Modelo de Variable Dependiente Discreta

Modelos de datos contables

Felipe J. Quezada-Escalona
Departamento de Economía

Introducción

- **Regresión de Poisson**
  - Modelo de regresión para **datos de conteo**, donde la variable dependiente representa el **número de eventos en un tiempo** o espacio específico.
  - Los datos de conteo son no negativos y discretos (0, 1, 2,…).
- Ejemplos:
  - Número de cervezas que se venden en un bar entre las 23:00 y las 00:00 horas.
  - Número de accidentes de tráfico en una intersección por día.
  - Número de llamadas telefónicas recibidas por un centro de atención al cliente por minuto.

Función de Probabilidad

El modelo de probabilidad esta dado por:

- **Función de Probabilidad:**

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

donde Y es variable discreta que representa el número de eventos; λ es la **media**.

Modelo de regresion

Para estimar un modelo poisson:

- **Relación entre Media y Variables Independientes:** Generalmente uno asume que

$$\lambda = E(Y|X) = e^{X'\beta}$$

  donde $X = (X_1, X_2, ..., X_k)'$ es un vector de variables explicativas.
- Es decir, el modelo de Poisson es log-lineal $\ln(\lambda_i) = \beta'\mathbf{X}$
- **Supuesto Clave:** Y sigue una distribución de Poisson con media λ, donde λ es función de X. Es decir, $Y \sim \text{Poisson}(\lambda(X))$

Modelo de regresion

En Poisson log-lineal $\ln(\lambda_i) = \beta'\mathbf{X}$:

- **Coeficientes (β_j):**
  - Representan el cambio en el logaritmo de la media de Y por un cambio de una unidad en X_j, manteniendo constantes las demás variables.
  - $\exp(\beta_j)$ representa el cambio multiplicativo en la media de Y con un cambio unitario en X_j.
- **Ejemplo:** Si $\exp(\beta_1) = 1,2$, un aumento de una unidad en X_1 se asocia con un aumento del 20% en la media de Y.

Función de Verosimilitud

Considerando la función de probabilidad, la verosimilitud es:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

donde $\lambda_i = e^{X_i'\beta}$.

y la **Log-Verosimilitud** es:

$$\ell(\beta) = \sum_{i=1}^{n} (y_i X_i'\beta - e^{X_i'\beta} - \ln(y_i!))$$

Inferencia

- Los **Errores Estándar** se calculan a partir de la matriz de información de Fisher.
- Y las **Pruebas de Hipótesis** se pueden realizar mediante pruebas t para la significancia individual de los coeficientes y pruebas de Wald para restricciones conjuntas.
- Respecto a los **Intervalos de Confianza**, se pueden construir intervalos de confianza para los coeficientes de forma usual.

Sobredispersión

La sobredispersión ocurre cuando la varianza de Y es mayor que su media, violando el supuesto de Poisson (**equisdispersión**).

- **Causas:**
  - Heterogeneidad no observada.
  - Omisión de variables relevantes.
  - Eventos correlacionados.
- **Consecuencias:**
  - Sesgo en estimaciones de errores estándar.
  - Pruebas de hipótesis incorrectas.

Sobredispersión

Soluciones:

- **Regresión Binomial Negativa:** Modelo alternativo para permitir sobredispersión.
  - En este caso la varianza es $Var(y_i) = \mu + \alpha\mu_i^p$, donde $p = 1, 2$ (Tipo 1 $p = 1$, y en Tipo 2 es $p = 2$)
  - α es el parametro que mide la dispersión.
- **Estimación Robusta:** Ajuste de errores estándar para sobredispersión.

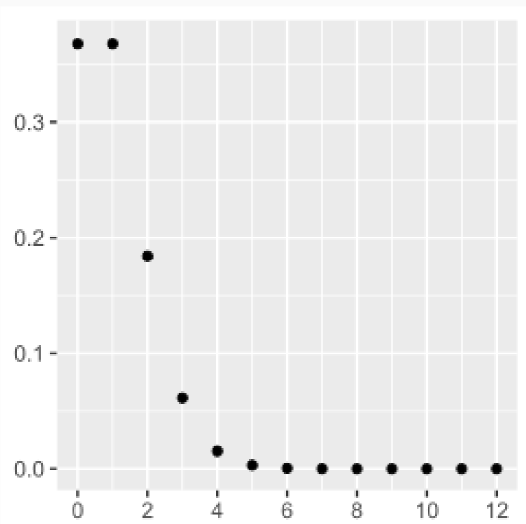
Ejemplo a mano

Tenemos diez puntos de datos de una distribución Poisson, pero ¿cuál es el parámetro λ de la distribución?

