Generalized Linear Model

Valentina Zangirolami

2024-01-09

```
#Load libraries
library(car)
```

```
## Caricamento del pacchetto richiesto: carData
```

```
library(ggplot2)
```

Logistic regression

```
icu <- read.csv("ICU.csv")
head(icu)
```

```
##   stato eta causa coscienza
## 1     0  27     1         0
## 2     0  59     1         1
## 3     0  77     0         0
## 4     0  54     1         0
## 5     0  87     1         0
## 6     0  69     1         0
```

ICU Dataset contains data about 200 patients admitted at the Intensive Care Unit (ICU), with the following variables:

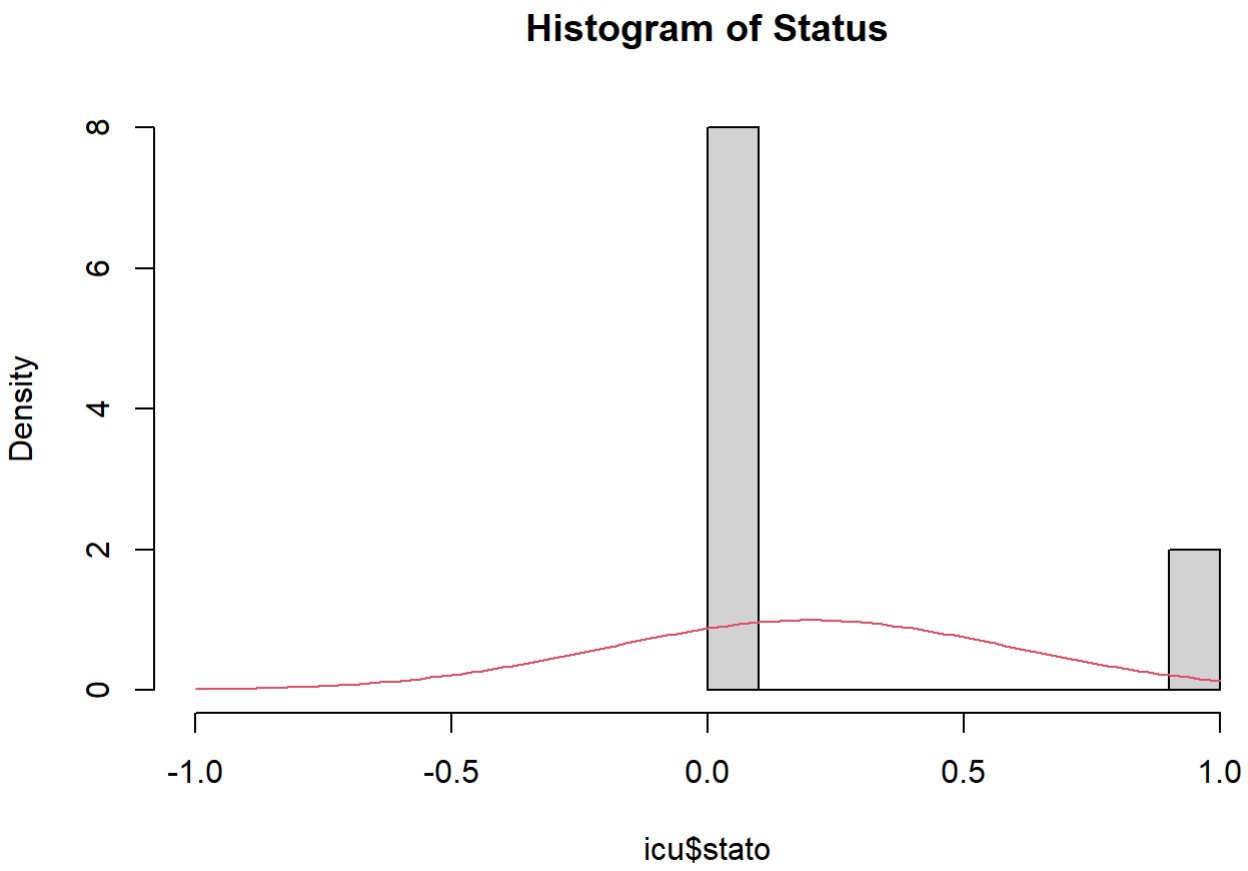
- *stato* means patient status, which is a binary variable: 0 (Alive) and 1 (Dead)
- *eta* means patient age
- *causa* means the reason of the hospitalization: 0 (planned) and 1 (emergency)
- *coscienza* means the level of consciousness: 0 (no coma) and 1 (coma)

Plot of the Dependent variable

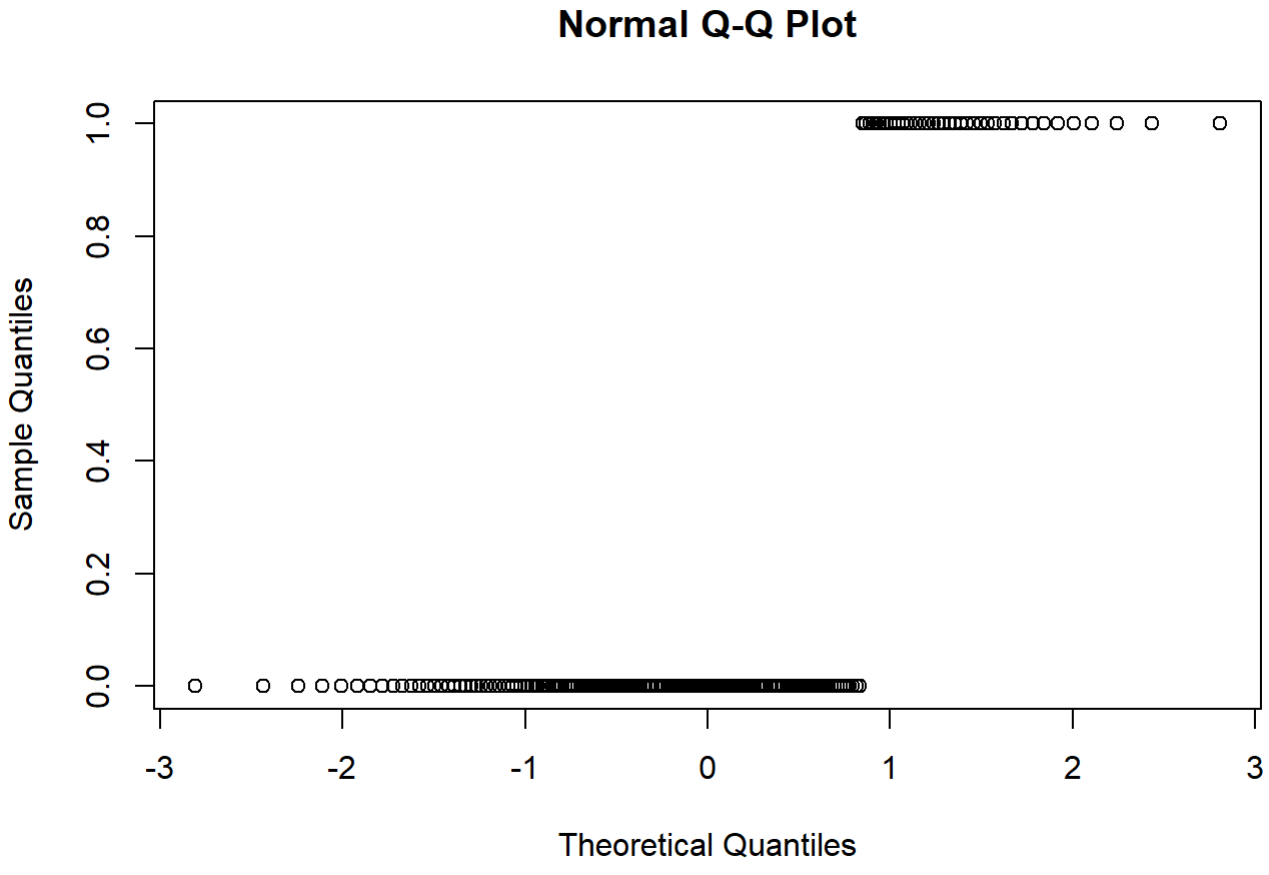
In this case, can the variable *status* be normally distributed?

