

EL MODELO DE POISSON

Si la variable respuesta es una variable de conteo el mejor modelo es el modelo de Poisson.

Las variables de conteo tienen la particularidad que son siempre positivas y son discretas.

Un modelo gaussiano o de regresión como el que has visto no es suficiente.

En este ejercicio quiero que trabajes el modelo que has visto en los vídeos para que lo entiendas paso a paso.

¡Empezamos!

EL OBJETIVO DEL MODELO

Se han listado diferentes colegios dónde se han recogido las variables siguientes:

- **num_awards** es la variable respuesta (de salida) e indica la cantidad de premios obtenidos por los estudiantes en una escuela secundaria en un año,
- **Math**: las matemáticas son una variable predictora continua y representan las notas de los estudiantes en su examen final de matemáticas
- **Prog** es una variable predictora categórica con tres niveles el tipo de programa en el que los estudiantes se inscribieron. Está codificada como 1 = "General", 2 = "Académico" y 3 = "Vocacional".

La variable de salida es el num_awards

Las variables de entrada son Math y Prog

¿Según este modelo qué objetivo perseguimos? ¿Qué podremos ver?

Buscamos determinar la contribución de las variables math y prog en el número de premios que obtienen los estudiantes.

Calcula el modelo de Poisson y comprueba las restricciones del mismo.

En cuanto a las restricciones:

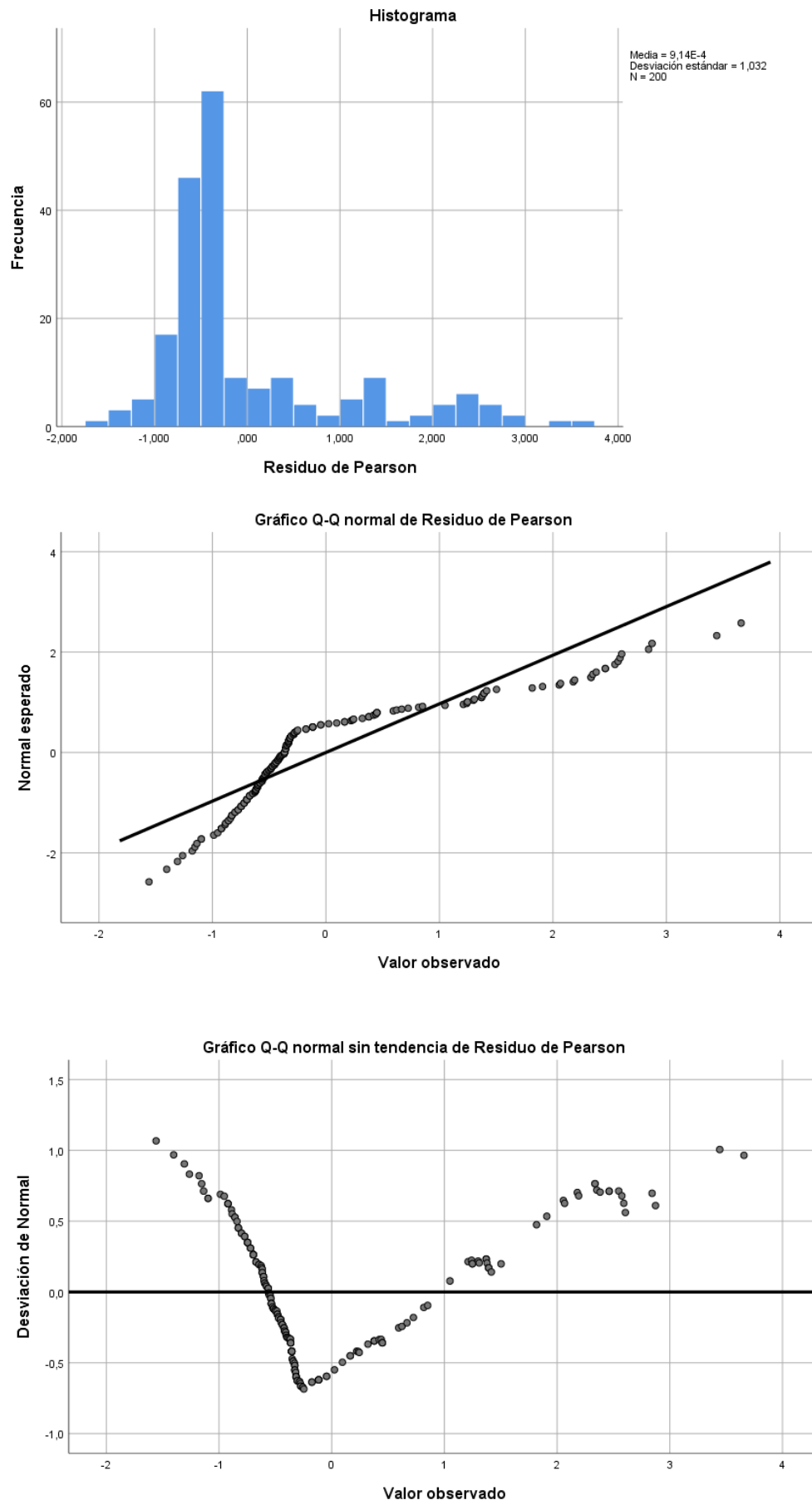
- La variable de salida es de conteo: lo cumple.
- Residuos normales con media = 0: No lo cumple. Los residuos tienen media próxima a 0, pero no son normales.
- Los residuos tienen varianzas constantes: No lo cumple.