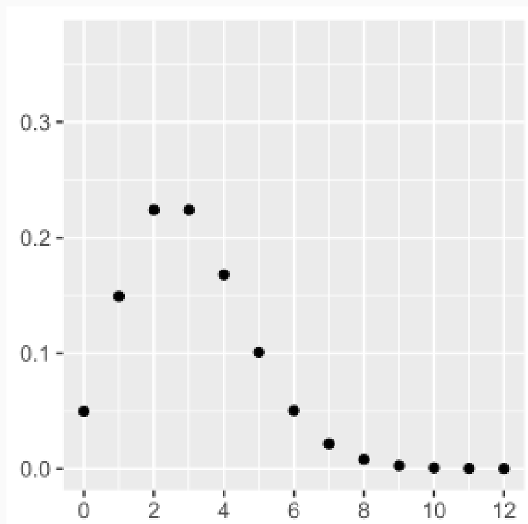
$$y = 2, 0, 1, 2, 2, 2, 0, 2, 1, 1$$

$$f(y \mid \lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

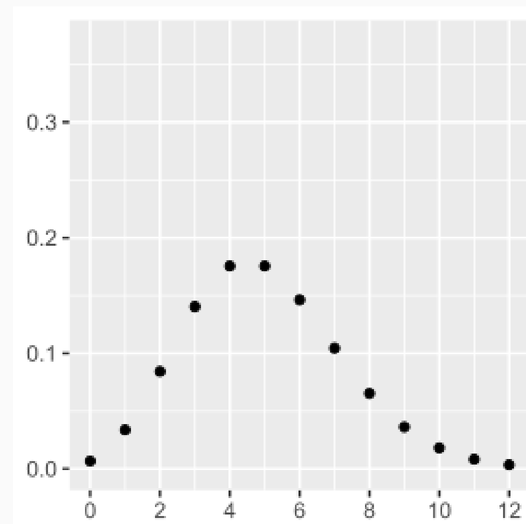
$\lambda = 1$



$\lambda = 3$



$\lambda = 5$



Ejemplo a mano

Tenemos diez puntos de datos de una distribución Poisson, pero ¿cuál es el parámetro λ de la distribución?

$$\mathbf{y} = \{2, 0, 1, 2, 2, 2, 0, 2, 1, 1\}$$

$$L(\lambda \mid \mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

$$\ln L(\lambda \mid \mathbf{y}) = -n\lambda + \ln\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!)$$

$$\frac{\partial \ln L(\lambda \mid \mathbf{y})}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i$$

$$\frac{\partial \ln L(\lambda \mid \mathbf{y})}{\partial \lambda} = 0 \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i = 1.3$$

Ejemplo a mano

Tenemos diez puntos de datos de una distribución Poisson, pero ¿cuál es el parámetro λ de la distribución?

$$\mathbf{y} = \{2, 0, 1, 2, 2, 2, 0, 2, 1, 1\}$$

$$L(\lambda \mid \mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} = \frac{e^{-10\lambda}\lambda^{13}}{32}$$

$$\ln L(\lambda \mid \mathbf{y}) = -n\lambda + \ln\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!) = -10\lambda + 13\ln\lambda - 3.47$$