Let's check

```
hist(icu$stato,prob=T, xlim=c(-1,1), main= "Histogram of Status")
curve(dnorm(x,mean(icu$stato), sd(icu$stato)),add=T, col=2)
```




```
qqnorm(icu$stato)
```



Of course, no. We can assume

$$Status_i \overset{\perp\!\!\!\perp}{\sim} Bernoulli(\pi_i)$$

and, then,

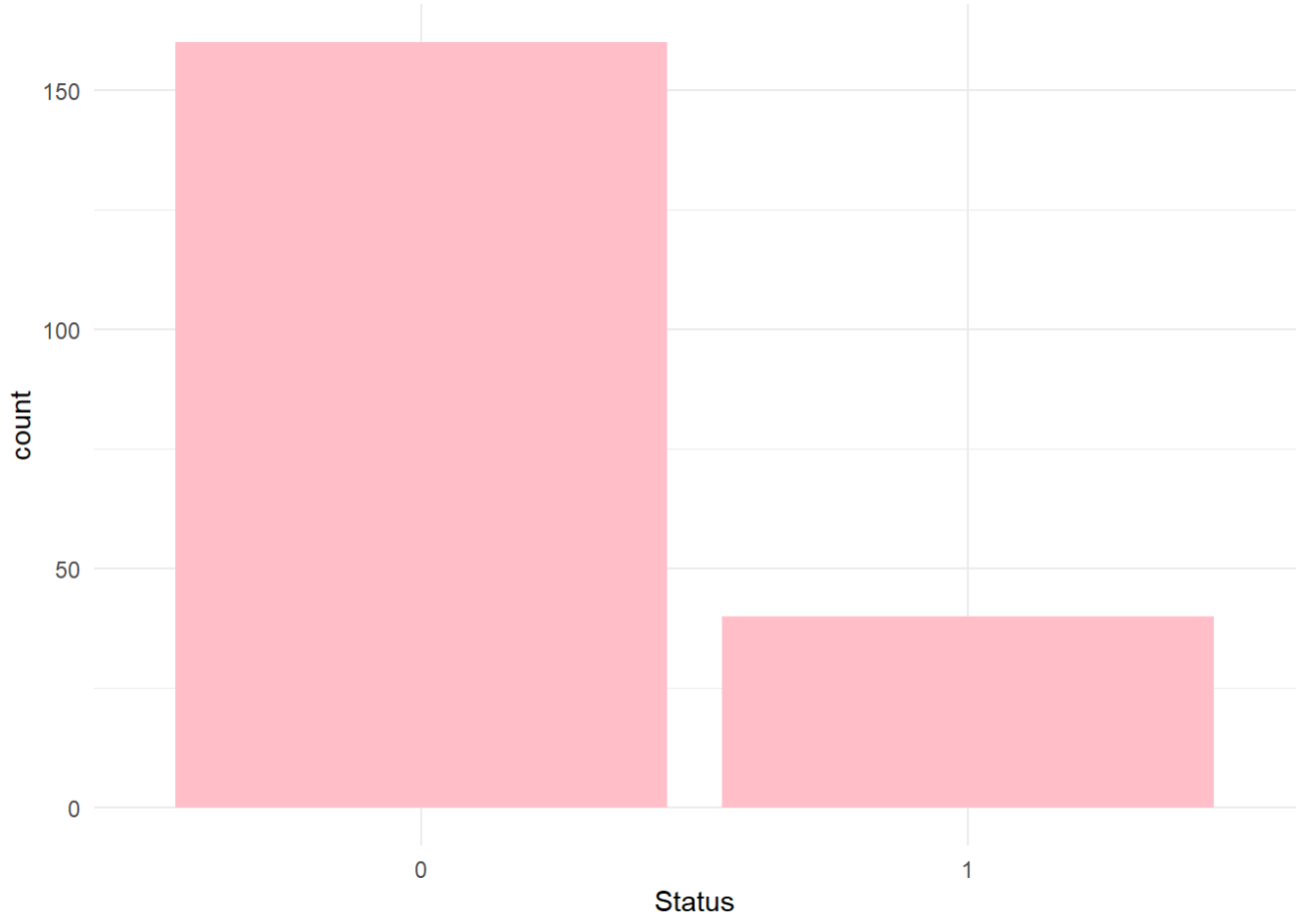
$$g(\pi_i) = \beta_1 + \beta_2 Age_i + \beta_3 D_{1i} + \beta_4 D_{2i},$$

where

$$D_{1i} = \begin{cases} 1, & \text{if } Causa_i = 1 \\ 0, & \text{otherwise} \end{cases} \qquad D_{2i} = \begin{cases} 1, & \text{if } Consciousness_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

Let's check the plot of our Dependent variable

```
ggplot(data=icu, aes(x=stato)) +
  geom_bar(fill="pink") + labs(x = "Status") +
  theme_minimal()
```



We can observe that our dependent variable is unbalanced. How try to solve this issue is beyond the scope of this course, for this reason we won't talk about unbalanced classes.

Our goal: We would like to use Generalized linear model to study the probability of death.