Pruebas de normalidad

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Residuo de Pearson	,270	200	,000	,802	200	,000

a. Corrección de significación de Lilliefors

EL MODELO DE POISSON



EL MODELO DE POISSON

- No sobredispersión. La varianza es discretamente mayor que la media y la desviación es < 1 .

Estadísticos descriptivos

	N	Media	Varianza
num_awards	200	,63	1,109
N válido (por lista)	200		

Bondad de ajuste^a

	Valor	gl	Valor/gl
Desviación	189,450	196	,967
Desviación escalada	189,450	196	
Chi-cuadrado de Pearson	212,144	196	1,082
Chi-cuadrado de Pearson escalado	212,144	196	
Logaritmo de la verosimilitud ^b	-182,752		
Criterio de información Akaike (AIC)	373,505		
AIC corregido para muestras finitas (AICC)	373,710		
Criterio de información bayesiana (BIC)	386,698		
AIC coherente (CAIC)	390,698		

Variable dependiente: num_awards

Modelo: (Intersección), prog, math

a. Los criterios de información están en un formato de cuanto más pequeño mejor.

b. La función de logaritmo de la verosimilitud completa se visualiza y utiliza en el cálculo de los criterios de información.

Estimaciones de parámetro

Parámetro	B	Desv. Error	95% de intervalo de confianza de Wald		Contraste de hipótesis			Exp(B)	95% de intervalo de confianza de Wald para Exp(B)	
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.		Inferior	Superior
(Intersección)	-5,247	,6585	-6,538	-3,957	63,503	1	,000	,005	,001	,019
[prog=3]	,370	,4411	-,495	1,234	,703	1	,402	1,447	,610	3,436
[prog=2]	1,084	,3583	,382	1,786	9,153	1	,002	2,956	1,465	5,966
[prog=1]	0 ^a	.	.	.	.	.	.	1	.	.
math	,070	,0106	,049	,091	43,806	1	,000	1,073	1,051	1,095
(Escala)	1 ^b									

Variable dependiente: num_awards

Modelo: (Intersección), prog, math

a. Definido en cero porque este parámetro es redundante.

b. Fijado en el valor visualizado.

ECUACIÓN DEL MODELO:

$\text{Log}(\text{num_awards}) = - 5.247 + 0.37 \cdot \text{Prog_Voc} + 1.084 \cdot \text{Prog_Ac} + 0.07 \cdot \text{math}$

¿Tenemos sobre dispersión?

Podemos determinar que hay sobredispersión si:

EL MODELO DE POISSON

- La varianza es mayor que la media.
- La desvianza > 1.

En este caso podemos afirmar que no hay sobredispersión.

En ambos casos (con sobre dispersión o no) Calcula el modelo de binomial negativa o quasipoisson. Este modelo nos permite ser más finos en caso de sobre dispersión (>1):

Bondad de ajuste^a

	Valor	gl	Valor/gl
Desvianza	169,763	195	,871
Desvianza escalada	169,763	195	
Chi-cuadrado de Pearson	192,122	195	,985
Chi-cuadrado de Pearson escalado	192,122	195	
Logaritmo de la verosimilitud ^b	-181,905		
Criterio de información Akaike (AIC)	373,811		
AIC corregido para muestras finitas (AICC)	374,120		
Criterio de información bayesiana (BIC)	390,302		
AIC coherente (CAIC)	395,302		

Variable dependiente: num_awards

Modelo: (Intersección), prog, math

a. Los criterios de información están en un formato de cuanto más pequeño mejor.

b. La función de logaritmo de la verosimilitud completa se visualiza y utiliza en el cálculo de los criterios de información.

Estimaciones de parámetro

Parámetro	B	Desv. Error	95% de intervalo de confianza de Wald		Contraste de hipótesis			Exp(B)	95% de intervalo de confianza de Wald para Exp(B)	
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.		Inferior	Superior
(Intersección)	-4,926	,6764	-6,252	-3,601	53,041	1	,000	,007	,002	,027
[prog=1]	-,367	,4526	-1,254	,520	,657	1	,418	,693	,285	1,682
[prog=2]	,708	,3338	,054	1,362	4,501	1	,034	2,030	1,055	3,905
[prog=3]	0 ^a	.	.	.	.	.	.	1	.	.
math	,071	,0116	,048	,094	37,315	1	,000	1,074	1,049	1,098
(Escala)	1 ^b									
(Binomial negativa)	,164	,1492	,027	,977						

EL MODELO DE POISSON

Variable dependiente: num_awards

Modelo: (Intersección), prog, math

a. Definido en cero porque este parámetro es redundante.

b. Fijado en el valor visualizado.

