Chapter 13: Interactive Notebook for Students

Ram Gopal, Dan Philps, and Tillman Weyde

Summer 20202

Table of Contents

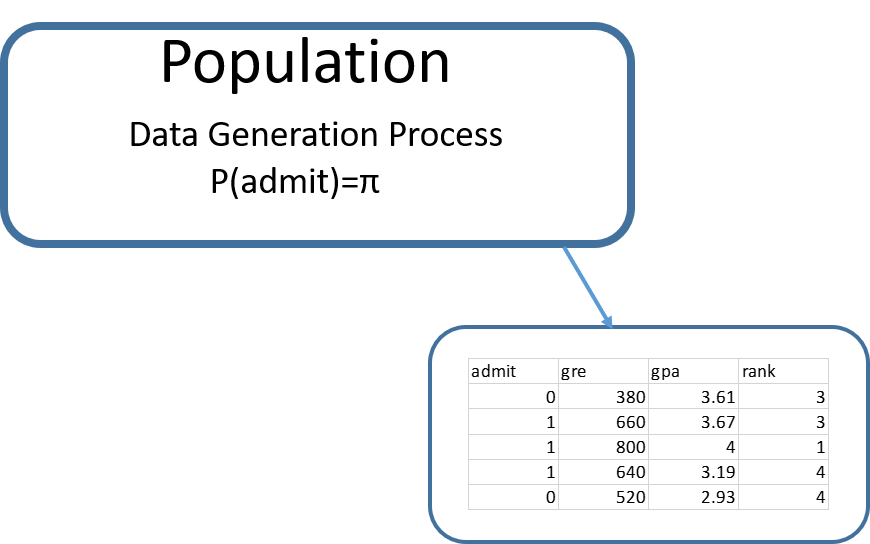
# Load packages

library(bbmle)

# Binary Outcomes

## Baseline Model

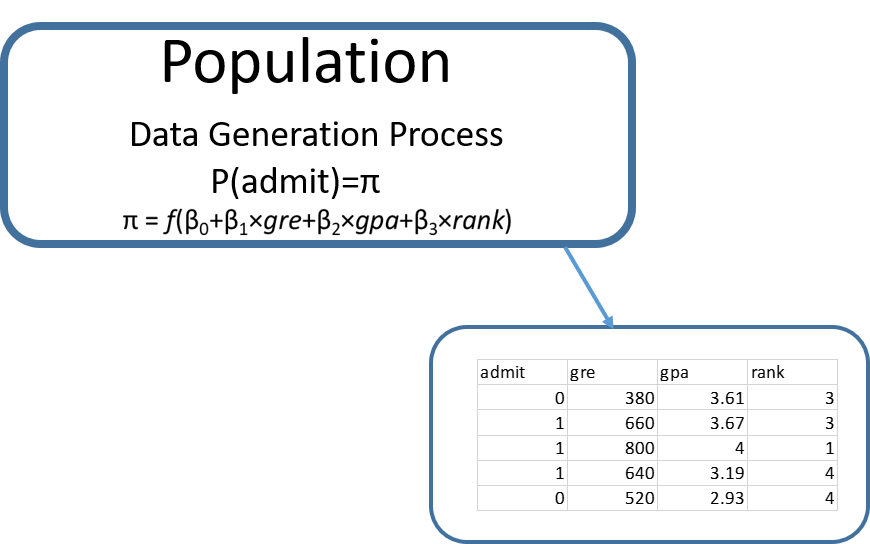
The baseline model that we want to estimate is shown in the following figure.



admit <- read.csv("../../data/admit.csv")  
LLbinary = function(pi){  
 p = ifelse(admit$admit == 1, pi, 1 - pi)  
 LL = sum(log(p))  
 return(-1\*LL)  
}  
  
res1 = mle2(minuslogl = LLbinary, start = list(pi= .5))  
summary(res1)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = LLbinary, start = list(pi = 0.5))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## pi 0.3175 0.0233 13.6 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 500

Now suppose we think that the probability of admission is determined by the variables gre, gpa, and rank. The data generation process is depicted in the following figure.