Grouped data

To make easier the transformation to grouped data, we can modify the variable *Age* as

$$D_{3i} = \begin{cases} 1, & \text{if } Age_i > 70 \\ 0, & \text{otherwise} \end{cases}$$

```
icu$cleta <- ifelse(icu$eta < 70, 0, 1)
```

Now, we need to find the number of deaths and alive for all the combinations of our explanatory variables and then build our new dataset.

```
table(icu$stato[which(icu$cleta==0 & icu$causa == 0 & icu$coscienza == 0)])
```

```
##
##  0  1
## 25  0
```

```
table(icu$stato[which(icu$cleta==0 & icu$causa == 0 & icu$coscienza == 1)])
```

```
##
## 0 1
## 6 0
```

```
table(icu$stato[which(icu$cleta==0 & icu$causa == 1 & icu$coscienza == 0)])
```

```
##
##  0  1
## 64  9
```

```
table(icu$stato[which(icu$cleta==0 & icu$causa == 1 & icu$coscienza == 1)])
```

```
##
##  0  1
## 19 13
```

```
table(icu$stato[which(icu$cleta==1 & icu$causa == 0 & icu$coscienza == 0)])
```

```
##
##  0  1
## 15  0
```

```
table(icu$stato[which(icu$cleta==1 & icu$causa == 0 & icu$coscienza == 1)])
```

```
##
## 0 1
## 5 5
```

```
table(icu$stato[which(icu$cleta==1 & icu$causa == 1 & icu$coscienza == 0)])
```

```
##
##  0  1
## 21  7
```

```
table(icu$stato[which(icu$cleta==1 & icu$causa == 1 & icu$coscienza == 1)])
```

```
##
## 0 1
## 5 6
```

```
ICU.binomial <- data.frame(cleta=c(0,0,0,0,1,1,1,1),
                           causa=c(0,0,1,1,0,0,1,1),
                           coscienza=c(0,1,0,1,0,1,0,1),
                           morti=c(0,0,9,13,0,5,7,6),
                           ni=c(25,6,73,32,15,10,28,11))
```

Logit model with grouped data

Assumptions:

- $Status_i \sim Binomial(\pi_i)$
- $\log(\frac{\pi_i}{1-\pi_i}) = \beta_1 + \beta_2 D_{3i} + \beta_3 D_{1i} + \beta_4 D_{2i}$

```
mod_glm <- glm(I(morti/ni) ~ cleta + causa + coscienza, family="binomial", weights = ni, data
=ICU.binomial)
summary(mod_glm)
```

```
##
## Call:
## glm(formula = I(morti/ni) ~ cleta + causa + coscienza, family = "binomial",
##     data = ICU.binomial, weights = ni)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.7845     0.6530  -5.796 6.81e-09 ***
## cleta         1.0489     0.4180   2.509  0.01210 *
## causa         1.6285     0.5696   2.859  0.00425 **
## coscienza     1.7955     0.3990   4.500 6.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.8993  on 7  degrees of freedom
## Residual deviance:  7.6479  on 4  degrees of freedom
## AIC: 32.681
##
## Number of Fisher Scoring iterations: 4
```

From the output, we can find the estimates of our regression coefficients: $\hat{\beta}_1 = -3.7845$, $\hat{\beta}_2 = 1.0489$, $\hat{\beta}_3 = 1.6285$ and $\hat{\beta}_4 = 1.7955$ and the standard errors: $SE(\hat{\beta}_1) = 0.6530$, $SE(\hat{\beta}_2) = 0.4180$, $SE(\hat{\beta}_3) = 0.5696$ and $SE(\hat{\beta}_4) = 0.3990$.

Test about significance

Let consider the generic system of hypothesis as

$$\begin{cases} H_0 : \beta_r = 0 \\ H_1 : \beta_r \neq 0 \end{cases}$$

where $r \in \{1, 2, 3, 4\}$.

The related test statistic corresponds to

$$Z_r = \frac{\hat{\beta}_r - \beta_r}{SE(\hat{\beta}_r)} \overset{H_0}{\sim} N(0, 1)$$

(and in this case β_r=0 under the null hypothesis)

Therefore the observed test statistics for each coefficient are: $z_1^{obs} = -5.796$, $z_2^{obs} = 2.509$, $z_3^{obs} = 2.859$ and $z_4^{obs} = 4.500$.

The related p-value corresponds to

$$\alpha_r^{obs} = \mathrm{P}_{H_0}(|Z_r| \geq |z_r^{obs}|),$$

and for each coefficient we obtained $\alpha_1^{obs} = 6.81e-09$, $\alpha_2^{obs} = 0.01210$, $\alpha_3^{obs} = 0.00425$ and $\alpha_4^{obs} = 6.79e-06$.

- We reject the null hypothesis $H_0 : \beta_1 = 0$, $H_0 : \beta_3 = 0$ and $H_0 : \beta_4 = 0$ at 1%, 5% and 10% significance levels.
- We reject the null hypothesis $H_0 : \beta_2 = 0$ at 5% and 10% significance levels.

The *null deviance* corresponds to the deviance of the null model and the *residual deviance* corresponds to the deviance of our model.

We know that the following relationship holds

$$D(null) = 2\{\tilde{l}(saturated) - \hat{l}(null)\}$$

and the degree of freedom of the null deviance corresponds to $n - p_0 = 8 - 1 = 7$ (The saturated model has n coefficients and the null model has 1 coefficient).

Instead, in the case of *residual deviance* we know

$$D(model) = 2\{\tilde{l}(saturated) - \hat{l}(model)\},$$

hence the degree of freedom of the residual deviance corresponds to $n - p = 8 - 4 = 4$.

The *residual deviance* is equal to 7.6479 and it is greater than $n - p = 4$, hence our model is not good enough.

ODDS RATIO

The odds ratio for the variable *Age* is

$$\frac{\left(\frac{\pi_i}{1-\pi_i}\,\middle|\, Age_i = x_0 + 1\right)}{\left(\frac{\pi_i}{1-\pi_i}\,\middle|\, Age_i = x_0\right)} = e^{\beta_2},$$

which is equal to

```
exp(coefficients(mod_glm)[2])
```

```
##    cleta
## 2.854587
```

The odds ratio for those in the highest age group (keeping constant the other explanatory variables), is 2.85 times that of those younger than 70 years. This means that age is a risk factor for death.

The odds ratio for the variable *Causa* is

$$\frac{\left(\frac{\pi_i}{1-\pi_i}\,\middle|\, Causa_i = x_0 + 1\right)}{\left(\frac{\pi_i}{1-\pi_i}\,\middle|\, Causa_i = x_0\right)} = e^{\beta_3},$$

which is equal to

```
exp(coefficients(mod_glm)[3])
```

```
##    causa
## 5.096361
```

The odds ratio of those who have an emergency ICU admission are about 5 times higher than those who have a planned admission, given the same age and consciousness. So even the variable *Causa* represents a risk factor for death.

