UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión de Poisson

Mtr. Alcides Ramos Calcina

Regresión de Poisson

Modelo de Poisson Simple

Mtr. Alcides Ramos Calcina

Introducción

- La Regresión de Poisson es un tipo especial de regresión donde la particularidad es que la variable dependiente se ajusta bien a una **distribución Poisson** .

- Las variables de conteo o recuento se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido.

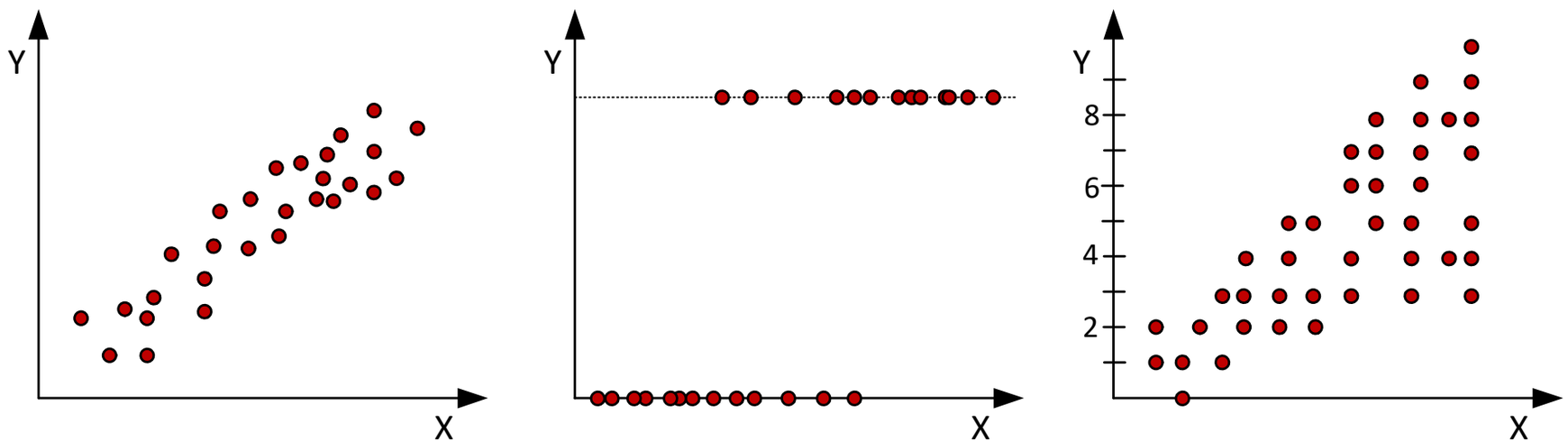
- Las variables de recuento presentan dos características que la diferencian de una variable cuantitativa continua, estas son su naturaleza **discreta** y **no negativa**.

- Por consiguiente, los modelos de regresión de Poisson se utilizan para modelar estos eventos en los que se **cuentan los resultados**.

Introducción

- Los modelos de Poisson resultan especialmente adecuados para modelar valores enteros **no negativos**, especialmente cuando la **frecuencia de ocurrencia es baja**.

- Los siguientes gráficos muestran los tres tipos de regresión comentados hasta ahora, para situarnos mejor.



1. Modelo de Poisson Simple

Ahora estamos interesados en estudiar la relación que existe entre la variable dependiente que sigue una distribución de Poisson y una o varias variables explicativas.

Iniciemos por el caso más sencillo, el modelo de regresión de Poisson simple, en el cual consideraremos una única variable explicativa.

Sea $Y \sim \text{Pois}(\lambda)$ una variable respuesta resultado de un conteo que toma valores en el conjunto $\{0, 1, 2, 3, \ldots\}$ y X una variable explicativa.

