Generalized Linear Models

Introduction to Statistical Learning
Bachelor in *Global Governance*
University of Rome - Tor Vergata

Marco Stefanucci
Department of Economics and Finance
University of Rome - Tor Vergata
marco.stefanucci@uniroma2.it

Introduction

The linear model is often adeguate to describe the relation between a set of explanatory variables $x_1, \ldots, x_p$ and the response y.

There are cases, however, where the linear model is not a good solution.

Generalized Linear Models (GLM) overcome some of the limits of the linear model, namely implicit (or not) gaussian assumption and homoscedasticity.

Examples

Imagine that your variable of interest is the presence (or absence) of a disease as a function of, for example, age. The only possible values for Y are 1 or 0 (presence or absence). Even if we think a number as the probability of having such disease, any number outside the interval $(0, 1)$ does not make sense.

The linear model will produce predictions that are not constrained to be 0 or 1 and not even in the interval $(0, 1)$.

The same will happen if you want to model the number of customers entering a shop as a function of the hour. Predicted customers should be a positive number (more, an integer!) and the linear model does not ensure this will happen.

The technical problem

The problem is that we defined the linear model as a model for *observations* instead of parameters. The general formulation

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

is the same as $y_i = \mu_i + \varepsilon_i$ where the mean is equal to the linear predictor.

Unfortunately, the relation "observation = mean + random noise" does not apply if the data are not symmetric with unbounded range of variation.

We could rephrase the linear model as

$$E(y_i) = \theta_i$$
$$\theta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

We are now directly modeling the parameter of the distribution.

Introducing GLMs

The last formulation is amenable of generalizations such as

$$E(y_i) = f(\theta_i)$$

$$\theta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

where $f(\cdot)$ is a nonlinear function, in order to deal with other data structures such as presence/absence and positive data. However, the usual GLM formulation is

$$g(E(y_i)) = \theta_i$$

$$\theta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

where $g(\cdot) = f^{-1}(\cdot)$

Some examples

Different choices of $g(\cdot)$ lead to different models:

- $g(x) = \log(x)$, log-linear regression
- $g(x) = \text{logit}(x) = \log(\frac{x}{1-x})$, logit regression

These functions are chosen to "force" predictions to be in some interval.

The inverse function is even more important:

- $y = g(x) = \log(x)$, $x = f(y) = g^{-1}(y) = \exp(y)$
- $g(x) = \text{logit}(x) = \log(\frac{x}{1-x})$, $x = f(y) = g^{-1}(y) = \frac{\exp(y)}{1+\exp(y)}$

Practical understanding:

- The function $\exp(\theta_i)$ constraints the linear predictor to be positive.
- The function $\frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ constraints the linear predictor to lie in the interval $(0, 1)$.

Link with random variables

By introducing random variables, we can enstabilish a link between them and a nonlinear function $g(\cdot)$.

Presence/absence data are treated as Bernoulli random variables, with parameter p_i. In this case $g(x) = \log\frac{x}{1-x}$ is an appropriate transform.

Positive counts data are treated as Poisson random variables, with parameter λ_i. In this case $g(x) = \log(x)$ is an appropriate transform.

In the end, it is the parameter (or a nonlinear function of it) that is modeled in a linear way.

What is an appropriate transform for the gaussian random variable?

The variance

As a by product, we also obtain new expressions for the variance:

- Linear model: $V(y_i) = \sigma^2$
- Log-linear (Poisson) model: $V(y_i) = \lambda_i$
- Logit (Binomial) model: $V(y_i) = p_i(1 - p_i)$

If the variance depends on the observation we can remove the constant variance assumption of the linear model!

Model construction

In order to define a GLM we have to:

- specify distribution for the dependent variable y;
- specify a link function $g(\cdot)$;
- specify a linear predictor;
- a model for the variance of the outcome (usually) automatically follows, hence heteroscedasticity.

Estimation procedure

The estimation procedure is based on maximum likelihood: the likelihood function $\mathcal{L}(\theta, y)$ is maximized.

For most GLMs the likelihood equations are nonlinear functions of $\boldsymbol{\beta}$: we need an iterative method to solve nonlinear equations and determine the maximum of a likelihood function.

Two main (similar) algorithms are used: Newton-Raphson and Fisher scoring.

Model comparison and the Deviance

To test the significance of the model, or the superiority of a model with respect to another we will use the deviance.

Essentially, the deviance is the likelihood-ratio statistic for testing the null hypothesis that the model M_0 holds against the alternative that a more general model M_1 holds.

If we denote by $\hat{\theta}_0$ the vector of estimated parameters under model M_0 and by $\hat{\theta}_1$ the vector of estimated parameters under model M_1, the deviance can be computed as

$$D = -2 \log \left(\frac{\mathcal{L}(\hat{\theta}_0, y)}{\mathcal{L}(\hat{\theta}_1, y)} \right) = -2 \left[\ell(\hat{\theta}_0, y) - \ell(\hat{\theta}_1, y) \right]$$

This statistic is large when M_1 fits better compared to M_0.

