Afi Escuela de Finanzas

GLM: Generalized Linear Models

Máster en Data Science y Big Data

Javier Nogales – PhD Matemáticas
Catedrático, Estadística e IO, UC3M
fcojavier.nogales@uc3m.es 2022

Organization

- Subject organized in 2 topics
- 8 hours in total: 2 sessions
- Practical course: 50% basic concepts + 50% computer labs (using R)
- Evaluation: 100% final exercise

Objectives

- Learn how to extend the linear models when output/target is not normally distributed
- Extend the knowledge of linear models to a broader class of models
- Model binary and count data
- Handle the R language for GLM

Outline

Some Complementary References

- Alan Agresti. Foundations of Linear and Generalized Linear Models. Wiley, 2015.
- Myers, Montgomery, Vining, and Robinson. Generalized Linear Models with Applications in Engineering and the Sciences. Wiley, 2010.
- G. James, D. Witten, T. Hastie and R. Tibshirani. An Introduction to Statistical Learning with Applications in R

1. Introduction

Machine/Statistical Learning

- **Supervised learning**: predict or estimate an output (*response*) from various inputs (*predictors*)

  Statistical tools: better understanding about the relationship between the response and the predictors (inference)

  Machine learning tools: better (prediction accuracy)
  - Widely used tools: regression, classification, text analytics, recommendation systems, time series

- **Unsupervised learning**: tools for understanding data, with no target attribute (no labels). Usually organize into some natural groups

  Difficult to know how well your are doing; useful as a pre-processing step for supervised learning
  - Most widely used tools: PCA, clustering/segmentation, association rules

- **Reinforcement learning**: learn to make decisions based on a reward signal

  Hence, learn a set of actions or policies in order to maximize a expected reward
  - Related with optimization (dynamic programming under uncertainty)

- Others: semi-supervised learning, optimization, simulation, ...

Supervised Learning

- Take inputs $x_1, x_2, \ldots, x_k$ and map them to an output y
  - Statistical learning uses probability assumptions to find this map or function
  - Machine learning does not use probabilities, just the data
- To know whether the map is working correctly, we need some kind of metric to measure performance
  - Loss function: count number of times the model is working wrong (classification), or measure how close the prediction is to y (regression)
- The objective is to **minimize the loss**

Supervised Learning

Some notation:

- Output y: target, dependent variable, response, ...
- Input $x_1, x_2, \ldots, x_k$: predictors, features, regressors, covariates, independent variables, ...
- Data organized by rows: observations, samples, examples, instances, ...

One framework, but mainly two categories:

- Regression
- Classification

Supervised Learning Framework

- Usual framework in **Machine Learning** / **Statistics**:

$$\text{Data} = \text{Model} + \text{Noise}$$

- When we focus on one variable predicted by others: **supervised learning**

$$y = g(x_1, \ldots, x_k) + \text{Noise}$$

  - Statistical (linear) approximation: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$
  - Machine learning approximation: $y = \text{map}(x_1, \ldots, x_p) + \epsilon$

- Note the variables in the true model, g, may be different in the approximations

- Two main categories:
  - Classification: y is categorical
  - Regression: y is numeric (usually continuous)

Linear regression: a brief review

- The target, y, is a continuous variable that depends linearly on predictors
- Although the linear assumption may seem simplistic, it is extremely useful in practice
- Linear regression can deal with interactions and non-linearities
- Estimation is based on least squares: the loss function is the sum of residual squares (training prediction error)

Some questions answered by Regression Models

- What will the electricity price be in Spain tomorrow? (depending on demand, wind, gas, etc.)
- How can we determine the credit scoring of a bank client? (income, consumption, age, etc.)
- What will the fourth quarter sales be for a given company? (?)
- How many new followers will I get next week? (?)
- How can I determine the value of my home? (?)
- What is the beta of a stock? (?)

Note to answer previous questions we need additional information:
how those variables are related with another ones

Linear regression: a brief review

One of the main tools in Statistics and Machine Learning

On the basis of a dataset, some questions to answer:

- Are there relationships between one variable and others?
  Analyze evidence of association
- How strong are those relationships?
  Analyze the strength of association
- Can we use those relationships to predict better a variable?
- How accurate is that prediction?
- How to assess whether a model is valid? What to do if it is not valid?