TEST ABOUT THE OVERALL SIGNIFICANCE

Let consider the following system of hypothesis

$$\begin{cases} H_0 : \underset{-}{\beta_2} = \beta_3 = \beta_4 = 0 \\ H_1 : H_0 \end{cases}$$

We need to estimate the null model as follows

```
mod_0 <- glm(I(morti/ni) ~ 1, family="binomial", weights = ni, data=ICU.binomial)
summary(mod_0)
```

```
##
## Call:
## glm(formula = I(morti/ni) ~ 1, family = "binomial", data = ICU.binomial,
##     weights = ni)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3863     0.1768  -7.842 4.43e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.899  on 7  degrees of freedom
## Residual deviance: 41.899  on 7  degrees of freedom
## AIC: 60.933
##
## Number of Fisher Scoring iterations: 5
```

The test statistic can be written as

$$W = 2(\hat{l}(model) - \tilde{l}(null)) \overset{H_0}{\sim} \mathcal{X}_{p-1},$$

where the observed value is equal to

```
(W <- 2*(as.numeric(logLik(mod_glm)) - as.numeric(logLik(mod_0))))
```

```
## [1] 34.25145
```

Then, the pvalue

$$\alpha^{obs} = \mathrm{P}(W > w^{obs})$$

is equal to

```
1-pchisq(W,3)
```

```
## [1] 1.753225e-07
```

We can reject H_0 at 1% significance level.

Evaluating the predictions

Let assume that we consider a patient to be “death” when the estimated probability is greater than 0.5.

```
(predicted <- ifelse(as.numeric(mod_glm$fitted.values) >= 0.5, ICU.binomial$ni, 0))
```

```
## [1]  0  0  0  0  0  0  0 11
```

```
ICU.binomial$morti
```

```
## [1]  0  0  9 13  0  5  7  6
```

Then, we can compute the quantities to build a confusion matrix (True Positive, False Positive, True Negative, False Negative).

```
(true_positive<- sum(ICU.binomial$morti[predicted != 0]))
```

```
## [1] 6
```

```
(false_positive <- sum(ICU.binomial$ni[predicted != 0]) - sum(ICU.binomial$morti[predicted!= 0]))
```

```
## [1] 5
```

```
(true_negative <- 160 - false_positive)
```

```
## [1] 155
```

```
(false_negative <- 40 - true_positive)
```

```
## [1] 34
```

Hence, the confusion matrix corresponds to

Predicted \ True values	Dead	Alive
Dead	6	5
Alive	34	155

Accuracy:

```
(6 + 155)/(6+155+5+34)
```

```
## [1] 0.805
```

The value of accuracy is really high, around 80.5%. This means that overall our model is good enough. However, in this case, we are interested in assessing whether the model predicts both classes well. In particular, we would like to understand whether the model can be used to have a good prediction of the “positive” class (deaths).

For this purpose, we can evaluate the sensibility and the specificity.

Sensibility:

```
6/(6+34)
```

```
## [1] 0.15
```

Specificity:

```
155/(155+5)
```

```
## [1] 0.96875
```

The model predicts very well the negative class (alive), indeed the specificity is around 96.88%. However, the positive class is predicted correctly by only 15%. This problem can be related to the presence of unbalanced classes.

Poisson regression

```
crabs <- read.csv("Granchi.csv")
head(crabs)
```

```
##   Satellites Width Dark GoodSpine
## 1          8  28.3    0         0
## 2          0  22.5    1         0
## 3          9  26.0    0         1
## 4          0  24.8    1         0
## 5          4  26.0    1         0
## 6          0  23.8    0         0
```

Crabs Dataset contains data about 173 female crabs with the following variables:

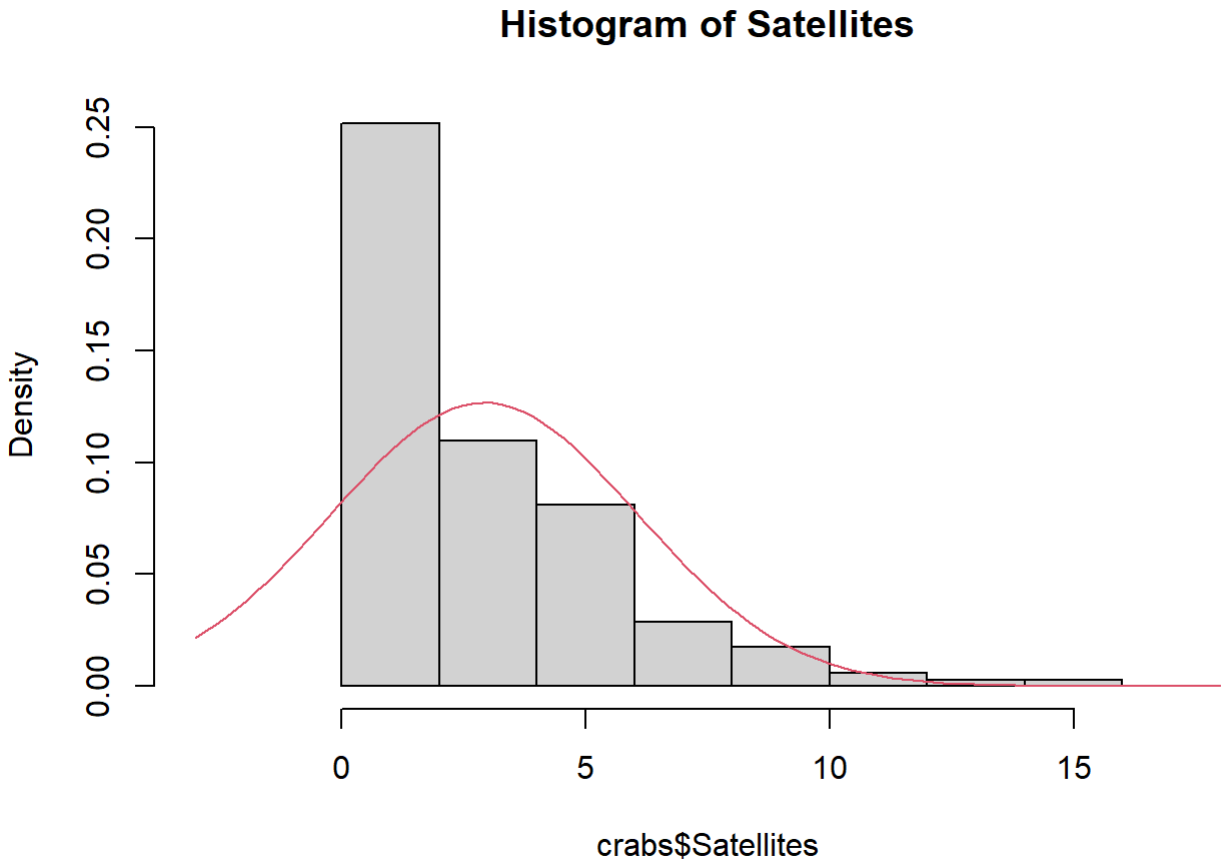
- *Satellites* refers to the number of male partners in addition to the primary partner
- *Width* is the width of the crab in centimeters
- *Dark* is a binary variable: 0 (no dark crab) and 1 (dark crab)
- *GoodSpine* refers to the crab shell defects: 0 (no) and 1 (yes)

