Survived** si los pasajeros sobrevivieron o no (No, Yes). Usar librería en R “titanic” y los datos se encuentran en la variable “Titanic”.

Considerar entonces un modelo log-lineal para analizar los posibles efectos:

- Class: Hay más pasajeros en algunas clases que en otras.
- Sex: Hay más pasajeros en un sexo que en otro.
- Age: Hay más pasajeros en un grupo de edad que en otro.
- Survived: Hay más pasajeros o vivos o muertos que la alternativa.
- Class \times Sex: Class y Sex no son independientes.
- Class \times Age: Class y Age no son independientes.
- Class \times Survived: Class y Survived no son independientes.
- Sex \times Age: Sex y Age no son independientes.
- Sex \times Survived: Sex y Survived no son independientes.
- Age \times Survived: Age y Survived no son independientes.
- Class \times Sex \times Age, Class \times Sex \times Survived, Class \times Age \times Survived, Sex \times Age \times Survived: hay interacción triple entre las variables.
- Class \times Sex \times Age \times Survived: hay interacción cuádruple entre las variables.

Problema 10

Se ha realizado un análisis sobre el valor terapéutico del ácido ascórbico (vitamina C) en relación a su efecto sobre la gripe común. Se tiene una tabla 2×2 con los recuentos correspondientes para una muestra de 279 personas:

	Gripe	No Gripe	Totales
Placebo	31	109	140
Acido Ascorbico	17	122	139
Totales	48	231	279

Aplicar un modelo lineal para determinar si existe evidencia suficiente para asegurar que el ácido ascórbico ayuda a tener menos gripe.