Let us denote

What should the function be that links X with ? The issue here is that the two terms have different domains. The domain for is . The domain for X on the other hand is . For the function that relates X to , we need to be able to match the two domains. The function is called the **link** function.

## Logistic Regression

The following table illustrates the creation of one possible link function.

| Variable | Domain |
| --- | --- |
| X |  |
|  |  |
|  | [0,1] |

The specification linking with X is the following.

Rearranging this gives us what is called the **logistic regression specification**.

In terms of the terminology, is the probability of **success**, is called the **odds**, and is the **log-odds**. The logistic regression specification essentially says that the log-odds are a linear function of the independent variables.

Now, let us write the code to estimate this.

LLbinary = function(b0,b1,b2,b3){  
 X = b0 + b1\*admit$gre + b2\*admit$gpa + b3\*admit$rank  
 pi = exp(X)/(1+(exp(X)))  
 p = ifelse(admit$admit == 1, pi, 1 - pi)  
 LL = sum(log(p))  
 return(-1\*LL)  
}  
  
res2 = mle2(minuslogl = LLbinary, start = list(b0 = 0, b1 = 0, b2 = 0, b3 = 0))  
summary(res2)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = LLbinary, start = list(b0 = 0, b1 = 0, b2 = 0,   
## b3 = 0))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## b0 -3.34596 1.13080 -2.96 0.0031 \*\*   
## b1 0.00169 0.00109 1.55 0.1204   
## b2 0.88048 0.32901 2.68 0.0074 \*\*   
## b3 -0.59679 0.12797 -4.66 0.0000031 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 459.8

The estimated logistic regression is the following.

The results indicate that gre is not statistically significant. A higher gpa increases the log-odds of admission, and higher rank lowers the log-odds of admission.

Let us try to understand the marginal effects. A unit increase in gpa increases the log-odds by 0.88. To understand this further, we can write

Rearranging this, we get

Thus, we can say that a unit increase in gpa increases the odds of admission by a factor of 2.41.

## Probit Regression

Another way to create the link function between X and is to use the following.

where is a cumulative probability of a standard normal distribution. This is called a probit regression. Let us code this.

LLbinary = function(b0,b1,b2,b3){  
 X = b0 + b1\*admit$gre + b2\*admit$gpa + b3\*admit$rank  
 pi = pnorm(X)  
 p = ifelse(admit$admit == 1, pi, 1 - pi)  
 LL = sum(log(p))  
 return(-1\*LL)  
}  
  
res3 = mle2(minuslogl = LLbinary, start = list(b0 = 0, b1 = 0, b2 = 0, b3 = 0))  
summary(res3)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = LLbinary, start = list(b0 = 0, b1 = 0, b2 = 0,   
## b3 = 0))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## b0 -2.084272 0.671707 -3.10 0.0019 \*\*   
## b1 0.000930 0.000645 1.44 0.1494   
## b2 0.558801 0.194190 2.88 0.0040 \*\*   
## b3 -0.352533 0.074458 -4.73 0.0000022 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 460.1

You will see that the results are pretty similar to the logistic regression. The coefficients in a probit regression are a bit harder to interpret, and therefore the logistic regression is more commonly used.