Plot of the Dependent variable

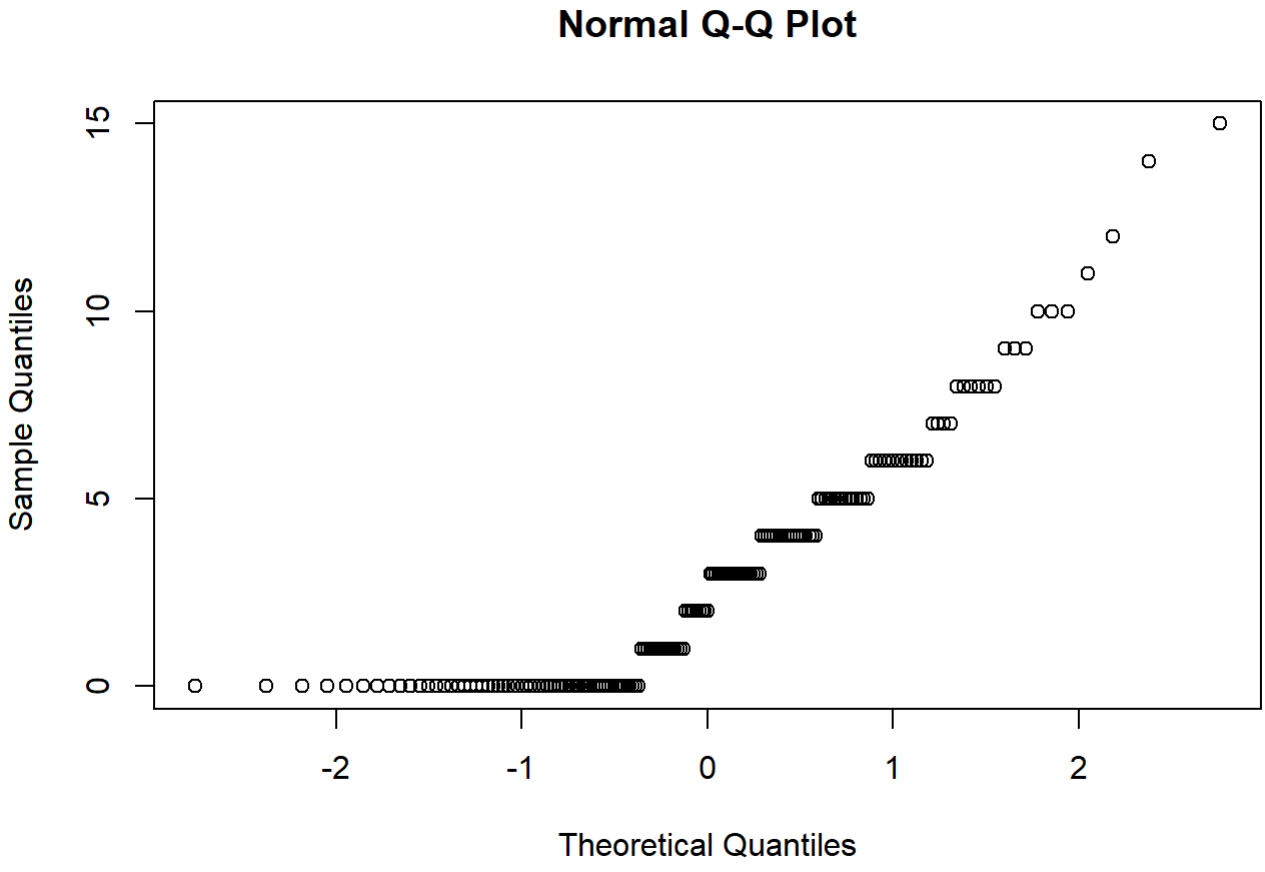
In this case, can the variable *Satellites* be normally distributed?

Let’s check

```
hist(crabs$Satellites,prob=T,xlim=c(-3,18), main= "Histogram of Satellites")
curve(dnorm(x,mean(crabs$Satellites),sd(crabs$Satellites)),col=2, add=T)
```