ECUACIÓN DEL MODELO:

$$\text{Log}(\text{num_awards}) = -4.926 - 0.367 \cdot \text{Prog_Gen} + 0.708 \cdot \text{Prog_Ac} + 0.071 \cdot \text{math}$$

¿Hay diferencias entre los coeficientes del modelo de Poisson y la binomial negativa o quasi poisson?

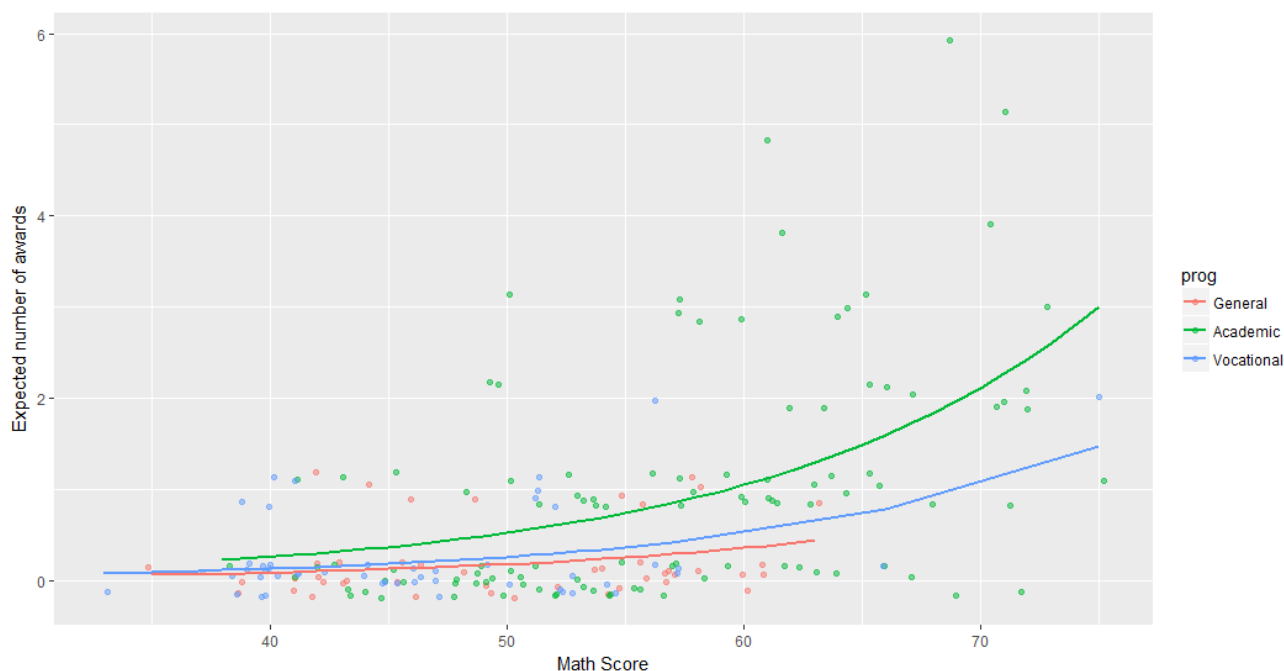
Existen diferencias todos los coeficientes, salvo el de la variable “math”, que se mantiene constante. Llama la atención que el primer modelo considera como referencia el programa general y el segundo modelo considera como referencia el vocacional.

INTERPRETA EL MODELO

Ahora que has calculado el modelo es momento de interpretarlo.

Te resumo los resultados del modelo para que te sea más fácil esta parte:

En el siguiente gráfico puedes ver el número de premios que predice el modelo en función de la nota en matemáticas y el tipo de programa.



Y este es el resultado del modelo:

EL MODELO DE POISSON

```
call:
glm(formula = num_awards ~ prog + math, family = "quasipoisson",
    data = varEstudio)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2043  -0.8436  -0.5106   0.2558   2.6796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.24712    0.68503  -7.660 8.36e-13 ***
progAcademic   1.08386    0.37272   2.908 0.00406 **
progVocational  0.36981    0.45888   0.806 0.42128
math           0.07015    0.01103   6.362 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.082366)

Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

¿Qué puedes afirmar sobre la influencia de la variable math en el modelo? ¿Cómo lo interpretas?

La variable math contribuye de forma estadísticamente significativa en el modelo, pero su coeficiente es muy bajo, lo que significa que cada punto tiene un impacto pequeño en el número de premios.

¿Qué puedes afirmar sobre la influencia de la variable prog en el modelo? ¿Cómo lo interpretas?

Aunque se considera como referencia el programa general, la contribución del programa vocacional no es estadísticamente significativa. El único programa que parece tener un impacto significativo en el número de premios es el académico, es decir, según este modelo, los estudiantes que pertenecen a este programa obtendrán más premios.

Ahora ya has aplicado por primera vez un modelo de Poisson.

Estos ejercicios te están ayudando a ver las posibilidades de los modelos y su flexibilidad.

¡Sigue así!