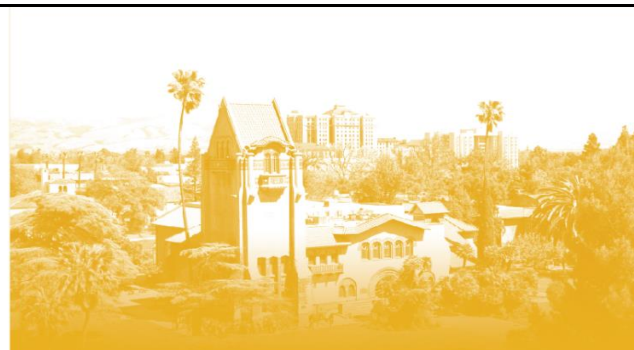


SJSU SAN JOSÉ STATE UNIVERSITY

Vector Data – Prediction
Generalized Linear Models (GLM)

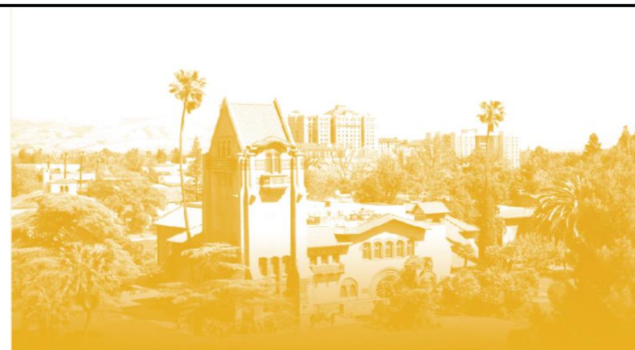
1

SJSU SAN JOSÉ STATE UNIVERSITY

Agenda

- Recap of Linear Regression Models & Basic Statistical Distributions
- Generalized Linear Model (GLM)
  - Background, Motivations and Assumptions
  - Basic Structure of GLMs
  - Types of GLMs
  - Formulation
  - Model Diagnostics
- Examples
- More Advanced Topics on GLM

2



SJSU SAN JOSÉ STATE UNIVERSITY

Reviews of Linear Regressions
and
Common Statistical Distributions

3

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Models Recap

Recall that the linear regression model assumes:

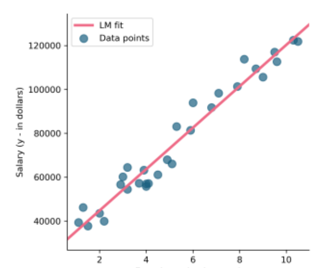
- Linear in parameters
- The response $Y|\boldsymbol{X}$ is continuous and normally distributed:

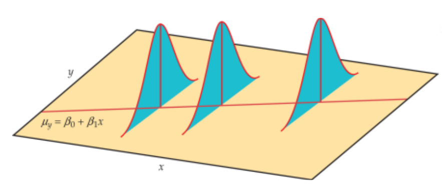
$$Y|\boldsymbol{X} \sim \boldsymbol{N}(\mu(\boldsymbol{X}), \sigma^2 \boldsymbol{I})$$

- The mean $\mu(\boldsymbol{Y})$ is simply given by:

$$E[Y|\boldsymbol{X}] = \mu(\boldsymbol{Y}) = \boldsymbol{X}^T \boldsymbol{\beta}$$

- Constant variance σ^2



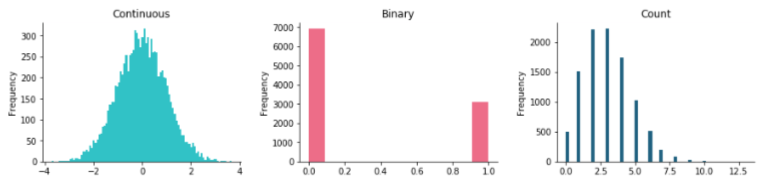


4

SJSU SAN JOSÉ STATE UNIVERSITY

Limitations of Linear Regression Models

- What if the response is not continuous (i.e. binary or count) ?

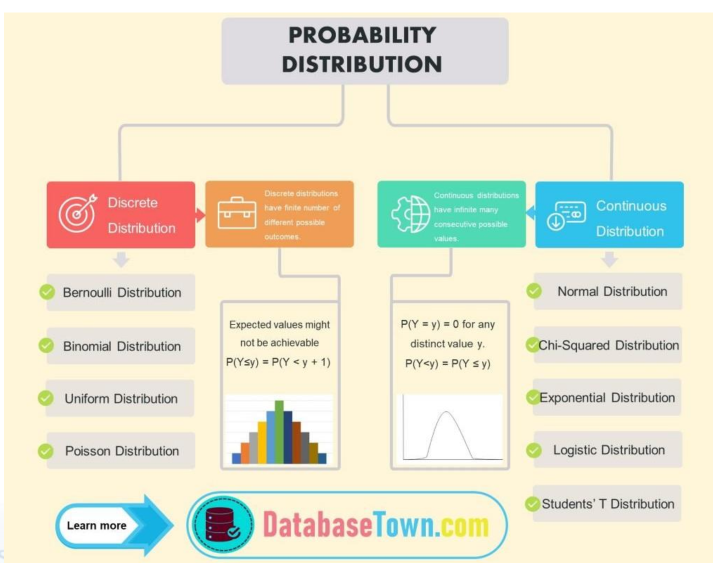


- What if the variance of Y isn't constant (i.e. depends on the mean) ?

5

SJSU SAN JOSÉ STATE UNIVERSITY

Review of Common Statistical Distributions

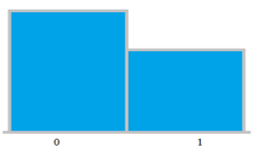


6

SJSU SAN JOSÉ STATE UNIVERSITY

Discrete Distributions

Bernoulli Distribution

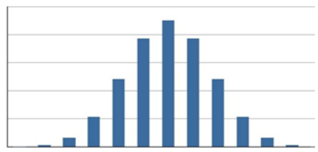


Y ~ Bern(p)

Characteristics
- Consists of a single trial
- 2 possible outcomes
- E(Y) = p
- Var(Y) = p × (1 – p)

Uses
- Guessing a single True/False question.

Binomial Distribution

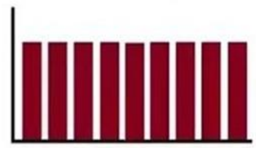


Y ~ B(n, p)

Characteristics
- Over the n trials, it measures the frequency of occurrence of one of the possible result.
- E(Y) = n × p
- P(Y = y) = C(y, n) × p^y × $(1 – p)^{n-y}$
- Var(Y) = n × p × (1 – p)

Uses
- How many heads obtained if a coin is flipped a coin n times.
- Predict an event occur over a series of trials

Uniform Distribution

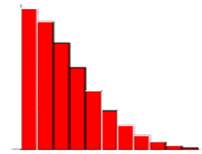


Y ~ U(a)

Characteristics
- All the outcomes are equally likely.
- All the bars are equally tall
- E(Y) & Var(Y) have no predictive power

Uses
- Results obtained after rolling a die
- Shuffling algorithms

Poisson Distribution



Y ~ Po(λ)

Characteristics of
- It measures the frequency over an interval of time or distance.
- E(Y) = λ
- P(Y = y) = λ^y / (λ! $e^{-\lambda}$)
- Var(Y) = λ

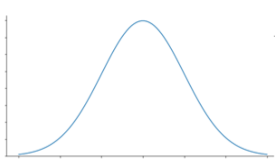
Uses
- Likelihood a certain event occur over a given interval of time or distance.
- Marketing analysis

7

SJSU SAN JOSÉ STATE UNIVERSITY

Continuous Distributions

Normal Distribution

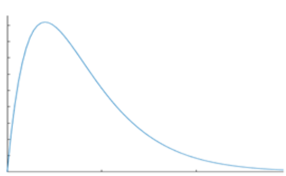


Y ~ N(μ, σ^2)

Characteristics
- Bell shaped w/ 68% within $\mu \pm \sigma$
- E(Y) = μ
- Var(Y) = σ^2

Uses
- Many…

Chi-Squared Distribution

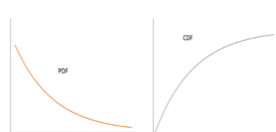


Y ~ X^2(k)

Characteristics
- Chi-Squared distribution is asymmetric and skewed to the right.
- Square of the t-distribution
- E(Y) = k
- Var(Y) = 2k

Uses
- Use to test wow of fit.
- x^2 – table

Exponential Distribution

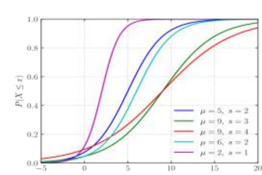


Y ~ Exp(λ)

Characteristics
- PDF & CDF plateau after a certain point
- E(Y) = 1/ λ
- Var(Y) = 1/ λ^2

Uses
- Use with dynamically changing variables, such as online websites traffic

Logistic Distribution



Y ~ Logistic(μ, s)

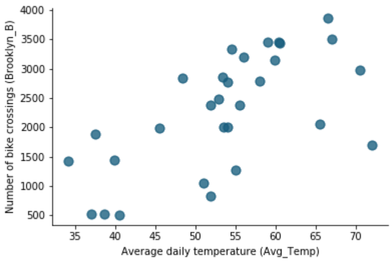
Characteristics of
- E(Y) = μ
- Var(Y) = s2*π^3/3

Uses
- Use to observe how continuous variable inputs can affect the probability of a binary result.

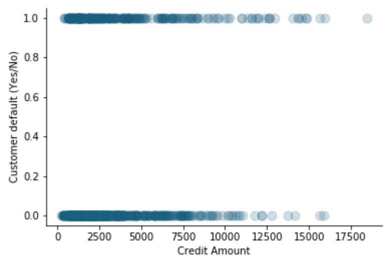
8

SJSU SAN JOSÉ STATE UNIVERSITY

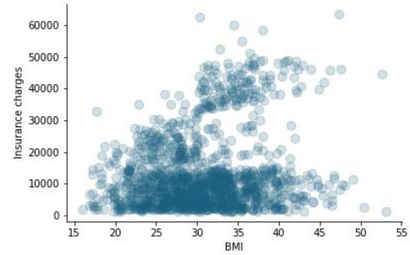
Data Type & Distribution Family?



Predict the number of bike crossings over the Brooklyn bridge in New York City given daily temperature



Predict the probability of default using Age of the customer

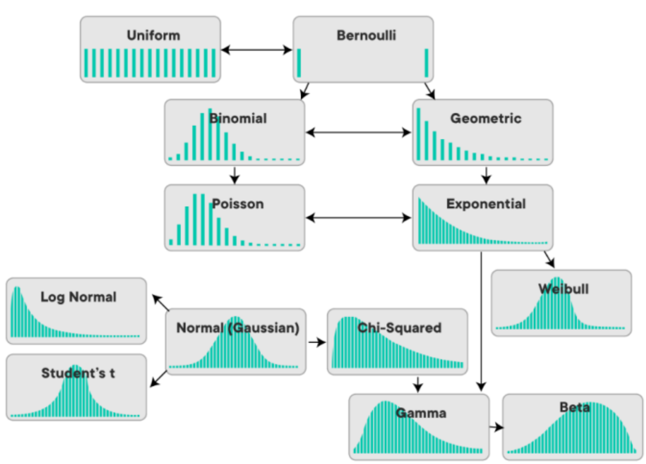


Predict insurance costs given the BMI of the individual.

9

SJSU SAN JOSÉ STATE UNIVERSITY

Different Statistical Distributions…

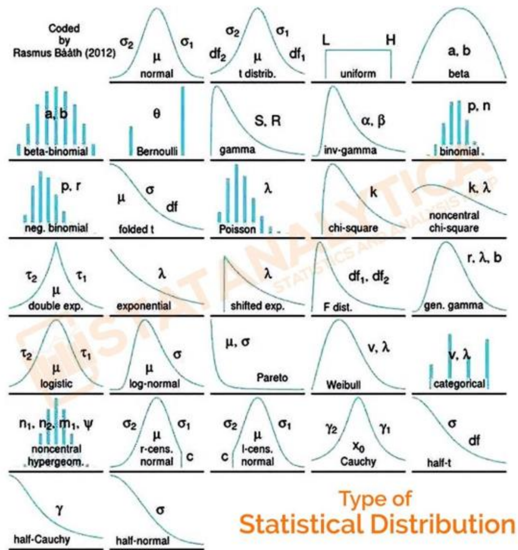


https://medium.com/mytake/understanding-different-types-of-distributions-you-will-encounter-as-a-data-scientist

10

SJSU SAN JOSÉ STATE UNIVERSITY

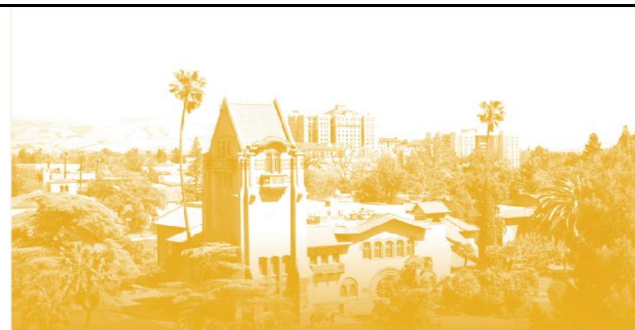
More Comprehensive List!!!



11

SJSU SAN JOSÉ STATE UNIVERSITY

Generalized Linear Models (GLMs)



12

What are GLMs?

Generalized Linear Model (GLM):

- a class of models popularized by McCullagh and Nelder in 1970-1980s
- unification of various statistical models like linear, logistic and Poisson regressions under a common framework.
- a flexible extension of traditional linear regression models.
- handle various types of response variables (unified approach to modeling of diverse data types)

13

Why are GLMs Important in Data Mining ?

GLMs are Important in data mining for several reasons:

- Flexibility to accommodate different types of response variables (e.g. binary, counts etc)
- Unified framework of various statistical models
- Handling of Non-Normal Data vs linear regression
- Use of Link Functions to model nonlinear relationships between the predictors and the response variables
- Robustness – robust to violations of assumptions that would otherwise invalidate simpler models
- Wide applications in many fields.

https://statisticseasily.com/generalized-linear-models/

14

Motivations

Generalized Linear Models (GLMs) are quite versatile and are used in various fields. Here are some common applications:

- Healthcare: GLMs are used to model patient outcomes, such as predicting the probability of disease occurrence or recovery rates based on patient characteristics and treatment plans.
- Finance: In risk management and insurance, GLMs help in modeling claim frequencies and severities, as well as in credit scoring to predict the likelihood of loan defaults.
- Marketing: GLMs are used to analyze customer behavior, such as predicting purchase probabilities, customer segmentation, and response rates to marketing campaigns.

https://statisticseasily.com/generalized-linear-models/

15

Motivations

- Environmental Science: They are used to model relationships between environmental factors and outcomes, such as pollution levels and health impacts, or species distribution based on habitat characteristics.
- Social Sciences: GLMs help in analyzing survey data, such as modeling voting behavior, social attitudes, and demographic influences on various outcomes.
- Manufacturing: In quality control and process optimization, GLMs are used to model defect rates and improve production processes.

https://statisticseasily.com/generalized-linear-models/

16

SJSU SAN JOSÉ STATE UNIVERSITY

GLM Assumptions

- The data $Y_1, Y_2, \ldots, Y_N$ are independently distributed, i.e., cases are independent.
- The dependent variable Y does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.).
- GLM does NOT assume a linear relationship between the response variable and the explanatory variables.
- It does assume a linear relationship between the transformed expected response (with the link function) and the explanatory variables.

https://statisticseasily.com/generalized-linear-models/

17

SJSU SAN JOSÉ STATE UNIVERSITY

GLM Assumptions (cont)

- Explanatory variables $X_1, X_2, \ldots, X_p$ can be nonlinear transformations of some original variables.
- The homogeneity of variance does NOT need to be satisfied.
- Errors need to be independent but NOT normally distributed.
- Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS).

18

SJSU SAN JOSÉ STATE UNIVERSITY

Linear Regression Models Recap

The linear (regression) model really has three components:

- Random Component: the response variable $Y|\boldsymbol{X}$ is continuous and normally distributed with mean $\mu = \mu(\boldsymbol{Y}) = E[Y|\boldsymbol{X}]$
- Systematic Component: explanatory variables $\boldsymbol{X}$ and its linear predictor is

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots = \boldsymbol{X}^T\boldsymbol{\beta}$$

- Link Function: The "link" between the random component & the systematic components:

$$\boldsymbol{X} = [X_1 \quad X_2 \quad \dots \quad X_p]^{\boldsymbol{T}}: \quad \mu(\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}$$

What is the link function in this case?

19

SJSU SAN JOSÉ STATE UNIVERSITY

Generalization of Linear Regression Models

A generalization to the linear regression model (GLM) is as follows:

- Random Component:

$$Y \sim \text{some exponential family distribution}$$

Normal Distribution: continuous response variables
Binomial Distribution: binary response variables
Poisson Distribution: count data

- Systematic Component (and the linear combination):

$$\eta = \boldsymbol{X}^T\boldsymbol{\beta}$$

- Link Function :

$$g(\mu(\boldsymbol{X})) = \boldsymbol{X}^T\boldsymbol{\beta} \quad \Longleftrightarrow \quad \mu(\boldsymbol{X}) = g^{-1}(\boldsymbol{X}^T\boldsymbol{\beta})$$

where g is the link function and $\mu(\boldsymbol{X}) = E[Y|\boldsymbol{X}]$.

20

The Exponential Family of Distributions

- In GLMs, the response variable is assumed to follow a distribution from the exponential family, which includes distributions like normal, binomial, Poisson, and gamma.
- The probability density function (pdf) for the exponential family can be written as:

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

canonical parameter

dispersion parameter

specific functions that define the distribution

	Normal	Poisson	Bernoulli
Notation	$\mathcal{N}(\mu,\sigma^2)$	$\mathcal{P}(\mu)$	$\mathcal{B}(p)$
Range of y	$(-\infty,\infty)$	$[0,-\infty)$	$\{0,1\}$
ϕ	σ^2	1	1
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1+e^\theta)$
$c(y,\phi)$	$-\frac{1}{2}(\frac{y^2}{\phi}+\log(2\pi\phi))$	$-\log y!$	1

21

Distribution	Domain	$\mu = E[Y\|x]$	$v(\mu)$	$\theta(\mu)$	$b(\theta)$
Binomial $B(n,p)$	$0,1,\ldots,n$	np	$\mu - \frac{\mu^2}{n}$	$\log\frac{p}{1-p}$	$n\log(1+e^\theta)$
Poisson $P(\mu)$	$0,1,\ldots,\infty$	μ	μ	$\log(\mu)$	e^θ
Neg. Binom. $NB(\mu,\alpha)$	$0,1,\ldots,\infty$	μ	$\mu+\alpha\mu^2$	$\log(\frac{\alpha\mu}{1+\alpha\mu})$	$-\frac{1}{\alpha}\log(1-\alpha e^\theta)$
Gaussian/Normal $N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	1	μ	$\frac{1}{2}\theta^2$
Gamma $N(\mu,\nu)$	$(0,\infty)$	μ	μ^2	$-\frac{1}{\mu}$	$-\log(-\theta)$
Inv. Gauss. $IG(\mu,\sigma^2)$	$(0,\infty)$	μ	μ^3	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$
Tweedie $p\geq 1$	depends on p	μ	μ^p	$\frac{\mu^{1-p}}{1-p}$	$\frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^\alpha$

22

SJSU SAN JOSÉ STATE UNIVERSITY

Link Functions

The link function connects the linear predictor to the mean of the distribution function.

Common link functions include:

Identity Link	Logit Link	Probit Link	Log Link	Inverse Link	Cloglog Link
$g(\mu) = \mu$	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$g(\mu) = \Phi^{-1}(\mu)$	$g(\mu) = \log(\mu)$	$g(\mu) = \frac{1}{\mu}$	$g(\mu) = \log(-\log(1-\mu))$
continuous	binary	binary	count	skewed cont.	binary

23

SJSU SAN JOSÉ STATE UNIVERSITY

Common Types of GLMs

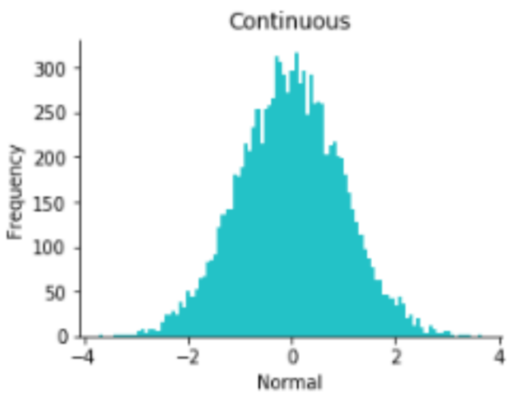
- Linear Regression (Continuous Outcome)
- Logistic Regression (Binary Outcome)
- Poisson Regression (Count Outcome)

- Gamma Regression (Positive Continuous Outcome)
- Probit Regression (Binary Outcome with Less Sensitivity to Extreme Values)
- Complementary Log-Log Regression (Binary Outcome with Small Probability of Success)

24

SJSU SAN JOSÉ STATE UNIVERSITY

Continuous Data ➔ Linear Regression



Data type: continuous
Domain: $(-\infty, \infty)$
Examples: house price, salary, person's height

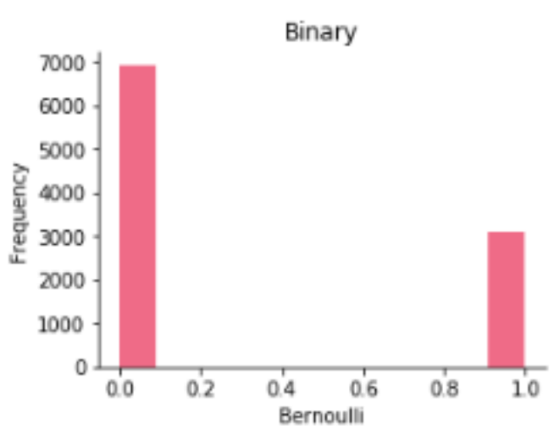
Family: `Gaussian()`
Link: identity
$g(\mu) = \mu = E(y)$

Model = Linear regression

25

SJSU SAN JOSÉ STATE UNIVERSITY

Binary Data ➔ Logistic Regression



Data type: binary
Domain: $0, 1$
Examples: True/False

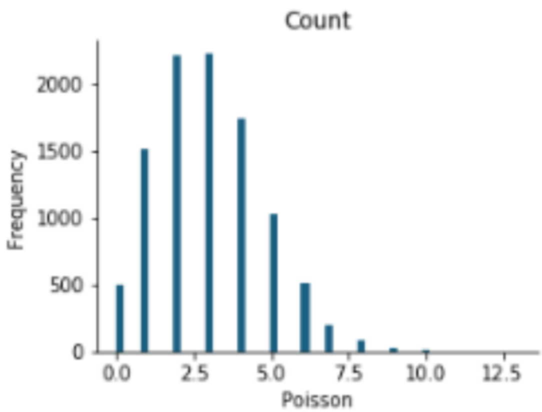
Family: `Binomial()`
Link: logit

Model = Logistic regression

26

SJSU SAN JOSÉ STATE UNIVERSITY

Count Data ➔ Poisson Regression



Data type: count
Domain: $0, 1, 2, \ldots, \infty$
Examples: number of votes, number of hurricanes

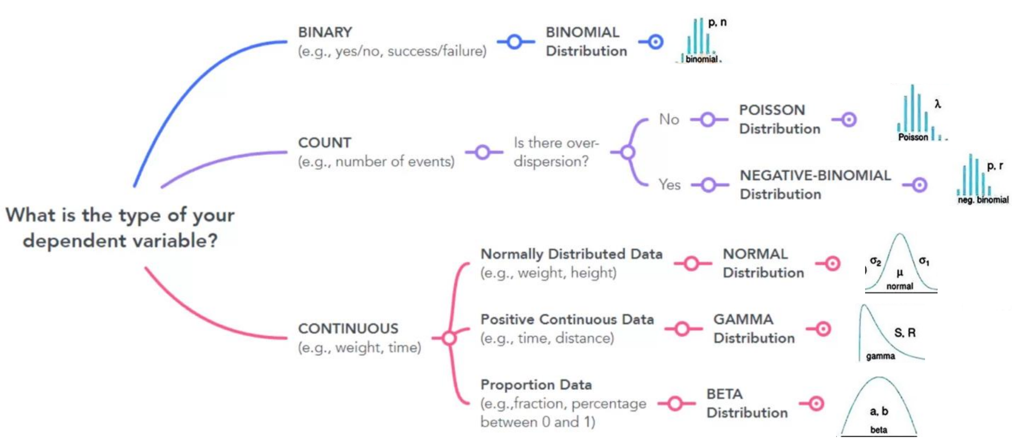
Family: `Poisson()`
Link: logarithm

Model = Poisson regression

27

SJSU SAN JOSÉ STATE UNIVERSITY

Dependent Variables & Distributions

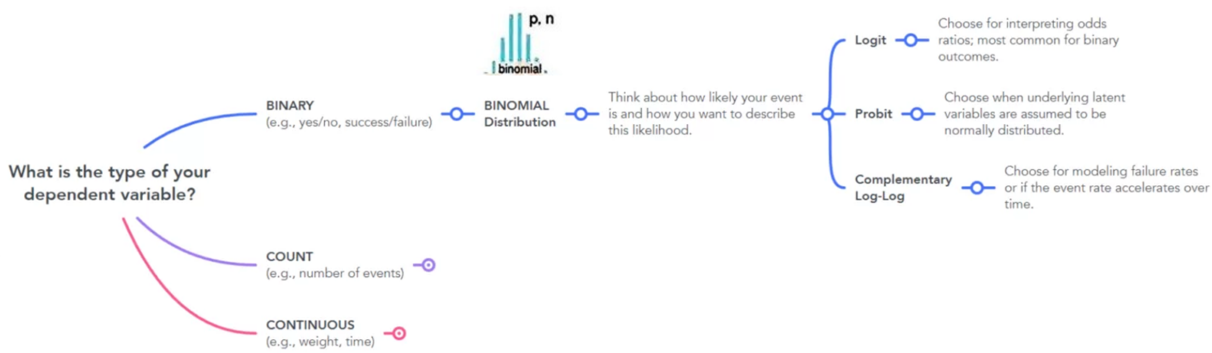


https://statisticseasily.com/generalized-linear-model-distribution-and-link-function/

28

SJSU SAN JOSÉ STATE UNIVERSITY

Binary Variables, Binomial Distribution & Link Functions

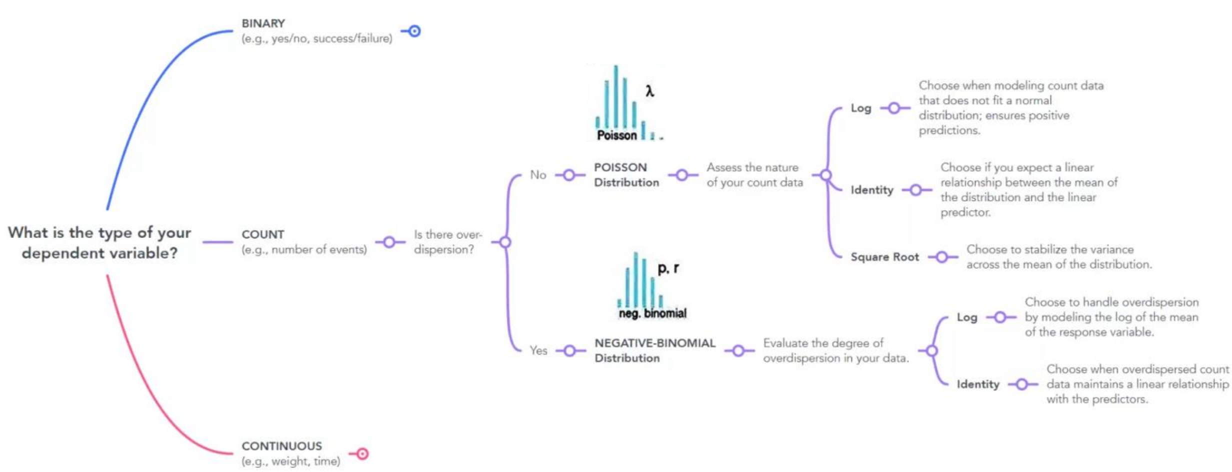


https://statisticseasily.com/generalized-linear-model-distribution-and-link-function/

29

SJSU SAN JOSÉ STATE UNIVERSITY

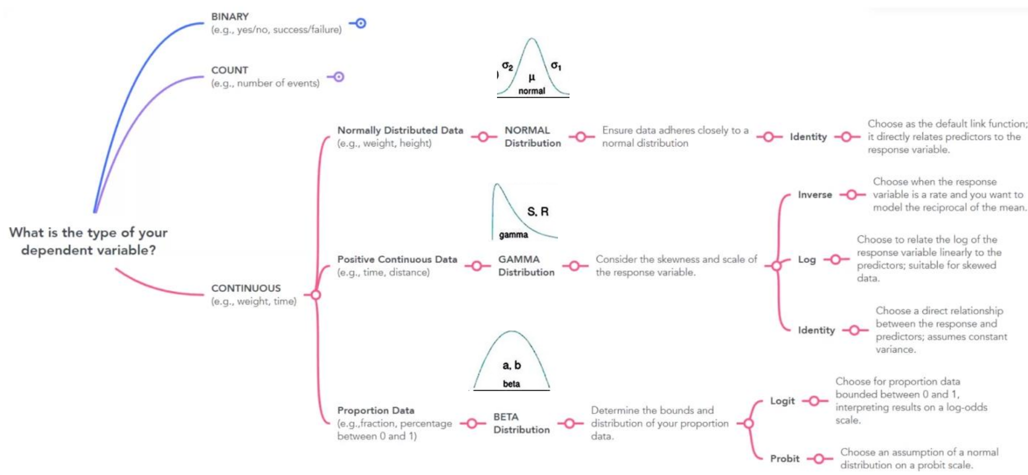
Count Variables, Distributions & Link Functions



https://statisticseasily.com/generalized-linear-model-distribution-and-link-function/

30

Continuous Variables, Distributions & Link Functions



https://statisticseasily.com/generalized-linear-model-distribution-and-link-function/

31

Summary Table of Different GLMs

Response Variable Type	Suggested Distribution	Common Link Functions	Use Case
Binary Outcome (e.g., success/failure)	Binomial	Logit, Probit, Complementary Log-Log	Modeling probabilities of binary outcomes, such as presence/absence of a disease.
Count Data (e.g., number of events)	Poisson	Log, Identity, Square Root	Counting occurrences in fixed intervals, such as the number of calls received by a call center per hour.
Count Data with Overdispersion	Negative Binomial	Log, Identity	Count data that exhibit variability exceeding Poisson assumptions, such as the number of insurance claims per client.
Continuous Proportions	Beta	Logit, Probit	Proportions that vary between 0 and 1, such as the fraction of an area affected by a certain condition.
Positive Continuous Data	Gamma	Inverse, Log, Identity	Modeling waiting times or service times, where the response variable is always positive.
Normally Distributed Data	Normal (Gaussian)	Identity	Continuous outcomes that are symmetrically distributed, such as test scores or heights.

32

SJSU SAN JOSÉ STATE UNIVERSITY

GLM Diagnostics

Model diagnostics are crucial for assessing the fit and validity of Generalized Linear Models (GLMs). Here are some key diagnostics and techniques used to evaluate GLMs:

- Residual Analysis using Pearson Residuals:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(\hat{\mu}_i)}}$$

estimated variance of y_i

- Influence Measures using Leverage & Cook's distance:

$$h_i = X_i(X^T W X)^{-1} X_i^T \qquad D_i = \frac{(r_i^2 h_i)}{p(1-h_i)^2}$$

diagonal matrix of weights

- Goodness-of-Fit using Deviance & Pearson Chi-Square Test:

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right] \qquad X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{V}(\hat{\mu}_i)}$$

33

SJSU SAN JOSÉ STATE UNIVERSITY

GLM Diagnositcs

- Model Validation using:
  - Cross-validation
  - Bootstrapping

- Collinearity Diagnostics using Variance Inflation Factor (VIF)
  - for highly correlated explanatory variables

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

coefficient of determination of the regression of predictor (j) on all other predictors

34

SJSU SAN JOSÉ STATE UNIVERSITY

Model Comparison

The two common criteria used for model selection and comparison for GLMs are:

- Akaike Information Criterion (AIC): lower values → better model
  - AIC penalizes models with more parameters to prevent overfitting

$$\text{AIC} = -2\log(L) + 2k$$

L - maximum likelihood of the model
k - # of parameters in model
n - # of observations

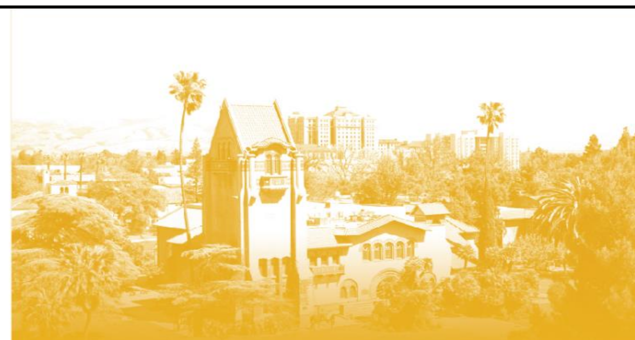
- Bayesian Information Criterion (BIC): lower values → better model
  - BIC penalizes models with more parameters more heavily than AIC, especially as the sample size (n) increases.

$$\text{BIC} = -2\log(L) + k\log(n)$$

35

SJSU SAN JOSÉ STATE UNIVERSITY

GLM Demos



36

Building GLMs in Python

- Using statsmodels library
  `import statsmodels.api as sm`
- Model fitting
  – Specify the model ➔ `model = sm.GTM(y, X, family=…)`
  – Fit the model ➔ `result = model.fit()`
  – Summarize the model ➔ `result.summary()`
  – Make model predictions➔ `result.predict()`

37

SJSU SAN JOSÉ STATE UNIVERSITY

Advanced GLM Topics

38

SJSU SAN JOSÉ STATE UNIVERSITY

More Complicated System Component

Here are some more complex forms of the systematic component:

- Generalized Additive Models (GAMs)

$$\eta = \beta_0 + f_1(X_1) + f_2(X_2) + \ \dots \ + f_p(X_p)$$

- Generalized Estimating Equations (GEEs)

$$\eta = X\beta$$

- Nonlinear Models

$$\eta = g(X, \beta)$$

- Interaction Terms

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$$

39

SJSU SAN JOSÉ STATE UNIVERSITY

Summary

- GLMs are an extension of linear models that allow for response variables that have error distribution models other than a normal distribution
- A unified framework that includes logistic, Poisson, Gamma regression etc
- 3 main components: Random Component, Systematic Component and Link Function
- It's based on MLE rather than OLS
- Can check goodness of fit using deviance and Pearson residuals & X^2 tests
- Common model comparison criteria: AIC and BIC

40