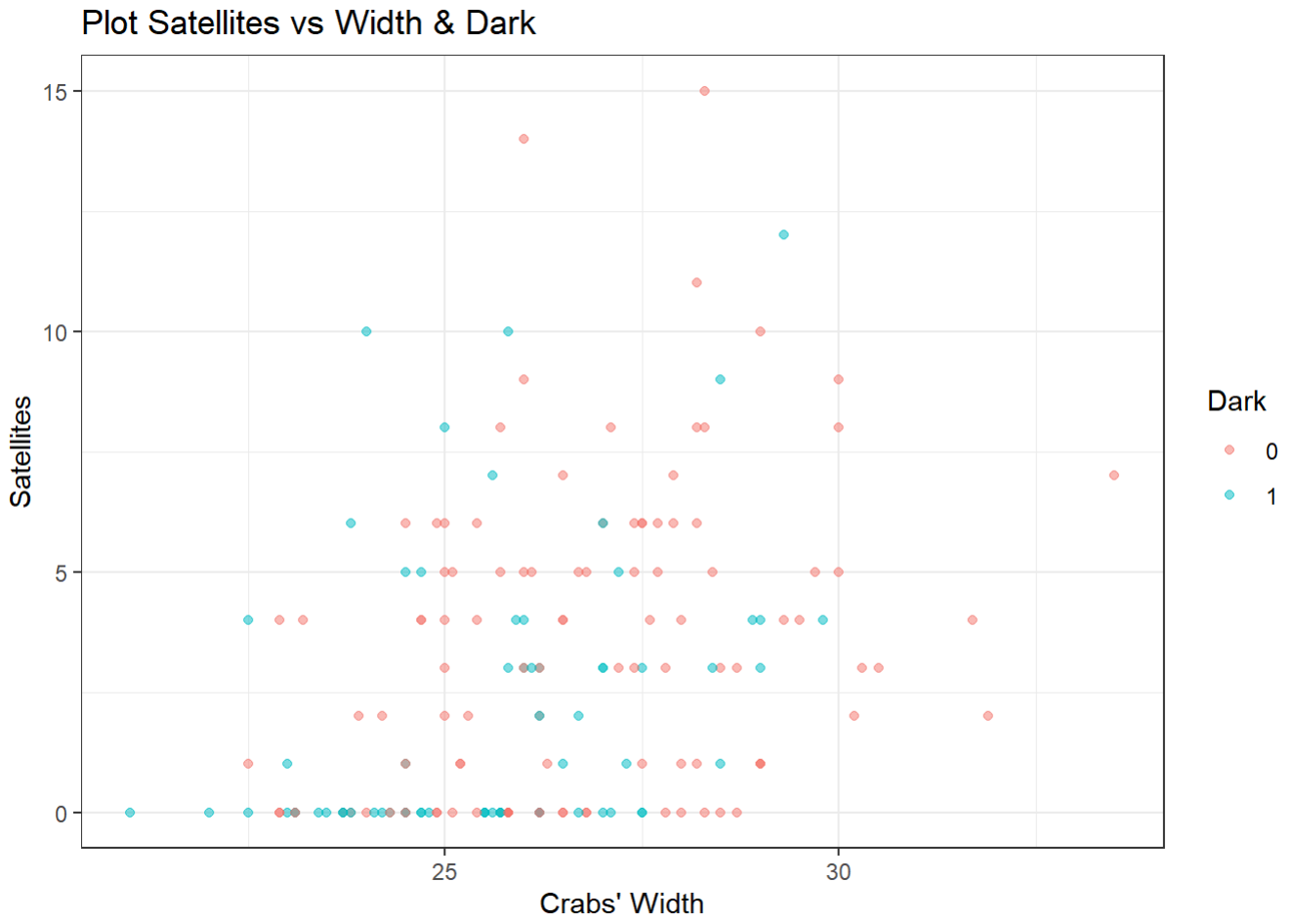

```
qqnorm(crabs$Satellites)
```



Satellites' support, due to the nature of the variable, is the non-negative integers instead of all reals. Looking at the plot, it does not seem to show a normal distribution (the histogram has obvious skewness and the q-q plot has a stepped shape), with a very anomalous trend on the left thing (due to the difference between the sample and theoretical support).

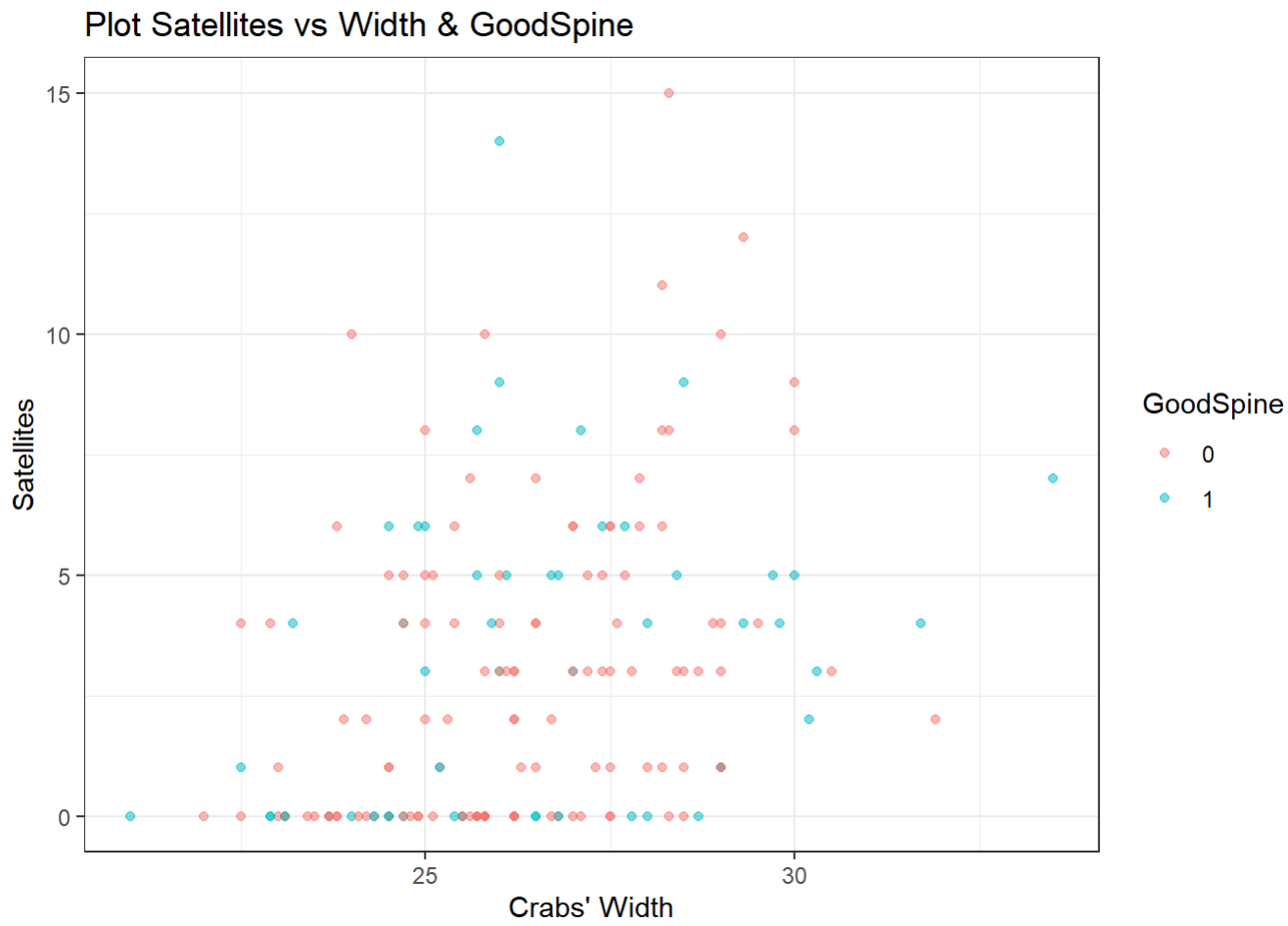
Data Exploration

```
theme_set(theme_bw())
ggplot(data = crabs, aes(x=Width, y=Satellites)) + geom_point( alpha = 0.5, aes(color= factor
(Dark))) + labs(x="Crabs' Width", y="Satellites", color="Dark", title = "Plot Satellites vs W
idth & Dark")
```



There does not seem to be much difference of the effect of Width on Satellites stratified by Dark. However, the plot is not very clear.

```
theme_set(theme_bw())
ggplot(data = crabs, aes(x=Width, y=Satellites)) + geom_point( alpha = 0.5, aes(color= factor
(GoodSpine))) + labs(x="Crabs' Width", y="Satellites", color="GoodSpine", title = "Plot Satel
lites vs Width & GoodSpine")
```



Even in this case, there does not seem to be much difference of the effect of Width on Satellites stratified by GoodSpine. However, the plot is not very clear.