Sabemos que, la función de regresión es la media condicionada de la variable respuesta en función de cada valor de la variable explicativa, es decir:

$$\lambda(x) = E[Y/X = x], \qquad \forall x$$

1. Modelo de Poisson Simple

La función de enlace es una función g de la esperanza de Y, $\mathrm{E}(Y) = \lambda$, que relaciona λ con el predictor lineal:

$$g(\lambda) = \beta_0 + \beta_1 x$$

Como función link parece razonable elegir el logaritmo porque la función de regresión está en el intervalo $[0; +\infty\rangle$, solo toma valores no negativos, entonces el modelo es el siguiente:

$$\log(\lambda) = \beta_0 + \beta_1 X$$

Al aplicar exponenciales, la función de regresión del modelo queda expresada de la siguiente forma:

$$\lambda = e^{\beta_0 + \beta_1 X} = e^{\beta_0} e^{\beta_1 X}$$

1. Modelo de Poisson Simple

1.1. Estimación de parámetros

Se considera una muestra aleatoria simple de tamaño n

$$(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$$

donde x_1, …, x_n son las obtenciones de la variable explicativa e Y_1, …, Y_n son las realizaciones de la variable respuesta y además $Y_i \sim \text{Pois}(\lambda(x_i))$.

Entonces, en este caso:

$$\log(\lambda) = \beta_0 + \beta_1 x_i \qquad \text{y} \qquad \lambda = e^{\beta_0 + \beta_1 x_i}$$

1. Modelo de Poisson Simple

1.2. Interpretación de los parámetros

De la expresión $E[Y] = \lambda = e^{\beta_0} e^{\beta_1 X}$, podemos interpretar los parámetros del modelo de regresión de Poisson de la siguiente manera:

- e^{β_0} : Valor inicial de la variable respuesta, es decir, el valor de la función de regresión cuando $x = 0$.

- e^{β_1} : Tasa de incremento de la respuesta esperada al incrementar una unidad la variable explicativa, esto es, pasando de x a $x + 1$.

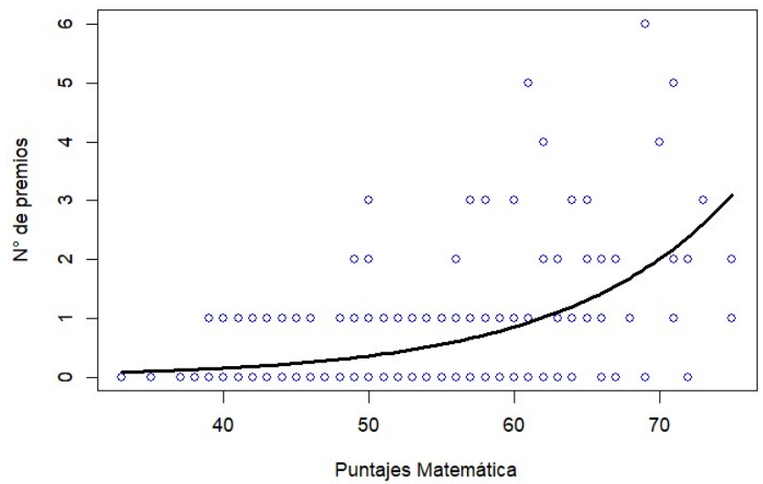
1. Modelo de Poisson Simple

1.2. Supuestos del modelo de regresión de Poisson

- **Variable de respuesta con distribución de Poisson.** La variable de respuesta es un recuento por unidad de tiempo o espacio, descrito por una distribución.
- **Independencia**. Las observaciones deben ser independientes entre sí.
- **Media = Varianza.** Por definición, la media de una variable aleatoria de Poisson debe ser igual a su varianza.
- **Linealidad**. El logaritmo de la tasa media, $\log(\lambda)$, debe ser una función lineal de x.

EJEMPLOS

$$\log(\lambda) = \beta_0 + \beta_1 x_i$$



Ejemplo 7

Se tiene la variable "**num_awards**" la cual es el resultado de indicar el número de premios obtenidos por los estudiantes en una escuela secundaria en un año. Los predictores son: los puntajes de los estudiantes en su examen final de matemáticas (**math**), y la variable categórica que indica el tipo de programa en el que se encontraban los estudiantes inscritos (**prog**), con tres niveles, la cual se codifica como: 1 = "General", 2 = "Académico" y 3 = "Vocacional".

Los datos se encuentran en el archivo ***awards.txt***, cada variable tiene 200 observaciones. Comencemos cargando los datos y mostrando los primeros datos.

