Aplicaciones del modelo de regresión de Poisson: cálculo de predicciones

Contents

1 Predicción del valor medio

Sea el modelo de regresión de Poisson

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, 3, \ldots, \quad i = 1, 2, \cdots, n$$

donde:

$$\lambda_i = exp(x_i^T \beta)$$

$$x_i = \begin{bmatrix} 1 \\ x_{1i} \\ x_{2i} \\ \cdots \\ x_{ki} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix}$$

Estamos interesados en el valor de la respuesta para los regresores $x_p^T = [1 \ x_{1p} \ x_{2p} \ \cdots \ x_{kp}]$. El valor predicho de λ_i en x_p es:

$$\hat{\lambda}_p = exp(x_p^T \hat{\beta})$$

donde $\hat{\beta}$ es el vector de parámetros estimados:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{bmatrix}$$

Una vez que hemos calculado $\hat{\lambda}_p$ podemos predecir probabilidades:

$$\hat{P}(Y_p = y) = \frac{e^{-\hat{\lambda_p}}\hat{\lambda_p}^y}{y!}, \quad y = 0, 1, 2, 3, \ldots$$

es decir

$$\hat{P}(Y_p = 0) = \frac{e^{-\hat{\lambda_p}}\hat{\lambda_p}^0}{0!}$$

$$\hat{P}(Y_p = 1) = \frac{e^{-\hat{\lambda_p}}\hat{\lambda_p}^1}{1!}$$

$$\ldots$$

2 Intervalo de confianza para λ_p

Se tiene que

$$\hat{\beta} \sim N(\beta, (X^T W X)^{-1})$$

Por tanto

$$x_p^T \hat{\beta} \sim N(x_p^T \beta, x_p^T (X^T W X)^{-1} x_p)$$

ya que

$$E[x_p^T \hat{\beta}] = x_p^T E[\hat{\beta}] = x_p^T \beta$$

y

$$Var[x_p^T \hat{\beta}] = x_p^T Var[\hat{\beta}] x_p = x_p^T (X^T W X)^{-1} x_p$$

Por tanto, el intervalo de confianza para $x_p^T \beta$ es

$$x_p^T \hat{\beta} - z_{\alpha/2}\sqrt{x_p^T (X^T W X)^{-1} x_p} \leq x_p^T \beta \leq x_p^T \hat{\beta} + z_{\alpha/2}\sqrt{x_p^T (X^T W X)^{-1} x_p}$$

Si llamamos:

$$L_p = x_p^T \hat{\beta} - z_{\alpha/2}\sqrt{x_p^T (X^T W X)^{-1} x_p} U_p = x_p^T \hat{\beta} + z_{\alpha/2}\sqrt{x_p^T (X^T W X)^{-1} x_p}$$

se tiene que

$$exp(L_p) \leq \lambda_p \leq exp(U_p)$$

donde se recuerda que

$$\lambda_p = exp(x_p^T \beta)$$

3 Ejemplos

```r
d = read.csv("datos/Aircraft_Damage.csv")
d$bomber = factor(d$bomber, labels = c("A4","A6"))
```

Primero estimamos el modelo:

```r
m = glm(damage ~ bomber + load + experience, data = d, family = poisson)
summary(m)
```

```
##
## Call:
## glm(formula = damage ~ bomber + load + experience, family = poisson,
##     data = d)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.406023   0.877489  -0.463   0.6436
## bomberA6     0.568772   0.504372   1.128   0.2595
## load         0.165425   0.067541   2.449   0.0143 *
## experience  -0.013522   0.008281  -1.633   0.1025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 53.883  on 29  degrees of freedom
## Residual deviance: 25.953  on 26  degrees of freedom
## AIC: 87.649
##
## Number of Fisher Scoring iterations: 5
```

Queremos calcular la predicción en bomber = A-4 (0), load = 6, experience = 75:

```r
xp = c(1,0,6,75)
beta_e = coef(m)
( lambda_p = exp(t(xp) %*% beta_e) )
```

```
##           [,1]
## [1,] 0.6520434
```

Por tanto, la probabilidad de que un A-4 cargado con 6 Tn de explosivos y 75 meses de experiencia de vuelo tenga 0 zonas con daño es:

$$\hat{P}(Y = 0) = \frac{e^{-\hat{\lambda_p}} \hat{\lambda_p}^0}{0!}$$

```r
dpois(0,lambda_p)
```

```
## [1] 0.5209801
```

La probabilidad de que tenga 1 zona con daño es:

$$\hat{P}(Y = 1) = \frac{e^{-\hat{\lambda_p}} \hat{\lambda_p}^1}{1!}$$

```
dpois(1,lambda_p)
```

```
## [1] 0.3397016
```