Multiple Regression

- Remember the usual framework in **Machine Learning / Statistics**:

$$\text{Data} = \text{Model} + \text{Noise}$$

- When we focus on one variable predicted by others:

$$y = g(x_1, \ldots, x_k) + \text{Noise}$$

- Most widely used tool (approximation) with p variables:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

- This is the Multiple Linear Regression tool

- Note the variables in the true model, g, may be different in the approximation

Multiple Regression: Assumptions

- For the model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$
- Assumptions

1. $E(\epsilon_i) = 0 \quad \forall i$
2. $\text{Var}(\epsilon_i) = \sigma^2 \quad \forall i$
3. $E(\epsilon_i \epsilon_j) = 0 \quad \forall i \neq j$
4. $\epsilon_i \sim \text{Normal} \quad \forall i$

1. $E(y_i|x_i) = \beta_0 + \beta_1\ x_{i1} + \cdots + \beta_k\ x_{ip} \quad \forall i$
2. $\text{Var}(y_i|x_i) = \sigma^2 \quad \forall i$
3. $\text{Cov}(y_i, y_j) = 0 \quad \forall i \neq j$
4. $y_i|x_i \sim \text{Normal} \quad \forall i$

- Note (3) and (4) imply the observations are independent
- We do not need (4) for estimation, but for inference

What happens if some assumptions do not hold?

- In practice, the output may not be continuous
- The errors may not be normal
- The relation between the output and the predictors may not be linear

Generalized Linear Models (GLM): try to extend linear models by relaxing previous assumptions in an unified way

Non-linear regression: specify a parametric non-linear (in parameters) relation and estimate the parameters using an optimization solver

Regression in high-dimension: OLS has high variance in high dimension, advanced tools try to reduce the variance while increasing the bias

Non-parametric regression / machine learning: a non-linear function is estimated (learned) that cannot be parametrized

GLM: History

- **Linear Regression:** least-squares published by Legendre in 1805, and by Gauss in 1809

  Term regression coined by Galton in XIX century

  Mathematical assumptions by Fisher at the beginning of the XX century

- **Logistic Regression:** developed by Cox in 1958 for binary outputs

- **GLM:** formulated by Nelder and Wedderburn in 1972

- Hence, logistic regression was developed earlier than GLM, although it is a particular case

What is the meaning of linear?

- Are these models linear?

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 x_{i2}^2 + \beta_3 x_{i1} x_{i2} + \epsilon_i$$
$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- Are these models linear?

$$y_i = \beta_0 + \beta_1 x_{i1}^{\beta_2} + \epsilon_i$$
$$y_i = \frac{\beta_0}{1 + \frac{x_{i1}}{\beta_1^2}} + \beta_2 x_{i2} + \epsilon_i$$

2. Generalized Linear Models

Generalized Linear Models

A GLM has three components:

- A systematic linear predictor: $\eta_i = \beta' x_i$
- A random family describing the probability distribution of the output, y, with $E(y_i|x_i) = \mu_i$
- A link function relating the mean with the predictor: $g(\mu_i) = \eta_i$

Generalized Linear Models

Linear models (with normal data) are a particular case:

- The linear predictor matches the mean: $\eta_i = \mu_i$
- Hence, the link function is: $g(\mu_i) = \mu_i$
- The random component is $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$

But there are more models:

- Binary and multinomial data (binomial, multinomial)
- Count data (poisson, negative binomial)
- Survival/reliability data (exponential, gamma)

Exponential Family

- The random family can be any distribution, but in practice GLM works with the exponential family of distributions

- Exponential family distribution:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

  where the parameter of interest is θ, ϕ is a positive scale (dispersion) parameter, and b and c are arbitrary functions

- Usual distributions in the exponential family:
  Normal, Bernoulli, Binomial, Poisson, Exponential, Gamma, von Mises, etc.

Exponential Family

- θ is the distribution parameter that depends on the predictors through a linear function
- The conditional mean is: $E(y|x) = \mu = b'(\theta)$
- The conditional variance is: $\text{Var}(y|x) = \phi \cdot b''(\theta)$

Exponential Family: examples

Gaussian	$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\| x-\mu \|^2/(2\sigma^2)}$	$x \in \mathbb{R}$
Bernoulli	$p(x) = \alpha^x \left(1-\alpha\right)^{1-x}$	$x \in \{0,1\}$
Binomial	$p(x) = \binom{n}{x} \alpha^x \left(1-\alpha\right)^{n-x}$	$x \in \{0,1,2,\ldots,n\}$
Multinomial	$p(x) = \frac{n!}{x_1!x_2!\ldots x_n!} \prod_{i=1}^{n} \alpha_i^{x_i}$	$x_i \in \{0,1,2,\ldots,n\}\,,\ \sum_i x_i = n$
Exponential	$p(x) = \lambda\, e^{-\lambda x}$	$x \in \mathbb{R}^+$
Poisson	$p(x) = \frac{e^{-\lambda}}{x!}\, \lambda^x$	$x \in \{0,1,2,\ldots\}$
Dirichlet	$p(x) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1}$	$x_i \in [0,1]\,,\ \sum_i x_i = 1$