Ejemplo 7

```
datos <- read.csv("C:/…/awards.txt", sep="")
View(datos)
head(datos)
```

```
   id num_awards prog math
1  45          0    3   41
2 108          0    1   41
3  15          0    3   44
4  67          0    3   42
5 153          0    3   40
6  51          0    1   42
```

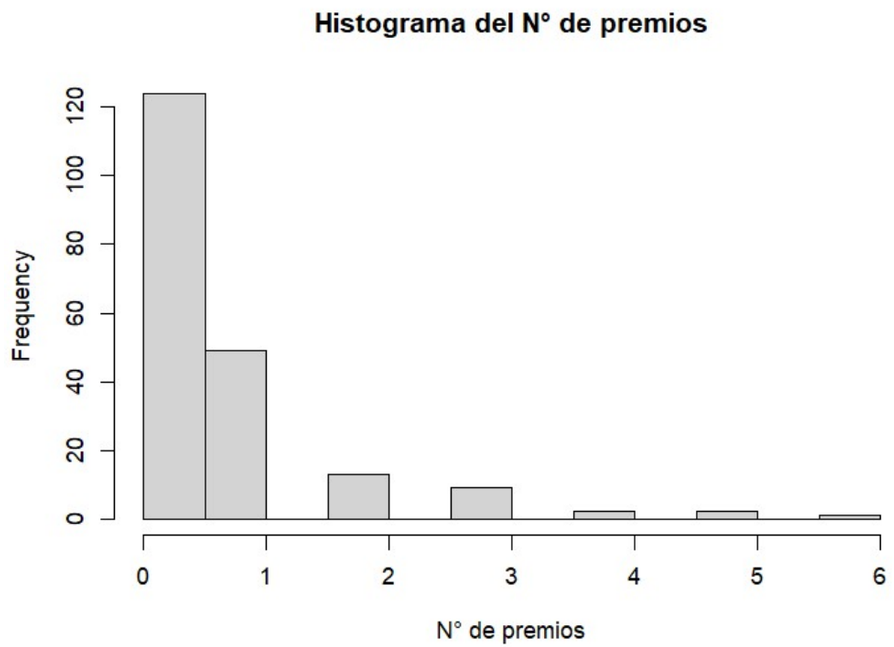
A efectos ilustrativos, estimaremos dos modelos de Poisson simples, el primero con el predictor cuantitativo y el segundo con el categórico.

Ejemplo 7

Solución

- Modelo: $\log(\lambda) = \beta_0 + \beta_1 math$

Iniciamos realizando un análisis descriptivo de los datos, y en primer lugar graficamos el histograma para ver la distribución de la variable dependiente.



Claramente, los datos no tienen la forma de una curva de campana como en una distribución normal. Por tanto, existen altos indicios de que $Y \sim \text{Pois}(\lambda)$.

Ejemplo 7

Ahora veamos la media y la varianza. .

```
mean(num_awards)
[1] 0.63

sd(num_awards)
[1] 1.052921
```