Summary

A generalized linear model (GLM) is a flexible generalization of ordinary linear regression.

The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Logistic regression

For a binary response variable, representing the success and failure outcomes by 1 and 0, observation i has probabilities $P(y_i = 1) = p_i$ and $P(y_i = 0) = 1 - p_i$.

For this random variable, the link function for p_i is $g(p_i) = \log(p_i/(1 - p_i))$, called the logit.

GLMs using the logit link function are called logistic regression models and have the form

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \quad i = 1, 2, \ldots, n$$

Parameters interpretation

To simplify notation, we focus on the case of a single quantitative explanatory variable x. The model is logit$[P(y_i = 1)] = \beta_0 + \beta_1 x_i$, for which

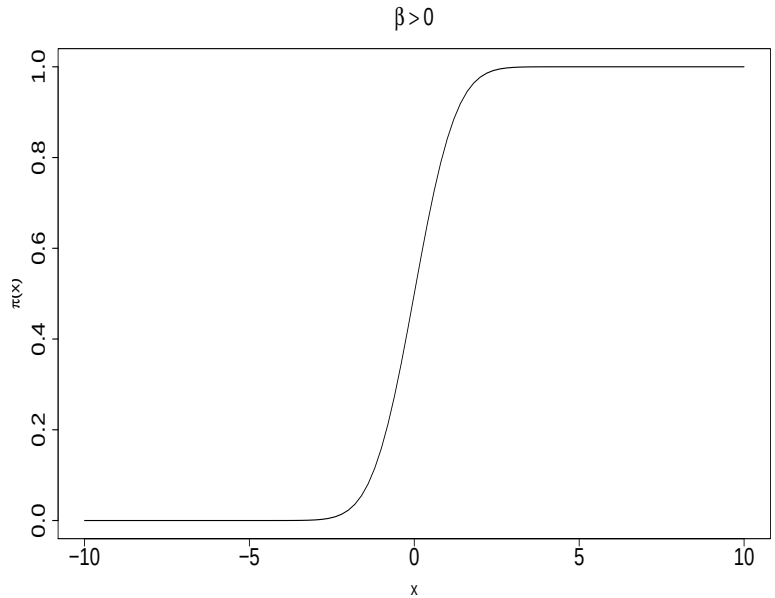
$$P(y_i = 1) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

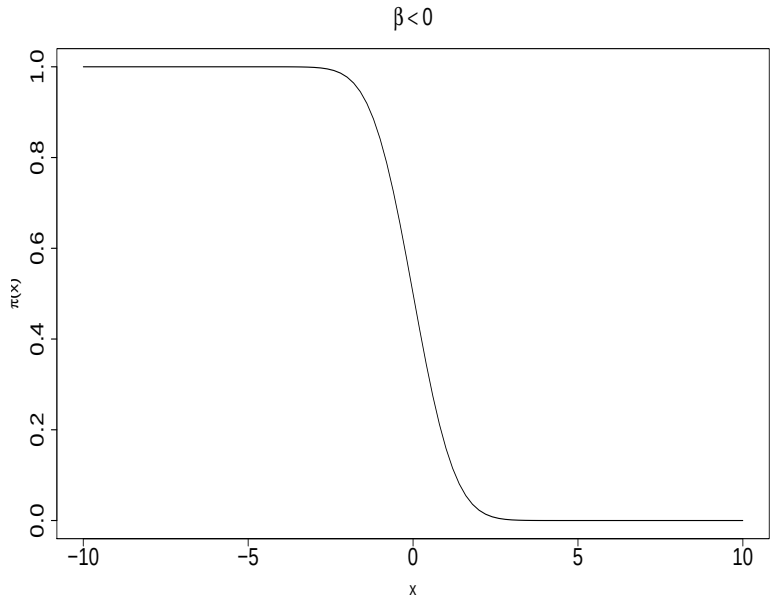
The curve for $P(y = 1)$ is monotone in x: When $\beta_1 > 0, P(y = 1)$ increases as x increases; when $\beta_1 < 0, P(y = 1)$ decreases as x increases.

When $\beta_1 = 0$, the logistic curve flattens to a horizontal line. As x changes, $P(y = 1)$ approaches 1 at the same rate that it approaches 0.

With multiple explanatory variables, $P(y = 1)$ is monotone in each explanatory variable according to the sign of its coefficient.

The rate of climb or descent increases as β_j increases. When $\beta_j = 0$, Y is conditionally independent of x_j, given the other explanatory variables.


β > 0
π(x)
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
x


β < 0
π(x)
0.0
0.2
0.4
0.6
0.8
1.0
−10
−5
0
5
10
x

Interpretation for β_1 uses the odds of success

$$\text{odds} = \frac{P(y=1)}{P(y=0)}$$

The odds can take any nonnegative value. With an odds of 3 we expect 3 successes for every failure; with an odds of $1/3$, we expect 1 success for every 3 failures.