The link

- The dependence of the conditional mean $E(y_i|x_i) = \mu_i$ on the regressors is specified via the link function: $g(\mu_i) = \beta' x_i$
- The link g can be any function, but in practice g must be differentiable and monotone increasing (invertible)
- Examples:
  - identity: x
  - log: $\log(x)$
  - logit: $\log(x/(1-x))$
  - probit: $\Phi^{-1}(x)$ where Φ is the distribution function of a standard Normal
  - complementary log-log: $\log(-\log(1-x))$
- Usually, the choice of link function is made based on assumptions derived from physical knowledge or simple convenience

Canonical link

- Relates directly the output mean to the parameter θ of the exponential family
- That is, if $g(\mu) = \theta$, then g is the canonical link. Equivalently, $g = (b')^{-1}$
- Usual canonical links for the most used distributions:
  - Normal: $g(\mu) = \mu$
  - Binomial: $g(p) = \log(p/(1-p))$
  - Poisson: $g(\lambda) = \log(\lambda)$
  - Exponential: $g(\lambda) = 1/\lambda$
  - Gamma: $g(\lambda) = 1/\lambda$
- But other link functions can be used besides the canonical ones
  For instance, for the Poisson process we can also use the sqrt link

GLM vs Transformations

- When the assumptions of linear regression are not met, we can always try to transform the output to make it more linear respect to the predictors, or to make its variance more homogeneous, etc.

- So, when is GLM better?

- The link function transforms the mean, not the output

  Example: if y is a price and we take the logarithm, then we can model $E(\log(y)) = \beta' x$ (we are assuming y follows a log-normal distribution)

  On the other hand, with GLM we model $\log(E(y)) = \beta' x$, which avoids difficulties when $y = 0$

- Transforming the mean often allows the results to be more easily interpreted

- GLMs are more flexible: they allow for unequal variance, non-linear relationships, binary/categorical/count outcomes, skewed distributions, etc.

3. Estimation and Diagnosis

Estimation and Inference

- There is no close solution for the β's in the linear predictor, and inference is only approximate
- Estimation is based on MLE: the canonical link and the exponential family distribution simplifies the estimation
- MLE can be done through iteratively reweighted least squares (IRLS) algorithm
- The estimation is approximately normal: this allows us to construct tests and confidence intervals for β

Diagnosis

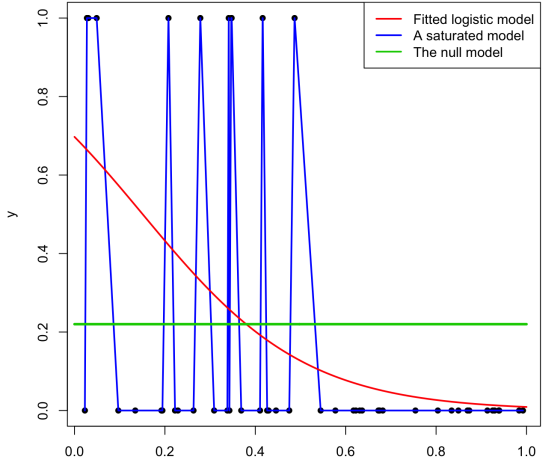
- How well does GLM fit the data?
- In linear models, diagnosis is based on residuals and R^2
- In GLM, different kind of residuals and R^2
- Pearson residuals: subtracting off the mean and dividing by the standard deviation

$$r_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i}$$

- Pearson statistic (sum of squares of Pearson residuals), $\sum_i r_i^2$, which follows a χ^2 distribution
- These residuals are less informative than in linear models

Diagnosis

- A better way is to compare the adequacy of a model is using a more general model
- The more general model is the saturated model: that with a maximum number of parameters to be estimated (one parameter per observation). It fits perfectly the sample



Diagnosis

- The comparison between the fitted model and the saturated one is made through the deviance:

$$D = -2 \log \left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}} \right]$$

- This is a likelihood ratio test, and $D \sim \chi^2$
- The deviance is then a measure of badness of fit: the smaller the better

Diagnosis

- Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean)
- Residual deviance corresponds to the fit of the complete model
- The Akaike Information Criterion (AIC) is basically te deviance penalized by number of parameters
- I.e. AIC penalizes complex models. The smaller AIC the better fit
- The AIC can be used to compare models (not necessarily nested)