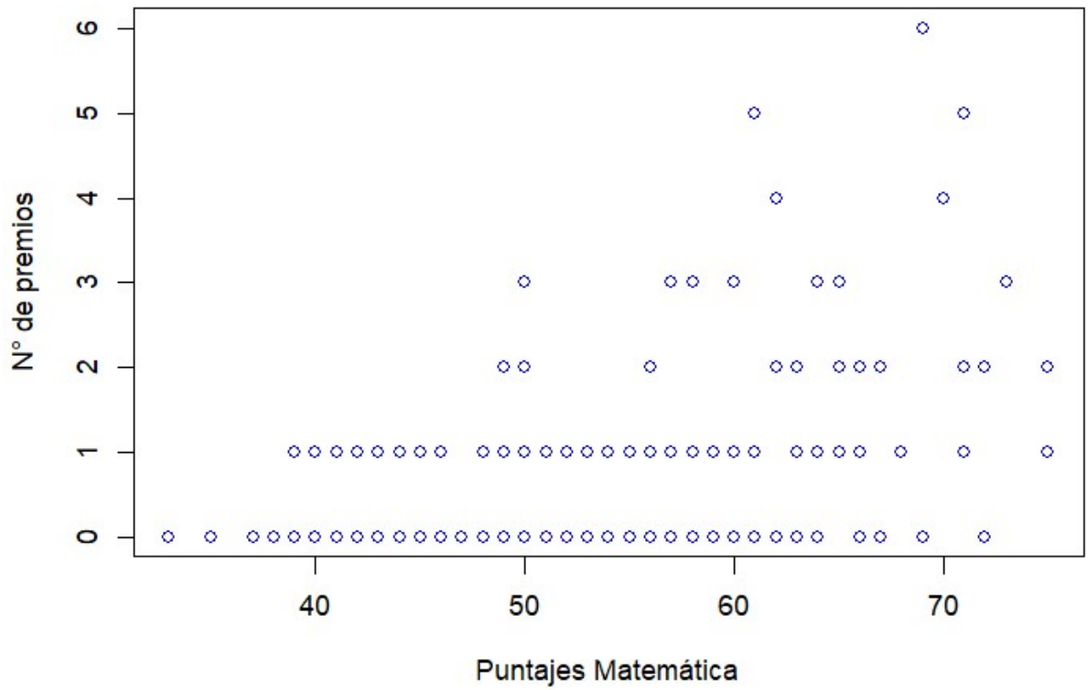
La media y la varianza no se diferencia mucho, por tanto, es un buen candidato para realizar el modelamiento con regresión de Poisson.

Si la varianza es mucho mayor que la media, esto sugiere que tendríamos una dispersión excesiva en el modelo, problemática que se abordará más adelante.

Ejemplo 7

Así mismo, es importante realizar un diagrama de dispersión, el cual nos ayude a visualizar el comportamiento de los datos de *Y* en función a *x*.

```
plot(math, num_awards, col = "blue", ylab = "N° de premios", xlab = "Puntajes Matemática")
```



La Figura nos da más indicios del ajuste de los datos a un modelo de Poisson

Ejemplo 7

Estimación del modelo de Poisson simple:

```
modelo1 <- glm(num_awards ~ math, family = "poisson", data = datos)
summary(modelo1)
Call:
glm(formula = num_awards ~ math, family = "poisson", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1853  -0.9070  -0.6001   0.3246   2.9529
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.333532   0.591261  -9.021   <2e-16 ***
math         0.086166   0.009679   8.902   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 204.02  on 198  degrees of freedom
AIC: 384.08

Number of Fisher Scoring iterations: 6
```

Ejemplo 7

El modelo estimado es el siguiente:

$$\log(\lambda) = -5.3335 + 0.0862 math$$

Con el propósito de realizar la interpretación de los coeficientes del modelo, calculamos las exponenciales de estos coeficientes:

$$\lambda = e^{-5.3335+0.0862 math}$$

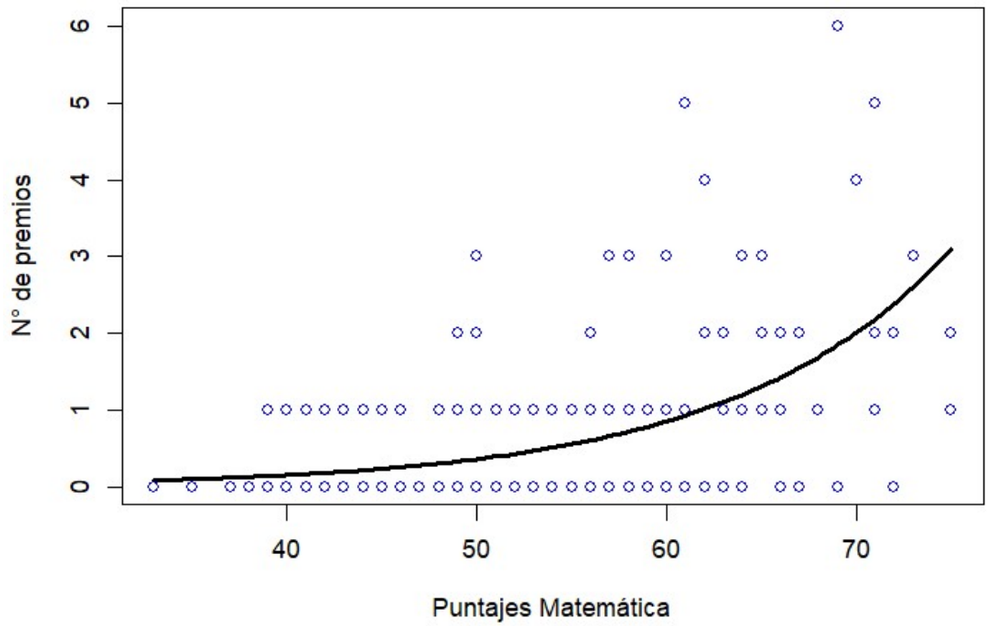
```
exp(coef(modelo1))
(Intercept)         math
0.004826991 1.089986813
```