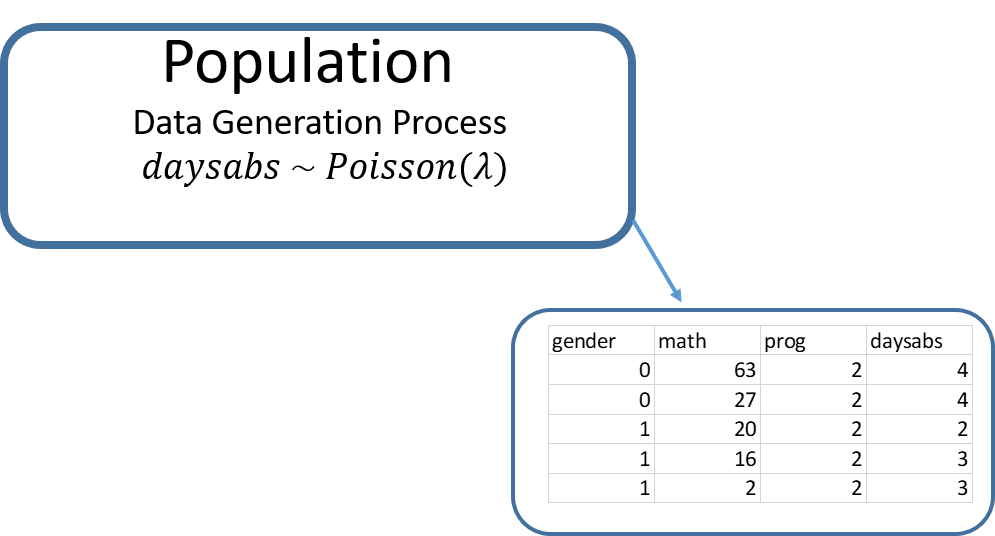
# Count Outcomes

Let us read the file student.csv.

student <- read.csv("../../data/student.csv")  
head(student)

## id gender math prog daysabs  
## 1 1001 0 63 2 4  
## 2 1002 0 27 2 4  
## 3 1003 1 20 2 2  
## 4 1004 1 16 2 3  
## 5 1005 1 2 2 3  
## 6 1006 1 71 2 13

## Baseline Model

The outcome of interest is the variable daysabs, the number of days a student was absent. The baseline model is depicted below. 

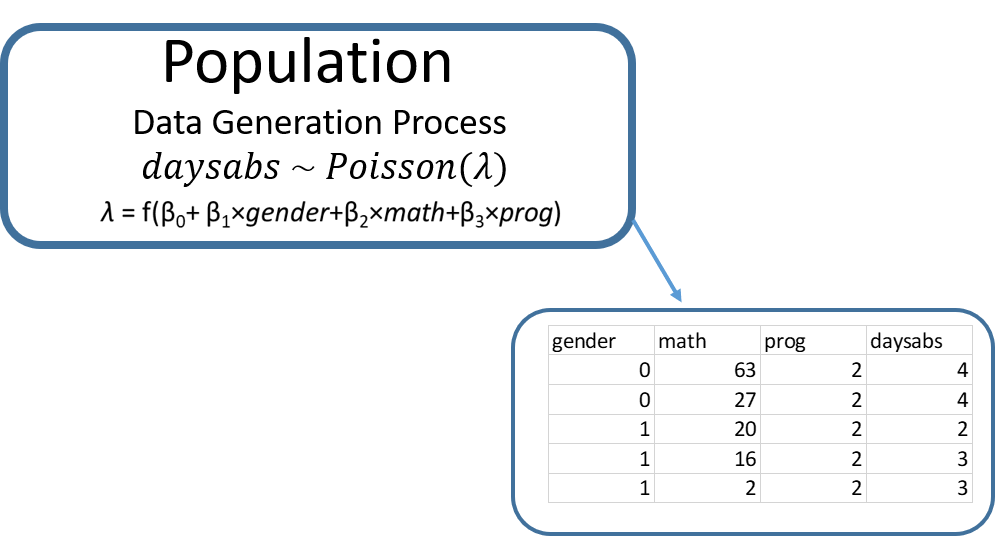
The following estimates the population parameter .

LLpois = function(lam){  
 p = dpois(x = student$daysabs, lambda = lam)  
 LL = sum(log(p))  
 return(-1\*LL)  
}  
  
res4 = mle2(minuslogl = LLpois, start = list(lam=10))  
summary(res4)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = LLpois, start = list(lam = 10))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## lam 5.955 0.138 43.2 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 3101

## Poisson Regression

Now suppose we think that the number of days absent is influenced by the variables gender, math, and prog. The data generation process is depicted in the following figure.



As the domain of is the link function is specified as . The following estimates the Poisson regression.