4. Particular Cases

Poisson regression

- Used for count data where the random component is Poisson, $y_i \sim \mathcal{P}(\lambda_i)$
- Examples: Number of goals in a soccer match, daily number of calls in a call center, number of insurance claims in a month, number of crimes in a region, ...
- The canonical link is the log one, $g(\lambda) = \log(\lambda)$, but other links may be used
- Hence, the Poisson rate depends on the predictors as:

$$\log(\lambda_i) = \beta' \boldsymbol{x}_i$$
$$\lambda_i = \exp(\beta' \boldsymbol{x}_i)$$

- The canonical link assures $\lambda_i > 0$
- Offset: a predictor with a fixed coefficient (beta) of 1. Used to model rates instead of counts

Over-dispersion

- In a Poisson distribution, $E(y) = \text{Var}(y)$, or equivalently $\phi = 1$, but sometimes the variability is larger than the mean

- There are two ways to deal with over-dispersion:
  - Quasi-likelihood: modify the score to add a scale parameter, $\phi \neq 1$

    In R, this is called the quasi-Poisson family

    We cannot use likelihood-based tools like tests, deviance, etc.
  - Negative binomial distribution

Negative binomial distribution

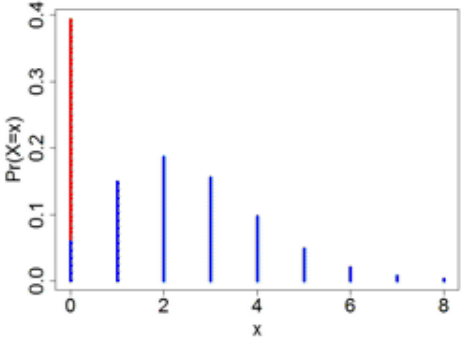
- The Negative Binomial distribution has mean λ and shape parameter θ
- Then,

$$E(y) = \lambda$$
$$\text{Var}(y) = \lambda + \lambda^2/\theta$$
$$\phi = 1$$

- That implies the variance is larger than the mean
- The main drawback is that the negative binomial distribution is not a member of the exponential family if θ is not known. In this case, estimation and inference is more difficult
- Special case: the Geometric distribution where $\theta = 1$

Zero-inflated models

- In count data, sometimes there are excess zeros: more zero observations than expected by the Poisson model



The blue 0's come from the Poisson distribution while the red ones are due to zero-inflation

Zero-inflated models

- Examples:
  number of physician-office visits per year (demand of medical care)
  number of days of absence of high-school juniors at schools per year (attendance behavior)
  number of children per family
  number of insurance claims within a population

- One way to deal with excess zeros is to use zero-inflated models

- They are mixture models that combine a count component and a point mass at zero

- Hence, they allow for two sources of zeros: zeros may come from both the point mass and from the count component

Zero-inflated models

- The probability of observing a zero count is inflated with probability π
- Equivalently, the distribution of y_i is

$$y_i \sim \begin{cases} 0 & \text{with probability } \pi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - \pi_i \end{cases}$$

- Hence, using the canonical log link for the Poisson count component, the regression equation for the mean is:

$$\lambda_i = \pi_i \cdot 0 + (1 - \pi_i) \exp(\beta' x_i)$$

- They are implemented in the pscl package for Poisson and Negative Binomial models

Logistic regression

- The output is usually binary: right/wrong, good/bad, diseased/healthy, etc.
- The Binomial distribution may be useful here: $\mathcal{B}(1, p_i)$
- The linear predictor is the same: $\eta_i = \beta' x_i$
- The canonical link is $g(p_i) = \log(p_i/(1 - p_i))$
- This link is called the logit and the GLM is called logistic regression
- But other links may be used: probit, complementary log-log link, etc.

Logistic regression

- As usual, estimation and inference is performed as in GLM
- Interpretation: the quantity $\frac{p}{1-p}$ is called the odds
  Hence, a $\beta = 1$ means the log of the odds that $y = 1$ occurs will go up by 1 after a change in the x by 1 (all the others variables fixed)
- Hence, more difficult to interpret...
- Logistic regression is one the most used techniques in practice
- We will see it in more detail in Advanced Regression
- It can be viewed as a regression tool or a classification one

Extensions: Generalized Additive Models (GAM)

- GLM: the predictor function is linear, i.e. $g(\mu) = \beta' x$
- GAM allows flexibility: $g(\mu) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$
- In GAM, the predictor function is additive but non-linear
- The predictor function is non-parametric (no betas) but smooth. The functions are usually estimated using splines
- In R, you can use the **gam** package

The end... Attained objectives (I hope...)

- Learn how to extend the linear models when output/target is not normally distributed
- Extend the knowledge of linear models to a broader class of models
- Model binary and count data
- Handle the R language for GLM

“[Statistics are] the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the science of man.”

Sir Francis Galton

Afi Escuela de Finanzas