Binomial Generalized Linear Models

Filippo Gambarota

University of Padova

2022/2023

Updated on 2023-04-26

Outline

Binomial GLM

Example: Passing the exam

We want to measure the impact of **watching tv-shows** on the probability of **passing the statistics exam**.

- `exam`: **passing the exam** (1 = "passed", 0 = "failed")
- `tv_shows`: **watching tv-shows regularly** (1 = "yes", 0 = "no")

```
head(dat)
```

```
##   tv_shows exam
## 1        1    1
## 2        1    1
## 3        1    0
## 4        1    1
## 5        1    0
## 6        1    1
```

Example: Passing the exam

We can create the **contingency table**

```
xtabs(~exam + tv_shows, data = dat) |>
    addmargins()
```

```
##      tv_shows
## exam    0   1 Sum
##   0    38  15  53
##   1    12  35  47
##   Sum  50  50 100
```

Example: Passing the exam

Each cell probability π_{ij} is computed as π_{ij}/n

```
(xtabs(~exam + tv_shows, data = dat)/n) |>
    addmargins()
```

```
##      tv_shows
## exam     0    1  Sum
##   0   0.38 0.15 0.53
##   1   0.12 0.35 0.47
##   Sum 0.50 0.50 1.00
```

Example: Passing the exam - Odds

The most common way to analyze a 2x2 contingency table is using the **odds ratio** (OR). Firsly let's define *the odds of success* as:

$$odds = \frac{\pi}{1 - \pi}$$
$$\pi = \frac{odds}{odds + 1}$$

- the **odds** are non-negative, ranging between 0 and $+\infty$
- an **odds** of e.g. 3 means that we expect 3 *success* for each *failure*

Example: Passing the exam - Odds

For the `exam` example:

```
odds <- function(p) p / (1 - p)
p11 <- mean(with(dat, exam[tv_shows == 1])) # passing exam | tv_shows
odds(p11)
```

```
## [1] 2.333333
```

Example: Passing the exam - Odds Ratio

The OR is a ratio of odds:

$$OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}$$

- OR ranges between 0 and $+\infty$. When $OR = 1$ the odds for the two conditions are equal
- An e.g. $OR = 3$ means that being in the condition at the numerator increase 3 times the odds of success

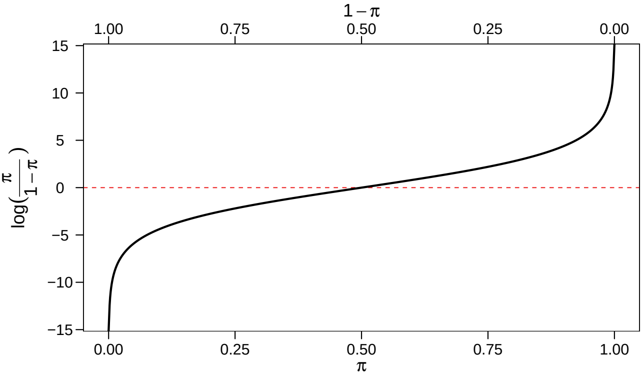
Example: Passing the exam - Odds Ratio

```
odds_ratio <- function(p1, p2) odds(p1) / odds(p2)
p11 <- mean(with(dat, exam[tv_shows == 1])) # passing exam | tv_shows
p10 <- mean(with(dat, exam[tv_shows == 0])) # passing exam | not tv_shows
odds_ratio(p11, p10)
```

```
## [1] 7.388889
```

Why using these measure?