Poisson model

Assumptions:

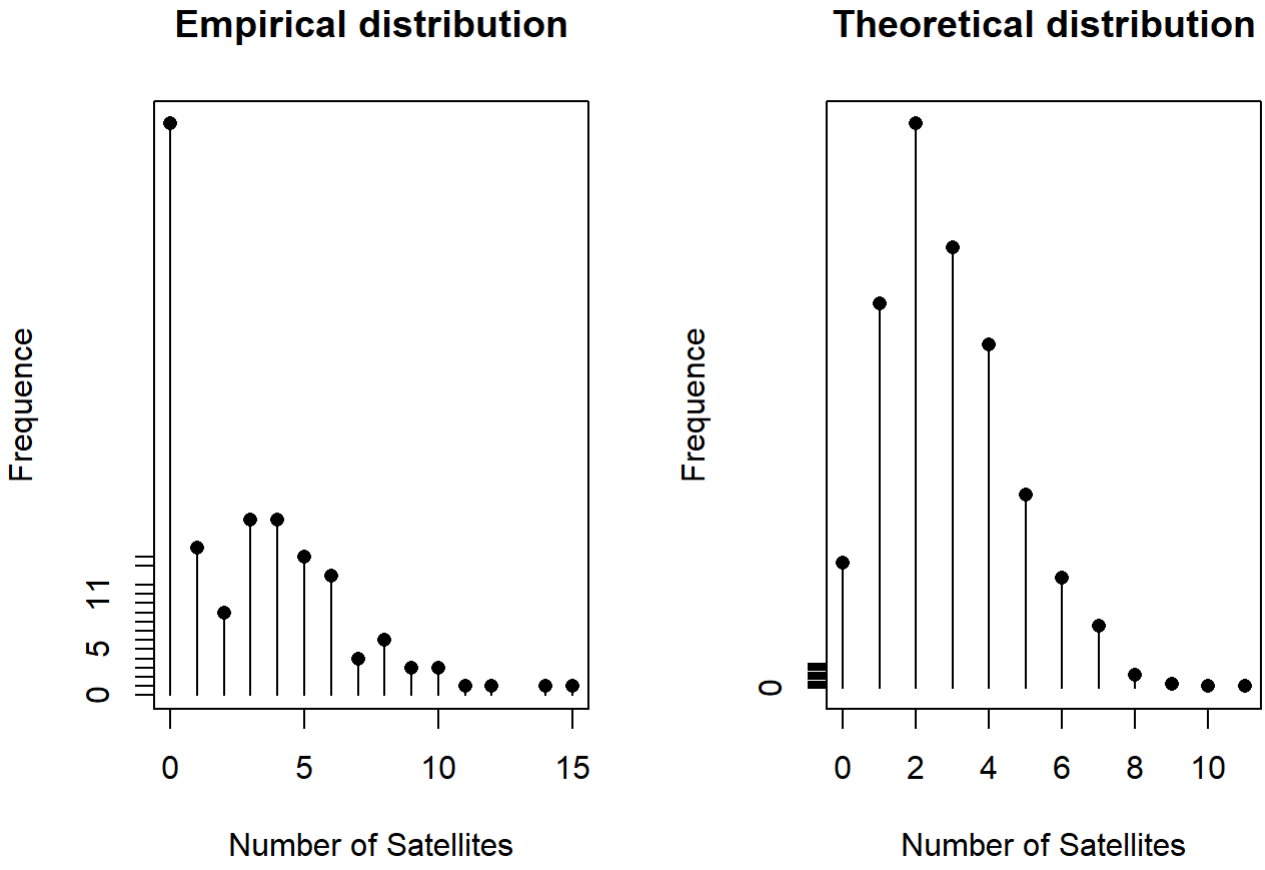
- $Satellites_i \overset{\perp\!\!\!\perp}{\sim} Poisson(\mu_i)$
- $log(\mu_i) = \beta_1 + \beta_2 Width_i + \beta_3 D_{1i} + \beta_4 D_{2i}$,

where

$$D_{1i} = \begin{cases} 1, & \text{if } Dark_i = 1 \\ 0, & \text{otherwise} \end{cases} \qquad D_{2i} = \begin{cases} 1, & \text{if } GoodSpine_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

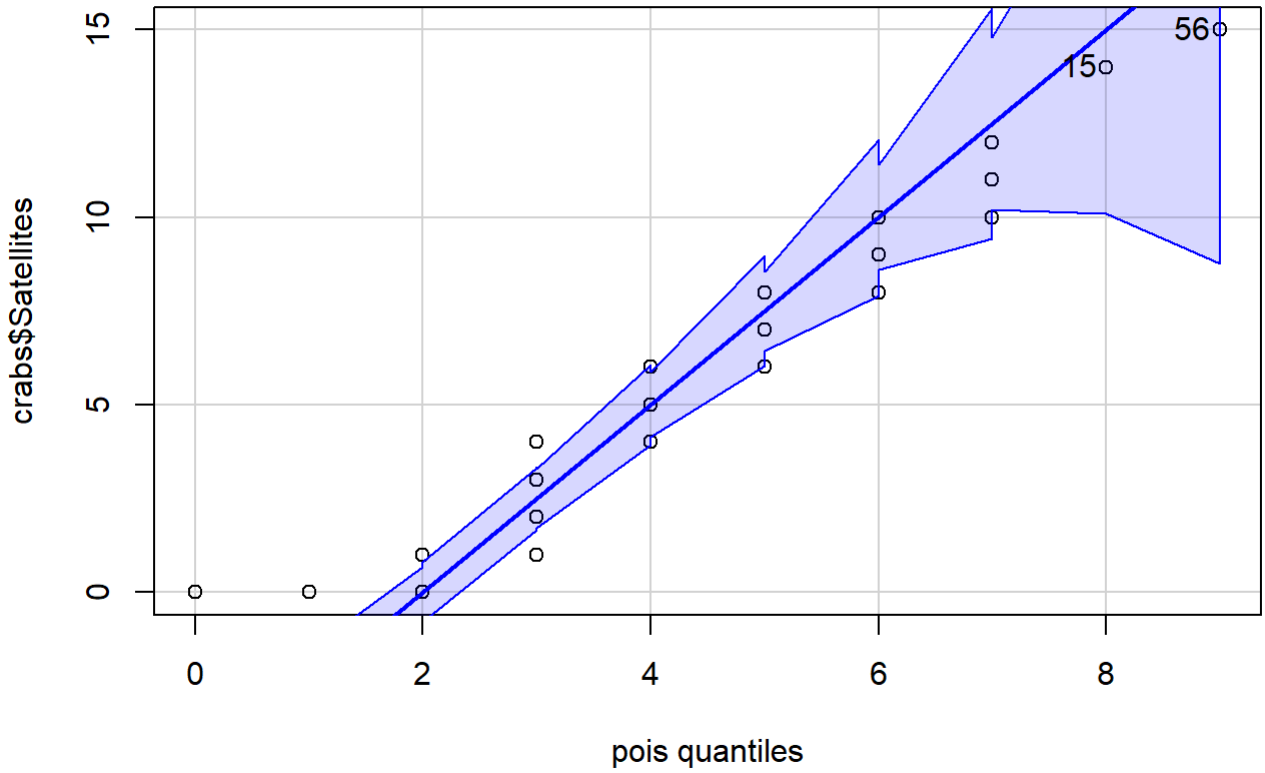
Let's check if the Poisson assumption can be reasonable

```
# Graficamente:
tab <- xtabs(~crabs$Satellites)
par(mfrow=c(1,2))
ascisse<-as.numeric(names(tab))

# Empirical distribution
plot(ascisse,tab,type="h",xlab="Number of Satellites",ylab="Frequence", main = "Empirical dis
tribution")
points(ascisse,tab,pch=16)

# Theoretical distribution
media<-mean(crabs$Satellites)
camp <-rpois(1000,media)
tab.camp<-xtabs(~camp)
ascisse.camp<-as.numeric(names(tab.camp))
plot(ascisse.camp,tab.camp,type="h",xlab="Number of Satellites",ylab="Frequence",main = "Theo
retical distribution")
points(ascisse.camp,tab.camp,pch=16)
```




```
par(mfrow=c(1,1))
```

```
qqPlot(crabs$Satellites,distribution="pois",lambda=mean(crabs$Satellites))
```




```
## [1] 56 15
```