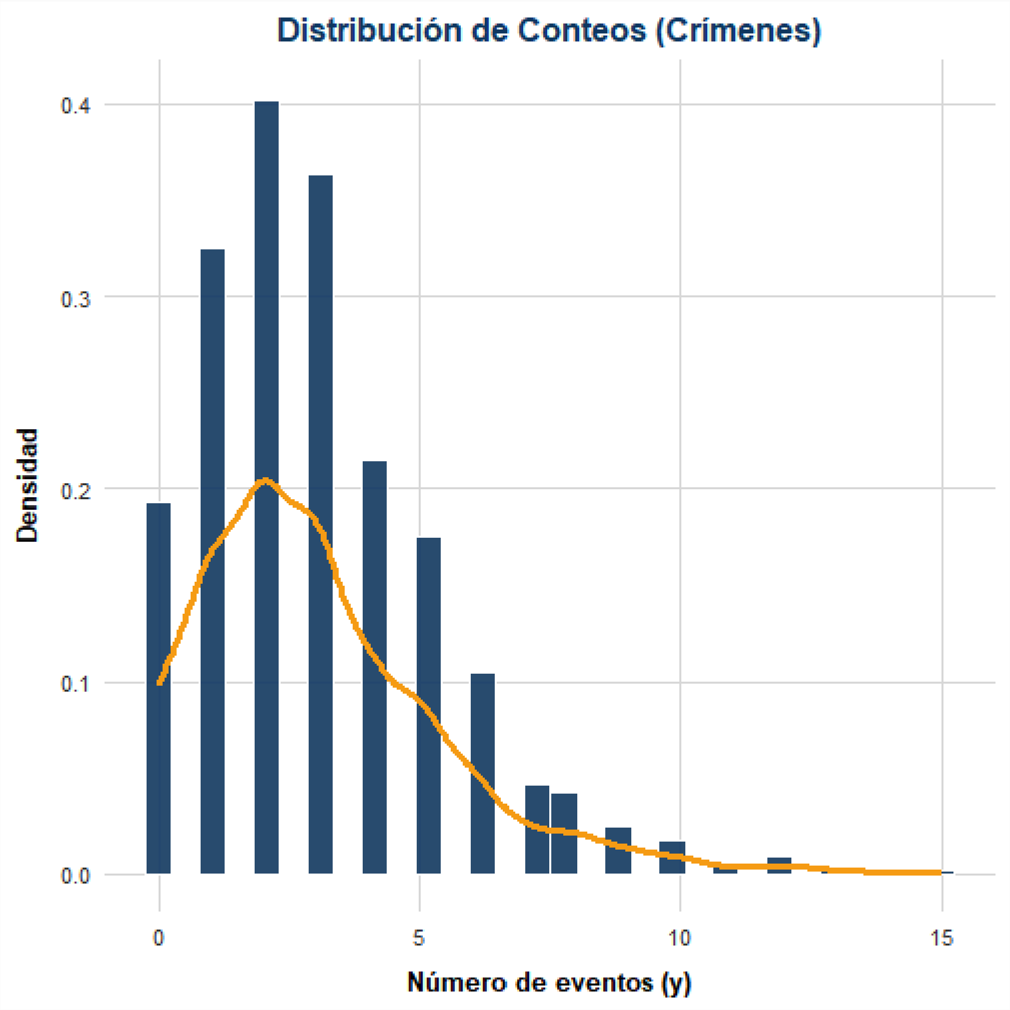
Software

- Stata: `rpoisson`
- R: `glm`

Ejemplo en R: Poisson

```r
# 1. Datos a usar
set.seed(123)
n      <- 1000
x      <- rnorm(n, mean = 0, sd = 1)
b0     <- 1
b1     <- 0.5
lambda <- exp(b0 + b1 * x)
y      <- rpois(n, lambda)      # (ej., número crímenes)
```

Histograma de Y



Estimación por Máxima Verosimilitud

```
library(maxLik)

ch_ll<- function(params) {
  b0 <- params[1]
  b1 <- params[2]
  xb <- b0 + b1 * x
  mu <- exp(xb)
  ll <- -mu + y * xb - lfactorial(y)
  return(sum(ll))
}

mle_poisson <- maxLik(logLik=ch_ll, start=c(0.1, 0.1))
summary(mle_poisson)
```

Estimación por Máxima Verosimilitud

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 4 iterations
## Return code 8: successive function values within relative tolerance limit (re
## Log-Likelihood: -1889.833
## 2  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]  1.00000    0.02012   49.71  <2e-16 ***
## [2,]  0.47950    0.01802   26.61  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

Comparación con glm

```r
glm(y ~ x, family = poisson(link = "log"))
```

```
##
## Call:  glm(formula = y ~ x, family = poisson(link = "log"))
##
## Coefficients:
## (Intercept)            x
##      1.0000       0.4795
##
## Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
## Null Deviance:       1845
## Residual Deviance: 1140      AIC: 3784
```

¡Muchas gracias!

Felipe J. Quezada-Escalona

Department of Economics

Universidad de Concepción

¿Preguntas?