La probabilidad de que tenga 2 zonas con daño es:

$$\hat{P}(Y=2) = \frac{e^{-\hat{\lambda}_p}\hat{\lambda}_p^{\,2}}{2!}$$

```
dpois(2,lambda_p)
```

```
## [1] 0.1107501
```

Por último, la probabilidad de que tenga 2 o menos zonas de daño se calcula como:

$$P(Y \leq 2) = \sum_{y=0}^{2} \frac{e^{-\hat{\lambda}_p}\hat{\lambda}_p^{\,y}}{y!}$$

```
# sumando
sum(dpois(0:2,lambda_p))
```

```
## [1] 0.9714319
```

```
# o directamente
ppois(2,lambda_p)
```

```
## [1] 0.9714319
```

Para calcular el intervalo de confianza:

```
source("funciones/poisson_funciones.R")
H = poisson_hess(coef(m),model.matrix(m))
xp = matrix(xp, ncol = 1)
(se = sqrt(- t(xp) %*% solve(H) %*% xp ))
```

```
##           [,1]
## [1,] 0.3168692
```

```
alfa = 0.05
Lp = t(xp) %*% beta_e - qnorm(1-alfa/2)*se
Up = t(xp) %*% beta_e + qnorm(1-alfa/2)*se
# limite inferior intrevalo confianza
exp(Lp)
```

```
##           [,1]
## [1,] 0.3503942
```

```
# limite superior intrevalo confianza
exp(Up)
```

```
##          [,1]
## [1,] 1.213378
```

Con R, podemos predecir el valor medio $\hat{\lambda}_p$:

```
xp_df = data.frame(bomber = "A4", load = 6, experience = 75)
(pred = predict(m, newdata = xp_df, type = "response"))
```

```
##         1
## 0.6520434
```

Para calcular el intervalo de confianza tenemos que trabajar con el *link*. En regresión de Poisson el link es:

$$log(\lambda_i) = x_i^T \beta$$

```r
(pred = predict(m, newdata = xp_df, type = "link", se.fit = T))
```

```
## $fit
##          1
## -0.4276442
##
## $se.fit
## [1] 0.3168688
##
## $residual.scale
## [1] 1
```

```r
alfa = 0.05
Lp = pred$fit - qnorm(1-alfa/2)*pred$se.fit
Up = pred$fit + qnorm(1-alfa/2)*pred$se.fit
# limite inferior intervalo confianza
exp(Lp)
```

```
##         1
## 0.3503945
```

```r
# limite superior intervalo confianza
exp(Up)
```

```
##        1
## 1.213377
```

Por tanto, la probabilidad predicha de que el avión A-4, con carga 6 y experiencia 75 tenga 2 daños es:

$$\hat{P}(Y=2) = \frac{e^{-\hat{\lambda}_p} \hat{\lambda}_p^2}{2!}$$

```r
dpois(2,lambda_p)
```

```
## [1] 0.1107501
```

Con intervalo de confianza

```r
c(dpois(2,exp(Lp)), dpois(2,exp(Up)))
```

```
## [1] 0.04324244 0.21877542
```

4 Intervalo de confianza para λ_p y para las predicciones utilizando bootstrap

```r
set.seed(99)
B = 500
n = nrow(d)
link_B = rep(0, B)
```

```r
lambda_B = rep(0,B)
pred_B = rep(0,B)
for (b in 1:B){
  pos_b = sample(1:n, n, replace = T)
  d_b = d[pos_b,]
  m_b = glm(damage ~ bomber + load + experience, data = d_b, family = poisson)
  link_B[b] = t(xp) %*% coef(m_b)
  lambda_B[b] = exp(t(xp) %*% coef(m_b))
  pred_B[b] = dpois(2, lambda_B[b])
}
```

- Prediccion puntual de λ_p calculada con bootstrap:

```r
mean(lambda_B)
```

```
## [1] 0.6249749
```

- Standard error calculado con bootstrap:

```r
(se_B = sd(link_B))
```

```
## [1] 0.2640387
```

- Invervalo de confianza calculados con bootstrap:

```r
alfa = 0.05
quantile(lambda_B, probs = c(alfa/2,1-alfa/2))
```

```
##      2.5%     97.5%
## 0.3419917 0.9250180
```

- Invervalo de confianza para los valores predichos:

```r
alfa = 0.05
quantile(pred_B, probs = c(alfa/2,1-alfa/2))
```

```
##       2.5%      97.5%
## 0.04154507 0.16964404
```