Although most of the points are within the confidence bands in the qqplot, the empirical and theoretical distributions seem to differ considerably.

```
mod_glm <- glm(Satellites ~ Width + Dark + GoodSpine, family=poisson, data=crabs)
summary(mod_glm)
```

```
##
## Call:
## glm(formula = Satellites ~ Width + Dark + GoodSpine, family = poisson,
##     data = crabs)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.820088   0.570859  -4.940 7.81e-07 ***
## Width        0.149196   0.020753   7.189 6.52e-13 ***
## Dark        -0.265665   0.104972  -2.531   0.0114 *
## GoodSpine   -0.002041   0.097990  -0.021   0.9834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 560.96  on 169  degrees of freedom
## AIC: 924.25
##
## Number of Fisher Scoring iterations: 6
```

From the output, we can find the estimates of our regression coefficients: $\hat{\beta}_1 = -2.820088$, $\hat{\beta}_2 = 0.149196$, $\hat{\beta}_3 = -0.265665$ and $\hat{\beta}_4 = -0.002041$ and the standard errors: $SE(\hat{\beta}_1) = 0.570859$, $SE(\hat{\beta}_2) = 0.020753$, $SE(\hat{\beta}_3) = 0.104972$ and $SE(\hat{\beta}_4) = 0.097990$.

Interpretation of regression coefficients:

- Width:

```
exp(coefficients(mod_glm)[2])
```

```
##    Width
## 1.160901
```

If the Width of the crabs increases by one unit, the average of Satellites increases by 16% (keeping the other explanatory variables constant).

- Dark:

```
exp(coefficients(mod_glm)[3])
```

```
##      Dark
## 0.7666962
```

When the color of the crabs changes from no dark to dark, the change in the mean response given all other covariates held constant is ≈ 0.77, hence a decrease of 23% of the average number of male partners.

- GoodSpline:

```
exp(coefficients(mod_glm)[4])
```

```
## GoodSpine
## 0.9979615
```

When crabs shell changes from no defect to defect, the change in the mean response given all other covariates held constant is ≈ 1.

Test about significance

Let consider the generic system of hypothesis as

$$\begin{cases} H_0 : \beta_r = 0 \\ H_1 : \beta_r \neq 0 \end{cases}$$

where $r \in \{1, 2, 3, 4\}$.

The related test statistic corresponds to

$$Z_r = \frac{\hat{\beta}_r - \beta_r}{SE(\hat{\beta}_r)} \overset{H_0}{\sim} N(0, 1)$$

(and in this case β_r=0 under the null hypothesis)

Therefore the observed test statistics for each coefficient are: $z_1^{obs} = -4.940$, $z_2^{obs} = 7.189$, $z_3^{obs} = -2.531$ and $z_4^{obs} = -0.021$.

The related p-value corresponds to

$$\alpha_r^{obs} = \mathrm{P}_{H_0}(|Z_r| \geq |z_r^{obs}|),$$

and for each coefficient we obtained $\alpha_1^{obs} = 7.81e-07$, $\alpha_2^{obs} = 6.52e-13$, $\alpha_3^{obs} = 0.0114$ and $\alpha_4^{obs} = 0.9834$.

- We cannot reject the null hypothesis $H_0 : \beta_4 = 0$, this means the coefficient is not significant.
- We reject $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ at 1%, 5% and 10$ significance levels.
- We reject $H_0 : \beta_3 = 0$ at 5% and 10$ significance levels.

The *null deviance* corresponds to the deviance of the null model and the *residual deviance* corresponds to the deviance of our model.

We know that the following relationship holds

$$D(null) = 2\{\tilde{l}(saturated) - \hat{l}(null)\}$$

and the degree of freedom of the null deviance corresponds to $n - p_0 = 173 - 1 = 172$ (The saturated model has n coefficients and the null model has 1 coefficient).

Instead, in the case of *residual deviance* we know

$$D(model) = 2\{\tilde{l}(saturated) - \hat{l}(model)\},$$

hence the degree of freedom of the residual deviance corresponds to $n - p = 173 - 4 = 169$.

The *residual deviance* is equal to 560.96 and it is greater than $n - p = 169$, hence our model is not good enough.

TEST ABOUT THE OVERALL SIGNIFICANCE

Let consider the following system of hypothesis

$$\begin{cases} H_0 : \underline{\beta_2 = \beta_3 = \beta_4 = 0} \\ H_1 : H_0 \end{cases}$$

We need to estimate the null model as follows

```
mod_0 <- glm(Satellites ~ 1, family=poisson, data=crabs)
summary(mod_0)
```

```
##
## Call:
## glm(formula = Satellites ~ 1, family = poisson, data = crabs)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0713     0.0445   24.07   <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: 632.79  on 172  degrees of freedom
## Residual deviance: 632.79  on 172  degrees of freedom
## AIC: 990.09
##
## Number of Fisher Scoring iterations: 5
```

The test statistic can be written as

$$W = 2(\hat{l}(model) - \tilde{l}(null)) \overset{H_0}{\sim} \mathcal{X}_{p-1},$$

where $p - 1 = 3$ and the observed value is equal to

```
(W <- 2*(as.numeric(logLik(mod_glm)) - as.numeric(logLik(mod_0))))
```

```
## [1] 71.83453
```

Then, the pvalue

$$\alpha^{obs} = \mathrm{P}(W > w^{obs})$$

is equal to

```
1-pchisq(W,3)
```

```
## [1] 1.776357e-15
```

We can reject H_0 at 1% significance level. We can obtain the same result using the following:

```
anova(mod_0,mod_glm,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Satellites ~ 1
## Model 2: Satellites ~ Width + Dark + GoodSpine
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       172     632.79
## 2       169     560.96  3   71.835 1.727e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```