The odds have an interesting property when taking the logarithm. We can express a probability π using a scale ranging $[-\infty, +\infty]$



Another example, **Teddy Child**

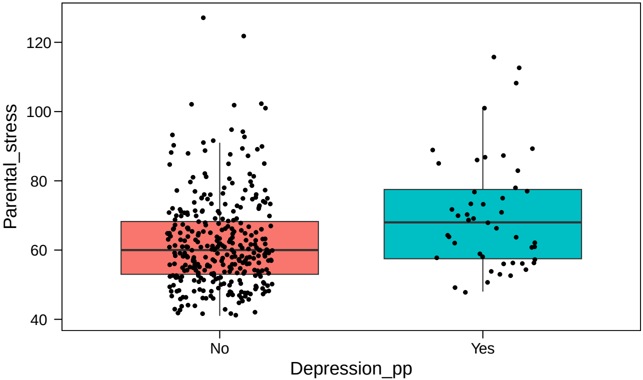
We considered a Study conducted by the University of Padua (TEDDY Child Study, 2020)[1]. Within the study, researchers asked the participants (mothers of a young child) about the presence of post-partum depression and measured the parental stress using the PSI-Parenting Stress Index.

ID	Parental.stress	Depression.pp
1	75	No
2	51	No
3	76	No
4	88	No
...	...	...
376	67	No
377	71	No
378	63	No
379	70	No

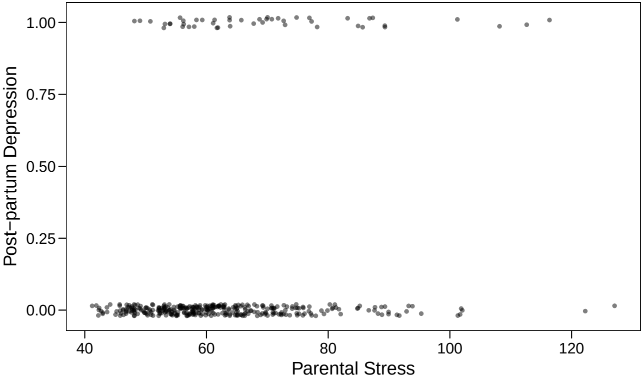
[1] Thanks to Prof. Paolo Girardi for the example, see https://teddychild.dpss.psy.unipd.it/ for information

Another example, **Teddy Child**

We want to see if the parental stress increase the probability of having post-partum depression:



Another example, **Teddy Child**



Another example, **Teddy Child**

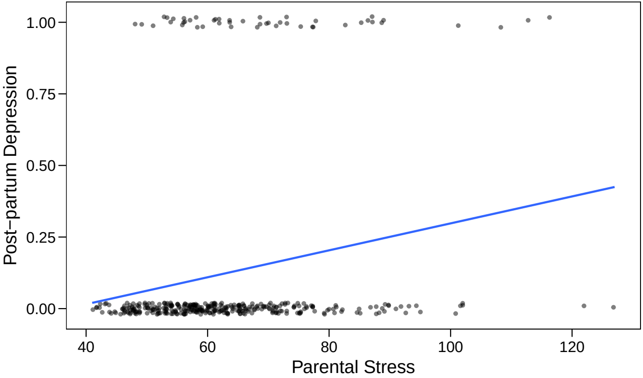
Let's start by fitting a linear model `Depression_pp ~ Parental_stress`. We consider "Yes" as 1 and "No" as 0.

```
fit_lm <- lm(Depression_pp01 ~ Parental_stress, data = teddy)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = Depression_pp01 ~ Parental_stress, data = teddy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42473 -0.13768 -0.10003 -0.05768  0.94702
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.172900   0.077561  -2.229 0.026389 *
## Parental_stress  0.004706   0.001201   3.919 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3239 on 377 degrees of freedom
## Multiple R-squared:  0.03915,    Adjusted R-squared:  0.0366
## F-statistic: 15.36 on 1 and 377 DF,  p-value: 0.0001054
```