LLpois = function(b0, b1, b2, b3){  
 X = b0 + b1\*student$gender + b2\*student$math + b3\*student$prog  
 lam = exp(X)  
 p = dpois(x = student$daysabs, lambda = lam)  
 LL = sum(log(p))  
 return(-1\*LL)  
}  
  
res5 = mle2(minuslogl = LLpois, start = list(b0 = log(5.96), b1 = 0, b2 = 0, b3 = 0))  
summary(res5)

## Maximum likelihood estimation  
##   
## Call:  
## mle2(minuslogl = LLpois, start = list(b0 = log(5.96), b1 = 0,   
## b2 = 0, b3 = 0))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(z)   
## b0 3.255504 0.081341 40.02 < 0.0000000000000002 \*\*\*  
## b1 0.235355 0.046742 5.04 0.00000048 \*\*\*  
## b2 -0.007665 0.000923 -8.30 < 0.0000000000000002 \*\*\*  
## b3 -0.606762 0.036192 -16.77 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## -2 log L: 2649

The Poisson regression equation is

All variables are statistically significant. Let us understand the marginal effects. The coefficient for gender is 0.24. Gender is coded as 0 for females and 1 for males.

With some alegbra we get,

and thus

This means that the average number of days a male student is absent is 1.27 times that of a female student.

# Model Fit

With linear regression is used as a measure of the models performance. The same concept does not translate to binary and count models.

For a binary regression, a few metrics are useful to assess the model fit. The first is the likelihood, which is normally reported as -2LL. Note that we want this to be smaller. For the baseline model the value of -2LL was 499.9765. For the logistic regression, it is 459.8384. Hence, the regression model is better than the baseline model. Two other metrics that are commonly used are AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria). There are defined as:

Where is the number of parameters to be estimated and is the number of observations.

Let us compute these for the Poisson model.

| Metric | Baseline Model | Poisson Regression |
| --- | --- | --- |
|  | 3101.018 | 2648.787 |
|  | 3103.018 (k = 1) | 2656.787 (k = 4) |
|  | 3112.517 | 2694.782 |

The metrics clearly indicate that the Poisson regression model is adding value over the baseline model.

# glm() function

The function to estimate a regression model with binary or count outcomes in R is glm(). The general syntax is:

glm(formula, family=familytype(link=linkfunction), data=)

| Family | Default Link Function |
| --- | --- |
| binomial | (link = “logit”) |
| gaussian | (link = “identity”) |
| Gamma | (link = “inverse”) |
| inverse.gaussian | (link = “1/mu^2”) |
| poisson | (link = “log”) |
| quasi | (link = “identity”, variance = “constant”) |
| quasibinomial | (link = “logit”) |
| quasipoisson | (link = “log”) |

The following code estimates the binary and count models with our data.

res6 = glm(admit~gre+gpa+rank,family="binomial",data=admit)  
summary(res6)

##   
## Call:  
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",   
## data = admit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.580 -0.885 -0.638 1.157 2.173   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.44955 1.13285 -3.05 0.0023 \*\*   
## gre 0.00229 0.00109 2.10 0.0356 \*   
## gpa 0.77701 0.32748 2.37 0.0177 \*   
## rank -0.56003 0.12714 -4.40 0.000011 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 499.98 on 399 degrees of freedom  
## Residual deviance: 459.44 on 396 degrees of freedom  
## AIC: 467.4  
##   
## Number of Fisher Scoring iterations: 4

res7 = glm(daysabs~gender+math+prog,family="poisson",data=student)  
summary(res7)

##   
## Call:  
## glm(formula = daysabs ~ gender + math + prog, family = "poisson",   
## data = student)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.073 -2.268 -0.970 0.781 7.292   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.254816 0.081352 40.01 < 0.0000000000000002 \*\*\*  
## gender 0.235542 0.046747 5.04 0.00000047 \*\*\*  
## math -0.007633 0.000923 -8.27 < 0.0000000000000002 \*\*\*  
## prog -0.607277 0.036192 -16.78 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 2217.7 on 313 degrees of freedom  
## Residual deviance: 1765.5 on 310 degrees of freedom  
## AIC: 2657  
##   
## Number of Fisher Scoring iterations: 5