La estimación de β_1 es 0.086166 (y su exponencial vale 1.08998), con lo cual, 1.09 es lo que se incrementan los premios recibidos cuando la variable **math** se incrementa en una unidad (es decir, cuando el estudiante sube un punto en sus calificaciones de matemática).

Ejemplo 7

Por último, representamos en la Figura 33 el ajuste del modelo sobre el diagrama de dispersión. El código necesario para obtener dicho diagrama de dispersión se muestra a continuación:

```
curve(exp(modelo1$coefficients[1] + modelo1$coefficients[2]*x), add = TRUE, lwd =3)
```



Ejemplo 7

- Modelo: $\log(\lambda) = \beta_0 + \beta_1 prog$

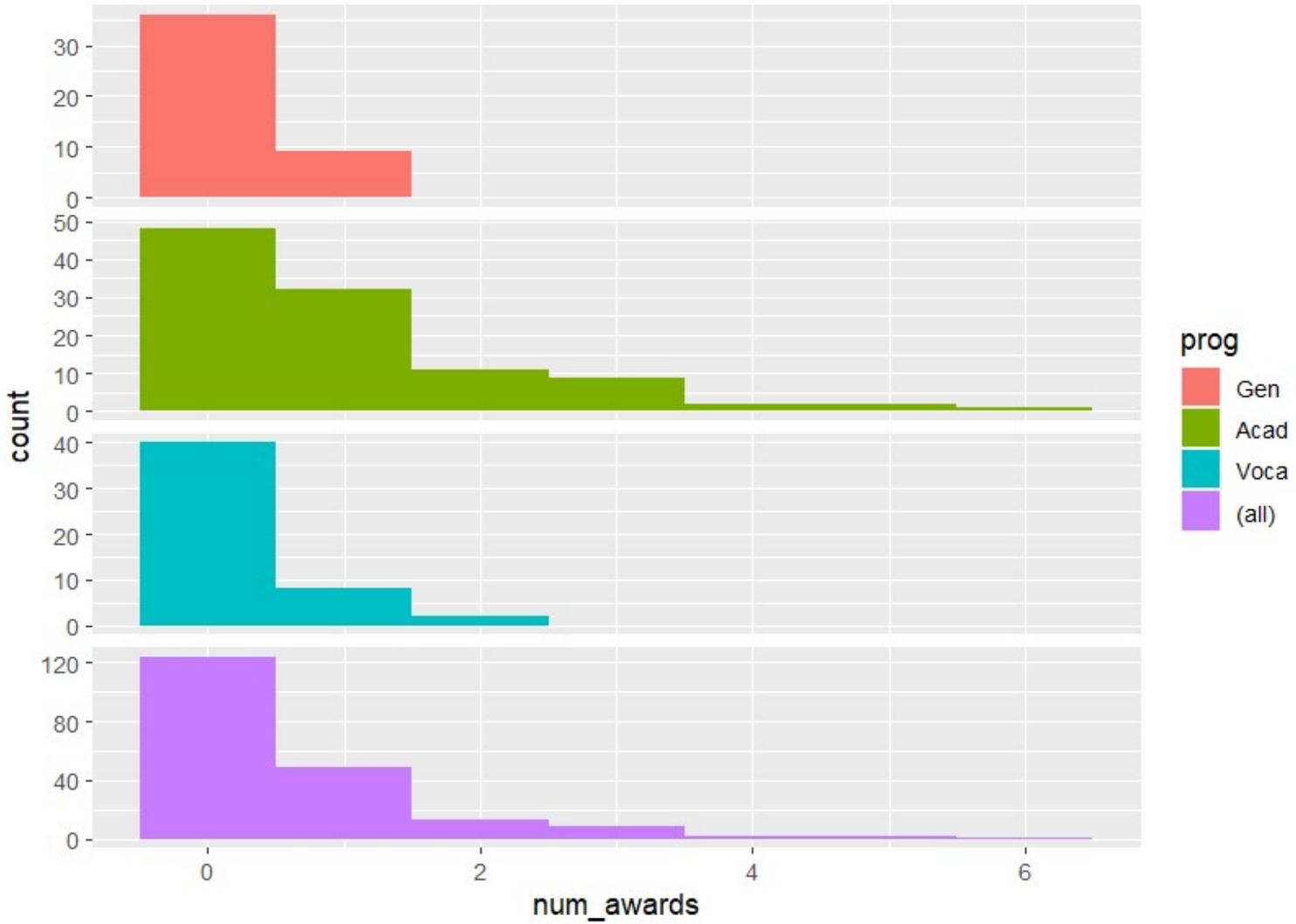
En primer lugar, convertimos la variable **prog** a factor, a través del siguiente código:

```
datos$prog <- factor(datos$prog, labels = c("Gen", "Acad", "Voca"))
```