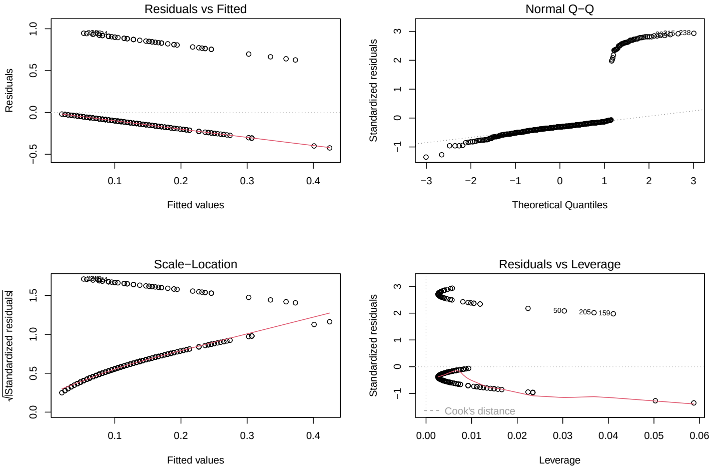
Another example, **Teddy Child**

Let's add the fitted line to our plot:



Another example, **Teddy Child**

… and check the residuals, pretty bad right?



Another example, **Teddy Child**

As for the exam example, we could compute a sort of contingency table despite the `Parental_stress` is a numerical variable by creating some discrete categories (just for exploratory analysis):

```
table(teddy$Depression_pp, teddy$Parental_stress_c) |>
    round(2)
```

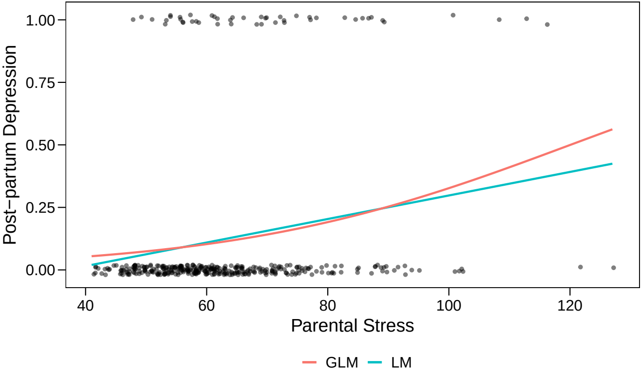
```
##
##        < 40 40-60 60-80 80-100 > 100
##   No      0   164   136     26     6
##   Yes     0    15    21      7     4
```

```
table(teddy$Depression_pp, teddy$Parental_stress_c) |>
    prop.table(margin = 2) |>
    round(2)
```

```
##
##        < 40 40-60 60-80 80-100 > 100
##   No         0.92  0.87   0.79  0.60
##   Yes        0.08  0.13   0.21  0.40
```

Another example, **Teddy Child**

Ideally, we could compute the increase in the odds of having the post-partum depression as the parental stress increase. In fact, as we are going to see, the Binomial GLM is able to estimate the non-linear increase in the probability.



Binomial GLM

- The **random component** of a Binomial GLM the binomial distribution with parameter π
- The **systematic component** is a linear combination of predictors and coefficients βX
- The **link function** is a function that map probabilities into the $[-\infty, +\infty]$ range.

Binomial GLM - Logit Link

The **logit** link is the most common link function when using a binomial GLM:

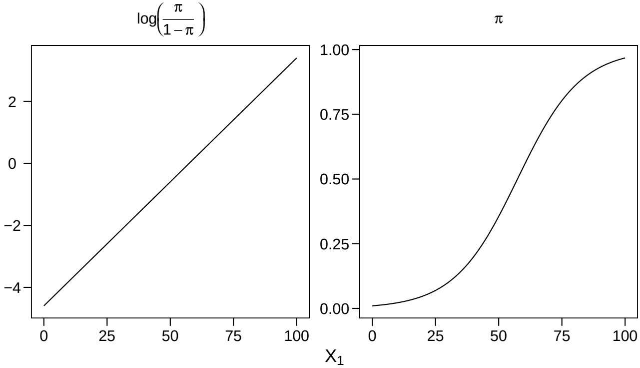
$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + ...\beta_p X_p$$

The inverse of the **logit** maps again the probability into the $[0, 1]$ range:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+...\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...\beta_p X_p}}$$

Binomial GLM - Logit Link

Thus with a single numerical predictor x the relationship between x and π in non-linear on the probability scale but linear on the logit scale.



Binomial GLM - Logit Link

The problem is that effects are non-linear, thus is more difficult to interpret and report model results