The log of the odds is the logit, so odds have an exponential relationship with x. A 1-unit increase in x has a multiplicative impact of e^{β_1}: The odds at $x = u + 1$ equals the odds at $x = u$ multiplied by e^{β_1}.

Poisson regression

Many response variables have counts as their possible outcomes. Counts $\{y_i\}$ have positive means. It is common to model the logarithm of the mean through a GLM. The loglinear model is

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n$$

The Poisson loglinear model assumes that the counts are independent Poisson random variables.

For loglinear models, the mean satisfies the exponential relation

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})$$

Interpretation: The mean of y at $x_{ij} + 1$ equals the mean at x_{ij} multiplied by e^{β_j}, adjusting for the other explanatory variables.

Data example

Data refer to a sample of 5000 people with age under 65, out of a random sample of 100,000 people taken in 2015 from the Tuscany region of Italy, using administrative sources collected and organized by Istituto Nazionale di Statistica (Istat).

We model $Y =$ whether the subject is employed, in terms of being present in any administrative source ($1 =$ yes,$0 =$ no). Explanatory variables are:

- $G =$ gender ($1 =$ female, $0 =$ male)
- $I =$ whether an Italian citizen ($1 =$ yes, $0 =$ no)
- $P =$ whether receiving a pension($1 =$ yes, $0 =$ no)

Using indicator variables for the binary explanatory variables, the next output shows that the logistic regression model fit is

$$\log\left(\frac{p(y_i = 1)}{p(y_i = 0)}\right) = 0.286 - 0.639G_i + 0.766I_i - 1.910P_i$$

```
out <- glm(employed ~ female + italian + pension, family = "binomial", data = employ)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6443  -1.3568   0.7739   1.0080   2.1739

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.28620    0.08475   3.377 0.000733 ***
female      -0.63996    0.06139 -10.424  < 2e-16 ***
italian      0.76620    0.08573   8.938  < 2e-16 ***
pension     -1.91033    0.11108 -17.198  < 2e-16 ***

    Null deviance: 6716.1  on 4999  degrees of freedom
Residual deviance: 6223.9  on 4996  degrees of freedom
AIC: 6231.9

Number of Fisher Scoring iterations: 4
```

For instance, adjusting for whether an Italian citizen and whether receiving a pension, the odds that a woman is employed are estimated to be $e^{-0.639} = 0.53$ times the odds that a man is employed.

Data example 2

The City of Chicago data portal provides open access to many datasets produced by the city. One of the largest and most well-known is the Chicago Crime dataset.

This contains one row for any crime reported, by either the police or through public tips, within the city limits. Variables include the **address** of the reported crime, a **description** of the crime type, the **neighborhood name** and a **timestamp**.

We have grabbed a subset of the data from the first six months of 2017. Then, we grouped all reported crimes by crime type, hour, month and community area. Here are the first few rows of data.

head(ca)

```
 community_area hour month    type n  hourb
7               53     5     1 Robbery 0 (3,17]
40              49     5     1 Robbery 1 (3,17]
47               7     5     1 Robbery 0 (3,17]
61              58     5     1 Robbery 1 (3,17]
68              41     5     1 Robbery 0 (3,17]
77              18     5     1 Robbery 0 (3,17]
```

For each group we have counted the number of crimes that occurred in a given grouping. Our goal is to model the number of crimes as a function of the month, bucketed hour, and crime type.

The variable of interest in this task is a count value. It seems reasonable then to fit a generalized linear model using a Poisson distribution. The next output shows that the poisson regression model fit is

$$\log(\lambda_i) = 0.09 - 0.16H_i^{21-24} - 0.37H_i^{3-17} - 0.63H_i^{0-3} -$$
$$-0.41B_i - 0.66N_i - 0.51R_i$$

```
model <- glm(n ~ factor(hourb) + type, data = ca, family = poisson())
summary(model)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4858  -1.0622  -0.8827   0.4394   9.0973

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.09881    0.04691   2.106   0.0352 *
factor(hourb)(21,24] -0.16264    0.07186  -2.263   0.0236 *
factor(hourb)(3,17]  -0.37515    0.04700  -7.982 1.44e-15 ***
factor(hourb)[0,3]   -0.63672    0.06820  -9.337  < 2e-16 ***
typeBurglary         -0.41023    0.04984  -8.231  < 2e-16 ***
typeNarcotics        -0.66637    0.05407 -12.324  < 2e-16 ***
typeRobbery          -0.50868    0.05169  -9.841  < 2e-16 ***

    Null deviance: 7675.4  on 4999  degrees of freedom
Residual deviance: 7379.6  on 4993  degrees of freedom
AIC: 11105

Number of Fisher Scoring iterations: 6
```

The highest rate of crimes occur in the baseline hour bucket, 17:00 to 21:00, corresponding to the evening hours. Assault is the most frequent crime.