Ahora, en el caso de variable catgorica, veremos la distribución del número premios por el tipo de programa (**prog**) y la media y varianza de los conteos en cada categoría.

```
#Activar library(ggplot2)
ggplot(datos, aes(num_awards, fill = prog)) +
  geom_histogram(binwidth = 1) +
  facet_grid(prog ~., margins = TRUE, scales = "free")+
  theme(text=element_text(size=12), strip.text.y = element_blank())
```

Ejemplo 7


count
30
20
10
0
50
40
30
20
10
0
40
30
20
10
0
120
80
40
0
0
2
4
6
num_awards
prog
Gen
Acad
Voca
(all)

Ejemplo 7

Media y varianza por tipo de programa.

```
d <- with(datos, tapply(num_awards, prog, function(x) {data.frame(Media
= mean(x), Varianza = sd(x)^2)}))
unlist(d)
    Gen.Media  Gen.Varianza
    0.2000000     0.1636364

 Acad.Media Acad.Varianza
 1.0000000     1.6346154

 Voca.Media Voca.Varianza
 0.2400000     0.2677551
```

Las medias y varianzas en cada categoría no se diferencian mucho, por tanto, el tipo de programa parecer ser buen candidato para predecir el número de premios.

Ejemplo 7

En este punto, estamos listos para realizar nuestro análisis de modelo de Poisson.

```
modelo2 <- glm(num_awards ~ prog, family = "poisson", data = datos)
summary(modelo2)
Call:
glm(formula = num_awards ~ prog, family = "poisson", data = datos)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6094     0.3333  -4.828 1.38e-06 ***
progAcad      1.6094     0.3473   4.634 3.59e-06 ***
progVoca      0.1823     0.4410   0.413    0.679
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 234.46  on 197  degrees of freedom
AIC: 416.51

Number of Fisher Scoring iterations: 6
```

Ejemplo 7

El modelo estimado es el siguiente:

$$\log(\lambda) = -1.6094 + 1.6094\, progAcad + 0.1823\, progVoca$$

Calculamos las exponenciales de estos coeficientes:

$$\lambda = e^{-1.6094 + 1.6094\, progAcad + 0.1823\, progVoca}$$

```
exp(coef(modelo2))
(Intercept)    progAcad    progVoca
        0.2         5.0         1.2
```

Ejemplo 7

Interpretación de coeficientes

- Para $e^{\beta 11}$ es 5.0, lo que nos indica que, la probabilidad de incrementar en una unidad los premios es 5 veces mas probable en los estudiantes en el programa “Académico” respecto a los del programa “General”.

- Para $e^{\beta 12}$ es 1.2, indica que, la probabilidad de incrementar en recibir un premio es 1.2 veces mas probable en los estudiantes en el programa “Vocacional” respecto a los del programa “General”.

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS