An Introduction to the Diversity of Variables, Data Generating Processes, and Statistical Methods (Regression Models) for Analysis

Nicola Righetti

Summary and Goals

- Gain a basic, intuitive but rigorous grasp of the fundational concepts for a deeper understanding of regression models.
- Clarify how various variables necessitate distinct statistical models. Our primary focus will be on regression models.
- In this process, we think “statistically” and familiarize ourselves with certain statistical concepts and notations.

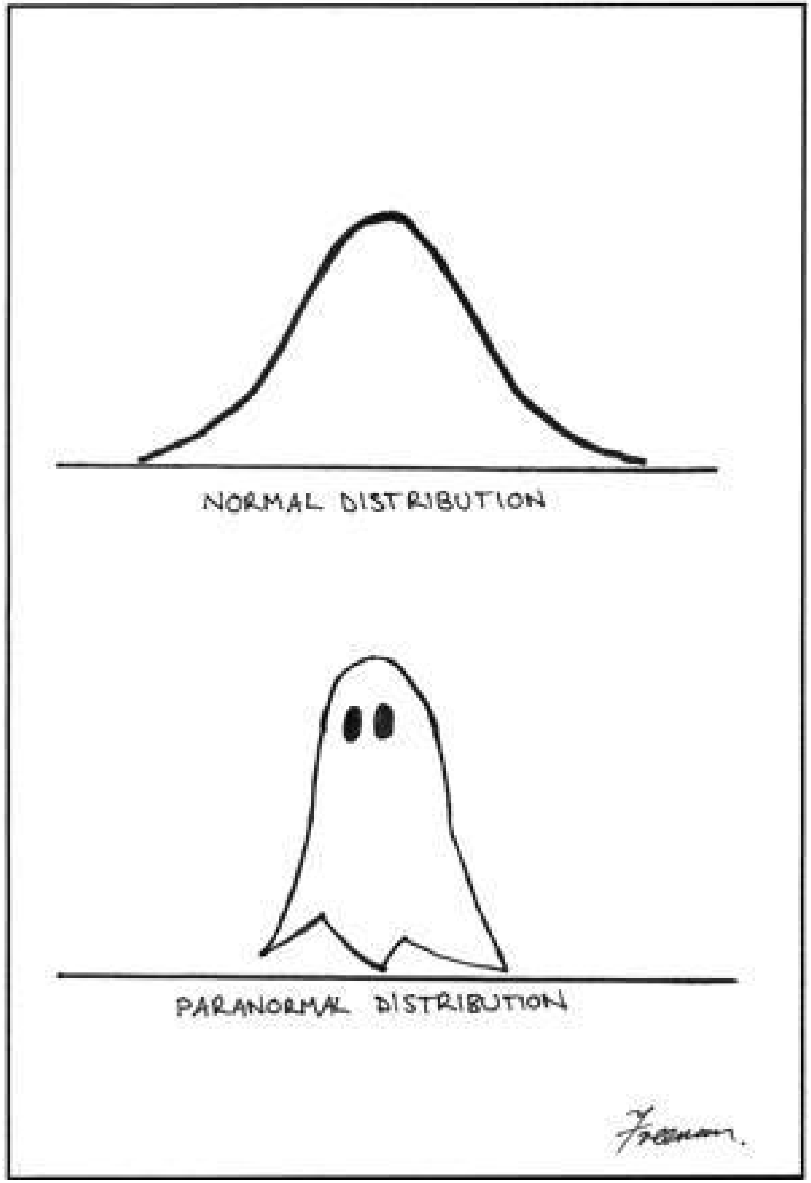
Advisory Notice

- The concepts we go over generally take time to assimilate and understand, and several readings and re-readings of reference texts.
- The lesson can be challenging to follow and require attention. Don’t panic and try to follow as possible.
- In this lecture we only aim to an introduction so that you have a map (and references) to help you go deeper when/if needed.
- Beyond the useful theoretical background, the main content you need to understand is practical/applicational in nature, and it is immediately understandable. The theoretical background is a plus.

Reference

- King, G. (1989). *Unifying political methodology: The likelihood theory of statistical inference.* Cambridge University Press (in particular, first part of chapter one and chapter five)
- Additional resources (advanced and complementary to the book): Gary King's lectures for *Quantitative Social Science Methods I, Gov2001 at Harvard University* on YouTube: https://www.youtube.com/watch?v=qs2uCuDL2OQ&list=PL0n492lUg2sgSevEQ3bLilGbFph4l92gH

Beyond Normality



Regression Models

- A regression model is a statistical tool used for quantifying the relationship between one or more independent variables (predictors or explanatory variables) and a dependent variable (outcome).
- It estimates how changes in the predictors are associated with changes in the outcome.
- This model is widely used to infer relationships (correlational, and under certain conditions, causal relationships), and for prediction.

Basic concepts and intuitions

Social System, Explanatory Variables, Output

- **Social system:** The ultimate object of study in the social sciences. It has features that are known or can be estimate, and other that are or remain unobserved and unknown.
- **Explanatory variables:** measures of the observed features of a social system, symbolized as X. X contains measures of each of these features for all n observations $(x_1, x_2, \ldots, x_n)$
- **Output (dependent variable):** consequences of the social system that can be observed and measured. For example, an election (social system) produces a victorious candidate, a communicative frame given other psycho-social characteristics (social system) could produce a specific opinion.

Experiment (1/2)

- **Experiment:** In experiments, the researcher can manipulate the explanatory variables (the known features of the social system) and make it produce output as needed.
- *Statistically*, the idea of an experiment is extended to include purely theoretical scenarios, such as imagining *multiple runs* of a presidential election *under the same conditions*.

Experiment (2/2)

- Let Y represent an n-dimensional vector of social system outputs (the set of candidates winning the election), where each output $\mathrm{Y_i}$ (a specific winning candidate) is theoretically produced by setting known system features ($\mathrm{x_i}$) and running the experiment.
- *Despite identical conditions* (identical values of the system features $\mathrm{x_i}$), the system generates *varying output* for each experiment due to the *probabilistic nature of output generation*, as opposed to deterministic.

Random Variables (1/2)

- **Random variables:** Operationally, a random variable is the assignment of numbers to events (Democratic President = 1, Republican President = 0; income = number of dollars, etc.) with a probability assigned to each number.
- Random variables are generated by random processes, which are not random at all (not “haphazard”) but adhere to very specific probabilistic rules.

Random Variables (2/2)

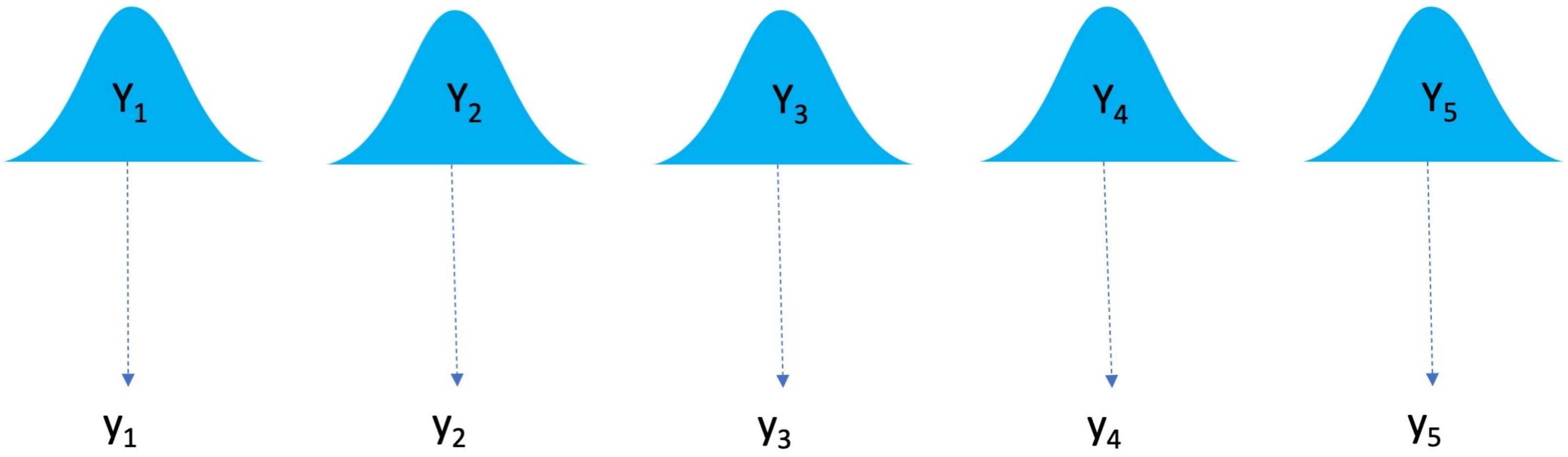
The output Y_i is a random variable since:

- The actual data are randomly produced from the social system's outputs according to each event's probability.
- It varies across an infinite number of hypothetical experiments, despite identical conditions of the social system (x_i) that generates the output.
- If the actual experiment were "run" again under essentially the same conditions, the observed values of the dependent variable (the "realizations" of the random variables) would differ but the nature of the experiment and the social system would remain constant.

Data

- **The data:** The data, y, are n observed realizations of the random variables Y. They are a set of numbers, each y_i being a random draw from a corresponding random dependent variable, Y_i.
- The process by which portions of output are chosen and measured is called *sampling*, and the data themselves are sometimes called “the sample”.

Data



Random Variables (Y) Y_1 Y_2 Y_3 Y_4 Y_5

Data (y) y_1 y_2 y_3 y_4 y_5

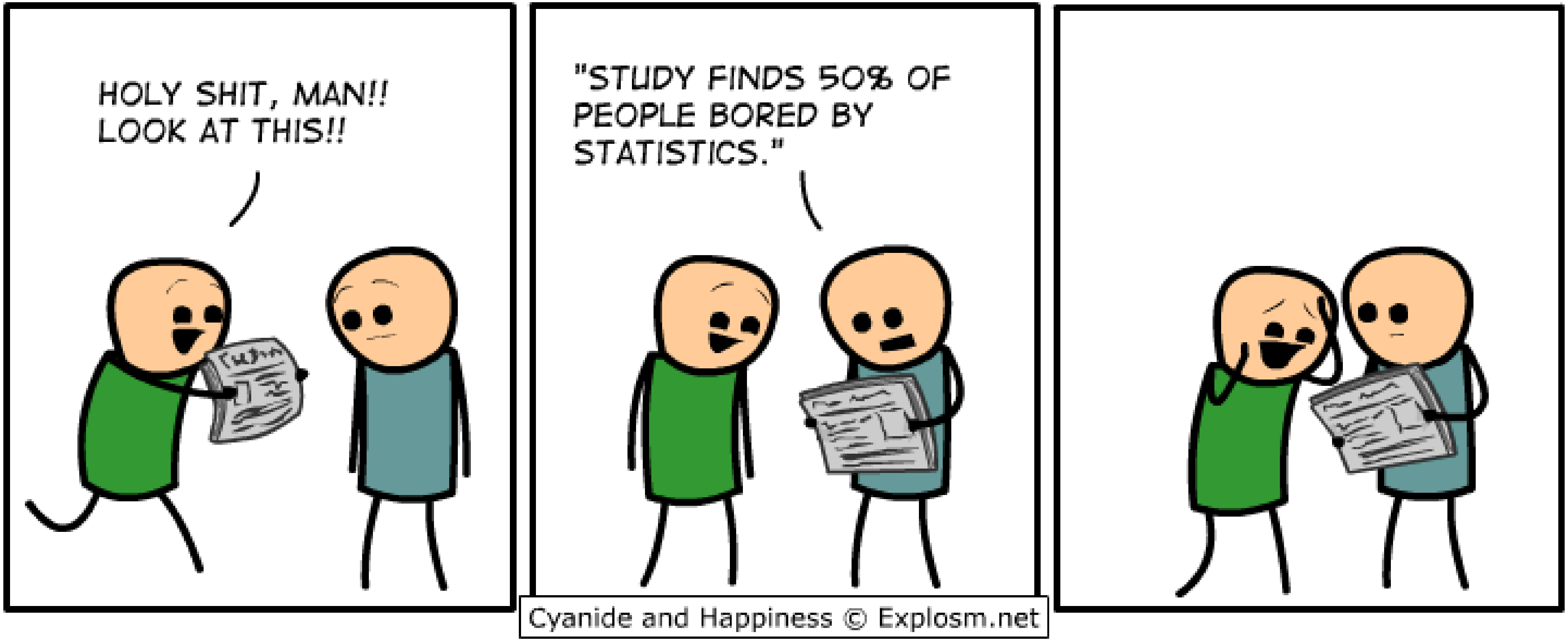
Models

- **Models:** A model is a mathematical simplification of, and approximation to, a more complex concept or social system.
- Models are never literally "true", although one often proceeds as if they were.

Statistical Models

- **Statistical models:** A statistical model is a formal representation of the *process by which a social system produces output*.
- The essential goal is to *learn about the underlying process* that generates output and hence the observed data.
- Statistical models are assumed to have both **_systematic_ and _stochastic_** components. In this lesson, we focus on a popular family of statistical models: regression models.

Boring?


HOLY SHIT, MAN!! LOOK AT THIS!!
"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."
Cyanide and Happiness © Explosm.net

Notation (1/6)

You can see regression models written in different ways.

- True population parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Estimated parameters from sample data (*hat* notation):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Alternative notation for estimated parameters (similar to Beta hat):

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2$$

Notation (2/6)

A more synthetic notation is as follows:

$$Y_i = x_i\beta + \epsilon_i$$

The systematic component is $x_i\beta$ and the stochastic component is ϵ.

In the common linear-Normal regression model, the stochastic component is represented by the Normal distribution with mean zero and constant variance σ^2

$$\epsilon_i \sim f_n(e_i|0, \sigma^2)$$

The subscript $_i$ is used to denote the values pertaining to a single observation or row in a dataset.

Notation (3/6)

$$Y_i = x_i \beta + \epsilon_i,$$
$$\epsilon_i \sim f_n(e_i|0, \sigma^2)$$

This equation represents a linear model. In the first part:

- Y_i is the dependent variable for the $i-th$ observation.
- x_i is the independent variable or predictor for the $i-th$ observation.
- β is the coefficient or parameter of the model that measures the impact of x_i.
- ϵ_i represents the error term or residual for the $i-th$ observation, capturing the deviation of the observed value from the value predicted by the model.

Notation (4/6)

$$Y_i = x_i\beta + \epsilon_i,$$
$$\epsilon_i \sim f_n(e_i|0, \sigma^2)$$

The second part describes the distribution of the error term ϵ_i. It states that:

- ϵ_i follows a distribution f_n with a mean of 0 and a variance of σ^2. The distribution f_n, in the case of the linear-Normal model, is the normal distribution

Notation (5/6)

A more general notation, is as follows:

$$\begin{aligned} Y_i &\sim (y_i | \mu_i, \sigma^2), \\ \mu_i &= x_i \beta \end{aligned}$$

- The first equation represents that the variable Y_i for the $i-th$ observation follows a distribution characterized by the parameters μ_i and σ^2.
- The second equation specifies that the mean μ_i of the distribution of Y_i for the $i-th$ observation is a linear function of the independent variables x_i and the parameter vector β.

Notation (6/6)

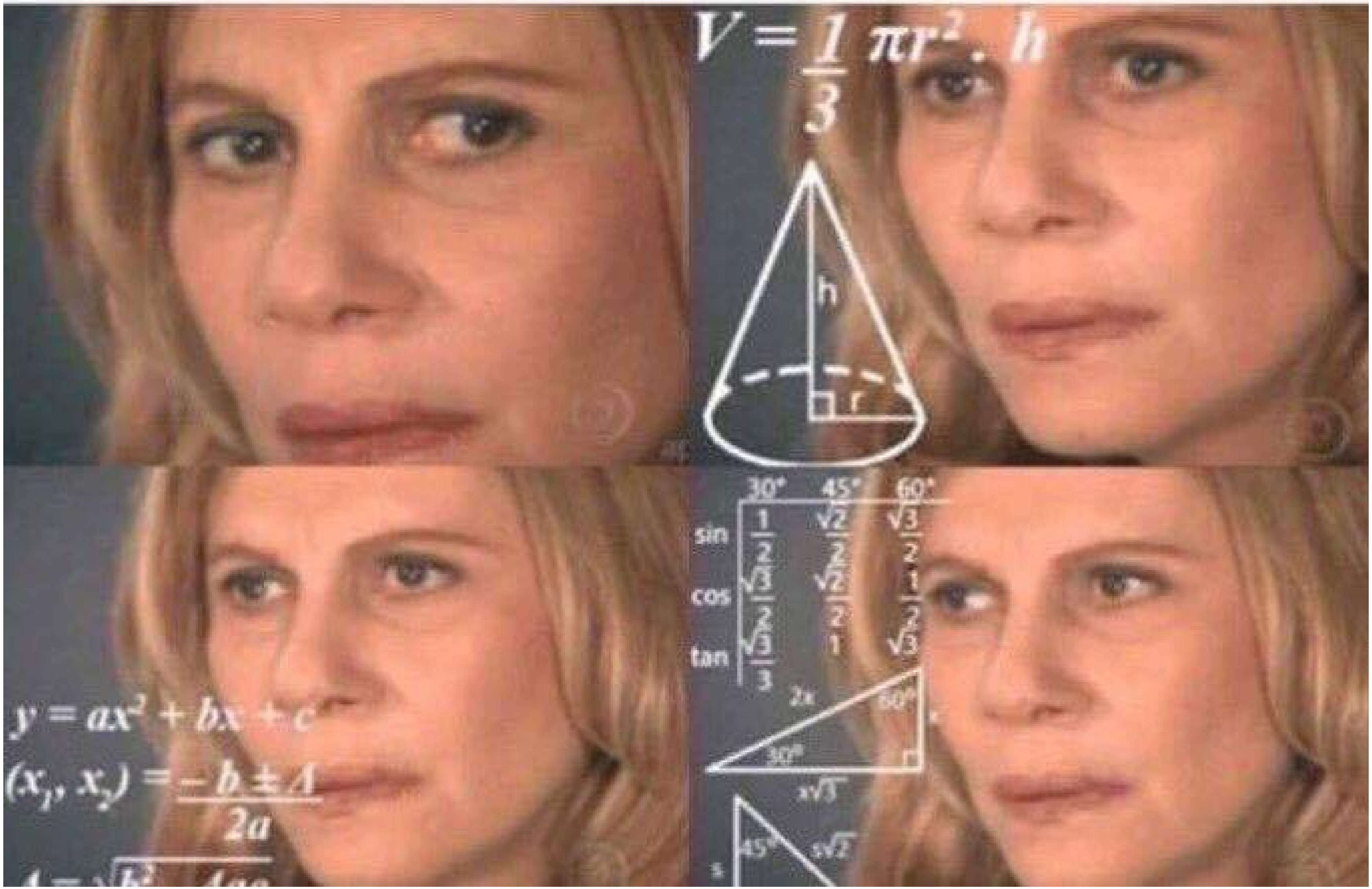
In this format, the randomness in the dependent variable Y_i is modeled directly, and the systematic component models variation in one of its parameters (the mean, μ_i) over the observations.

$$\begin{aligned} Y_i &\sim (y_i | \mu_i, \sigma^2), \\ \mu_i &= x_i \beta \end{aligned}$$

This notation is fine for the linear-Normal model. The general case, for almost any model, is as follows:

$$\begin{aligned} Y &\sim f(y|\theta, \alpha), \\ \theta &= g(X, \beta) \end{aligned}$$

No more math :)

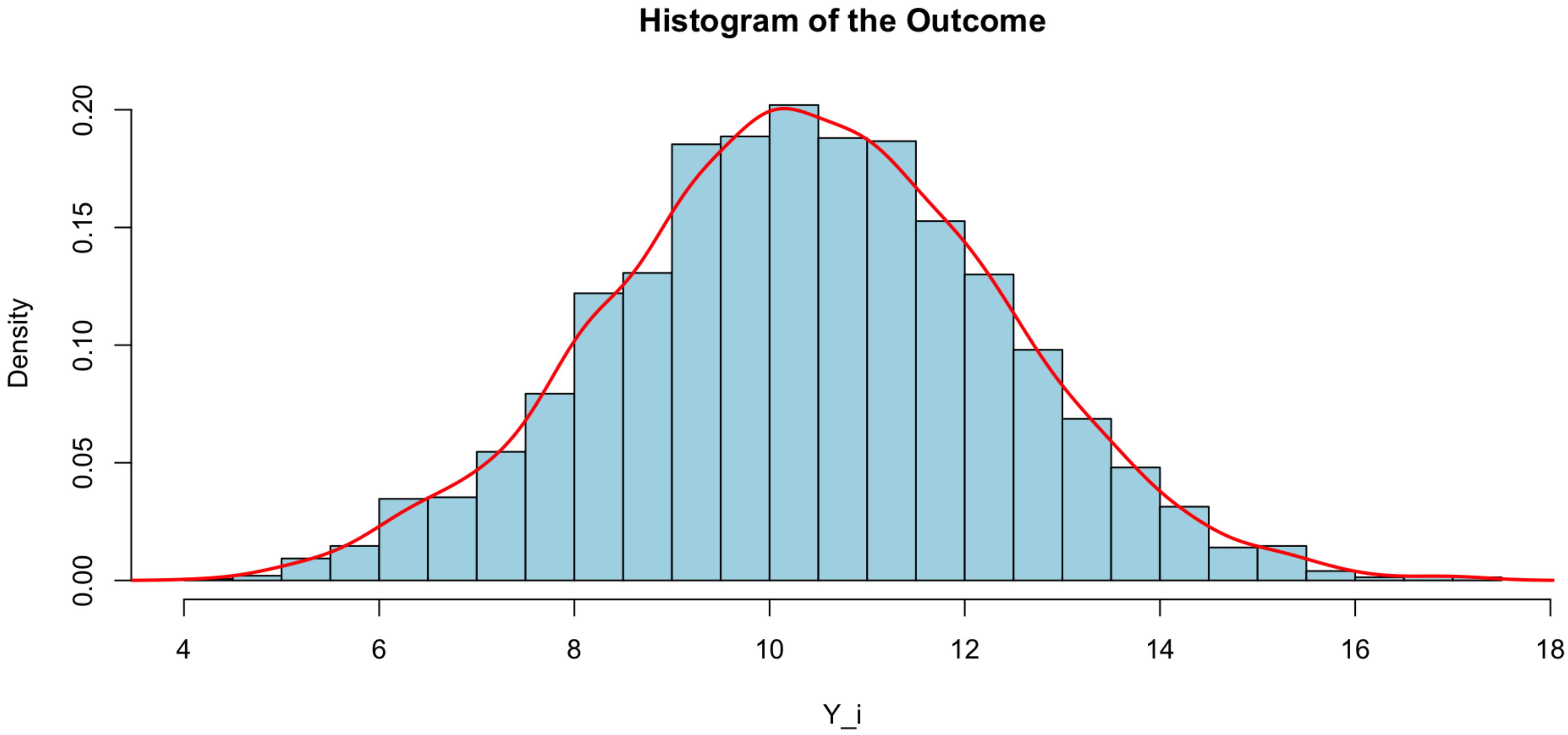

$V = \frac{1}{3}\pi r^2 . h$
h
r
30° 45° 60°
sin $\frac{1}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{\sqrt{3}}{2}$
cos $\frac{\sqrt{3}}{2}$ $\frac{\sqrt{2}}{2}$ $\frac{1}{2}$
tan $\frac{\sqrt{3}}{3}$ 1 $\sqrt{3}$
$y = ax^2 + bx + c$
$(x_1, x_2) = \frac{-b \pm \Delta}{2a}$
2x
60°
30°
$x\sqrt{3}$
45°
$s\sqrt{2}$
s

Simulation - Fixed X (1/3)

```r
# Set the seed for reproducibility
set.seed(123)

# Parameters for fixed x_i scenario
n <- 3000                    # Number of observations
beta <- c(10, 0.5, -0.3)     # Coefficients including intercept
sigma <- 2                   # Standard deviation of the error term
x_i <- c(1, 0.5, -0.3) # Fixed values for independent variables

# Calculate deterministic part (mean) using fixed x_i
mu_i <- x_i %*% beta

# Replicate fixed_mu_i for each observation
mu_i <- rep(mu_i, n)

# Generate response variable Y_i using fixed mu_i as the mean
Y_i <- rnorm(n, mean = mu_i, sd = sigma)

# Results
head(Y_i)

```

Simulation - Fixed X (2/2)

```
[1]  9.219049  9.879645 13.457417 10.481017 10.598575 13.770130
```

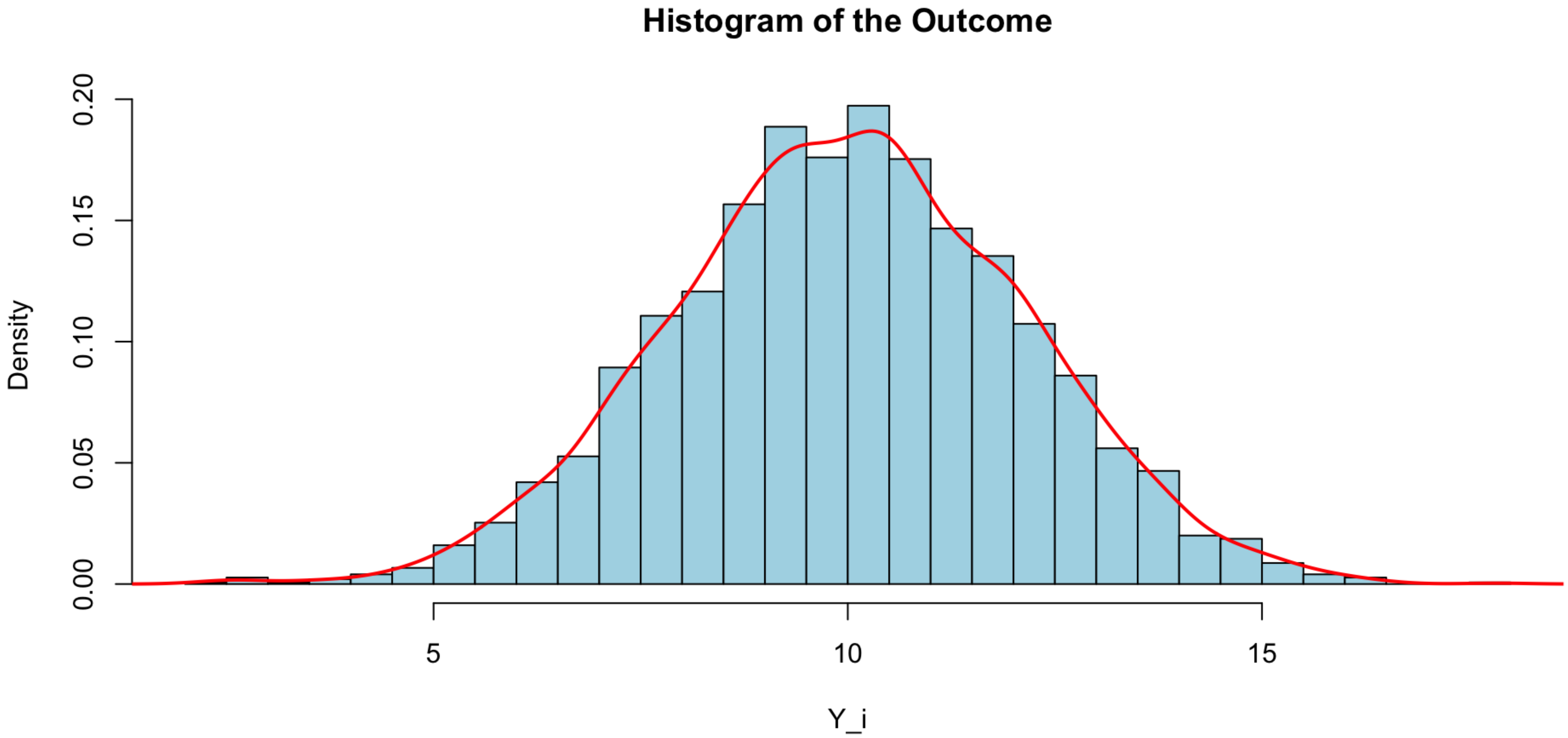


Simulation - Varying X (1/2)

```r
# Set the seed for reproducibility
set.seed(123)

# Parameters
n <- 3000                   # Number of observations
beta <- c(10, 0.5, -0.3)    # Coefficients including intercept
sigma <- 2                  # Standard deviation of the error term

# Simulate independent variables (including a column for the intercept)
x_i <- cbind(rep(1, n), matrix(rnorm(n * 2), ncol = 2))

# Calculate deterministic part (mean for each observation)
mu_i <- x_i %*% beta

# Generate response variable Y_i using mu_i as the mean
Y_i <- rnorm(n, mean = mu_i, sd = sigma)

# Results
head(Y_i)

# Create the histogram
```

Simulation - Varying X (2/2)

```
[1]  8.366398 11.976141  9.828313 10.037473 10.492829  9.652667
```



Stochastic component (1/2)

The stochastic component is represented in the first equation:

$$Y \sim f(y|\theta, \alpha),$$
$$\theta = g(X, \beta)$$

- f is a probability distribution, an explicit model of the form of uncertainty in the random variable Y across repeated statistical experiments.

Stochastic component (2/2)

$$Y \sim f(y|\theta, \alpha),$$
$$\theta = g(X, \beta)$$

- θ and α are parameters that define the distribution. We are usually more interested in θ. In the linear-Normal regression case, it is the mean μ, but in the logistic regression case, it is π, the probability of observing one of the two outcomes of a dichotomous variable. α is often an ancillary parameters, like the constant standard deviation σ^2 in the linear-Normal case.

Systematic component

The systematic component is represented in the second equation:

$$
\begin{gathered}
Y \sim f(y|\theta, \alpha), \\
\theta = g(X, \beta)
\end{gathered}
$$

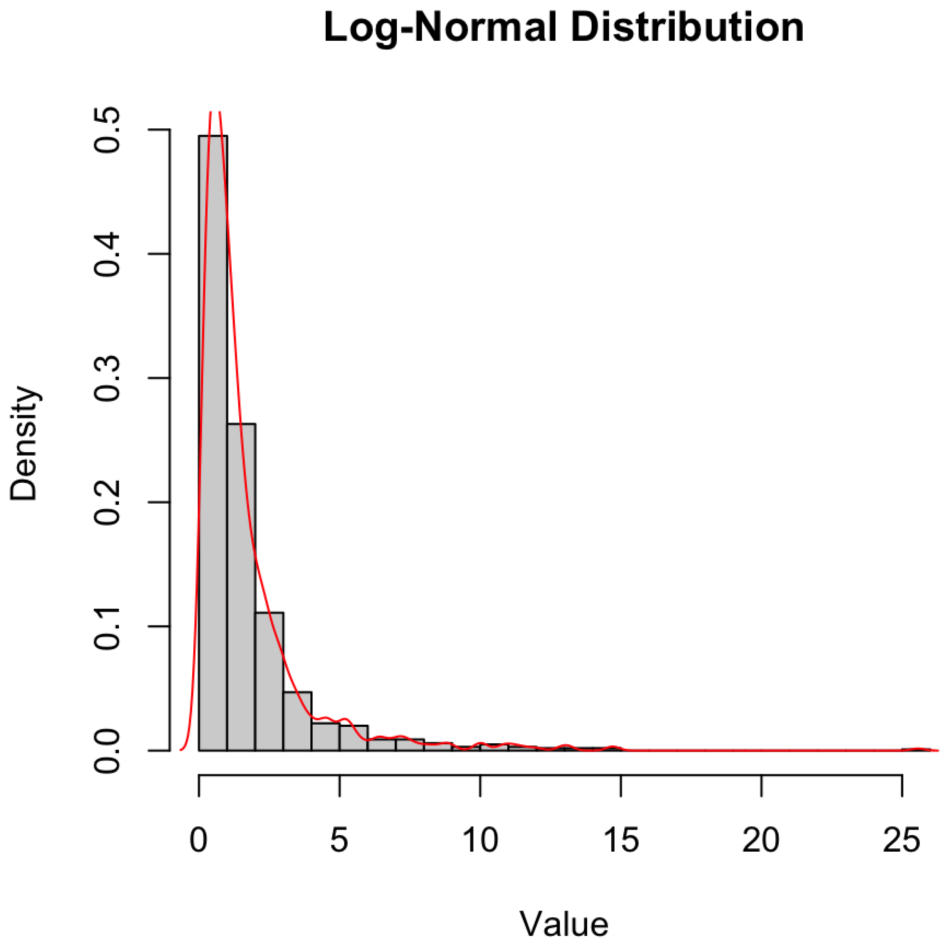
- It is a statement of how θ_i varies over observations as a function of a vector of explanatory variables.
- β are called the effect parameters and represent the degree and direction of dependence of θ on the explanatory variables X.
- g is the functional form that expresses the relationship between explanatory variables X and the output Y.
  - E.g.: it is a *linear function* for the common *linear-Normal regression* model, and a *logit function* for the *Logistic regression model*.

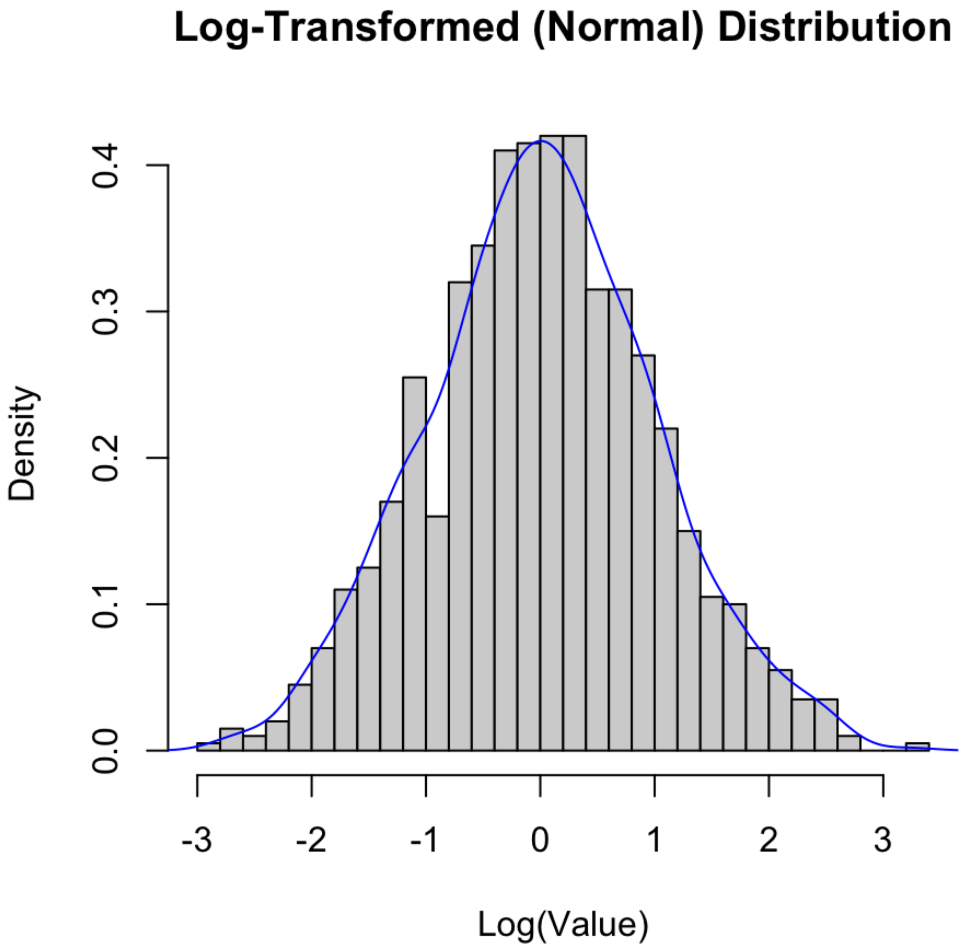
Regression models

Choose the Appropriate Model

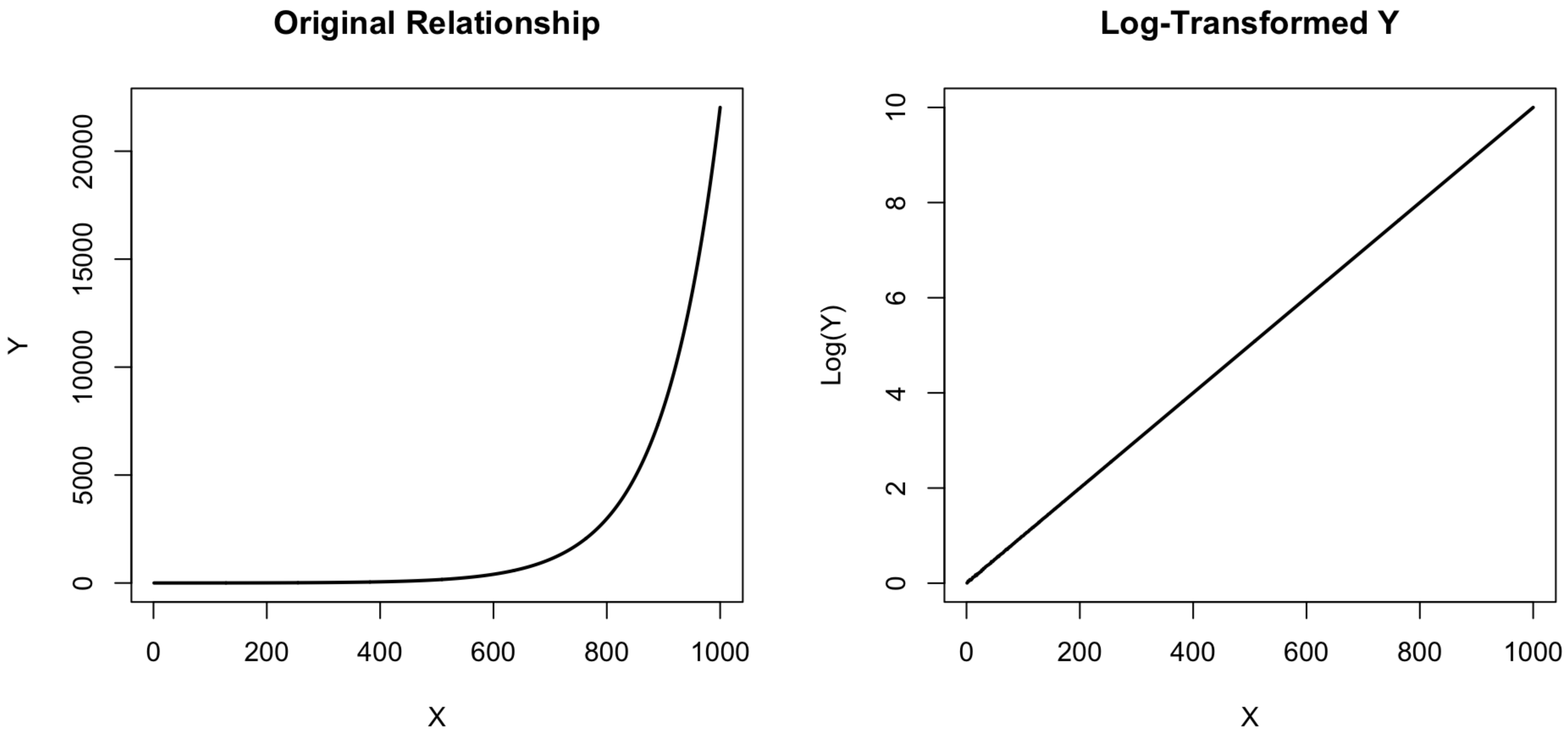
- The linear-Normal regression is only one among many models available.
- It is sometimes indiscriminately used in the most diverse situations, perhaps after “transforming” in more or less complex and counterintuitive ways the dependent variable “to make it normal” (a common “trick” is to take the log of the Y variable).
- However, it is always reccomended to model the variables as they are.

Transformations (1/2)





Transformations (2/2)



Models, Variables, Probability Distributions

- The choice of models depends on the nature of the variables.
- Different models are required to create statistical models of continuous variables (interval or ratio scale), or discrete variables ("count data," usually defined on a ratio scale), dichotomous, or multicategorical variables.

Models, Variables, Probability Distributions

Type of Variable	Probability Distribution	Appropriate Regression Model (Common Choices)
Continuous (Interval or Ratio)	Normal	Linear-Normal Regression Model
Discrete (Count)	Poisson, Negative Binomial	Poisson Regression or Negative Binomial Regression
Dichotomous (Nominal)	Bernoulli	Logistic Regression
Multi-Categorical (Nominal)	Multinomial	Multinomial Logistic Regression
Categorical (Ordinal)	Cumulative Logistic	Proportional Odds (or Ordinal Logistic) Regression Model

The Data Generating Process

Choosing the appropriate statistical model requires an understanding of the data generating process.

- A data generating process (DGP) refers to the *underlying mechanism* or series of mechanisms that *produce the observed data.*
- It encompasses the *probabilistic* (stochastic component) *and causal factors* (deterministic component) that govern how data points are generated and distributed.
- Understanding the DGP is crucial for selecting appropriate statistical models and methods for data analysis, as it guides assumptions about the structure and behavior of the data.

Choose the Appropriate Statistical Model

In other words, choosing the appropriate statistical model requires an understanding of two main elements:

- the nature of the variable Y that is to be explained or predicted, and specifically of its statistical distribution f;
- the relationship g that binds it to the explanatory variables X.

$$\mathrm{Y} = \mathrm{f} \sim (\mathrm{y}|\theta, \alpha),$$
$$\theta = \mathrm{g}(\mathrm{X}, \beta)$$

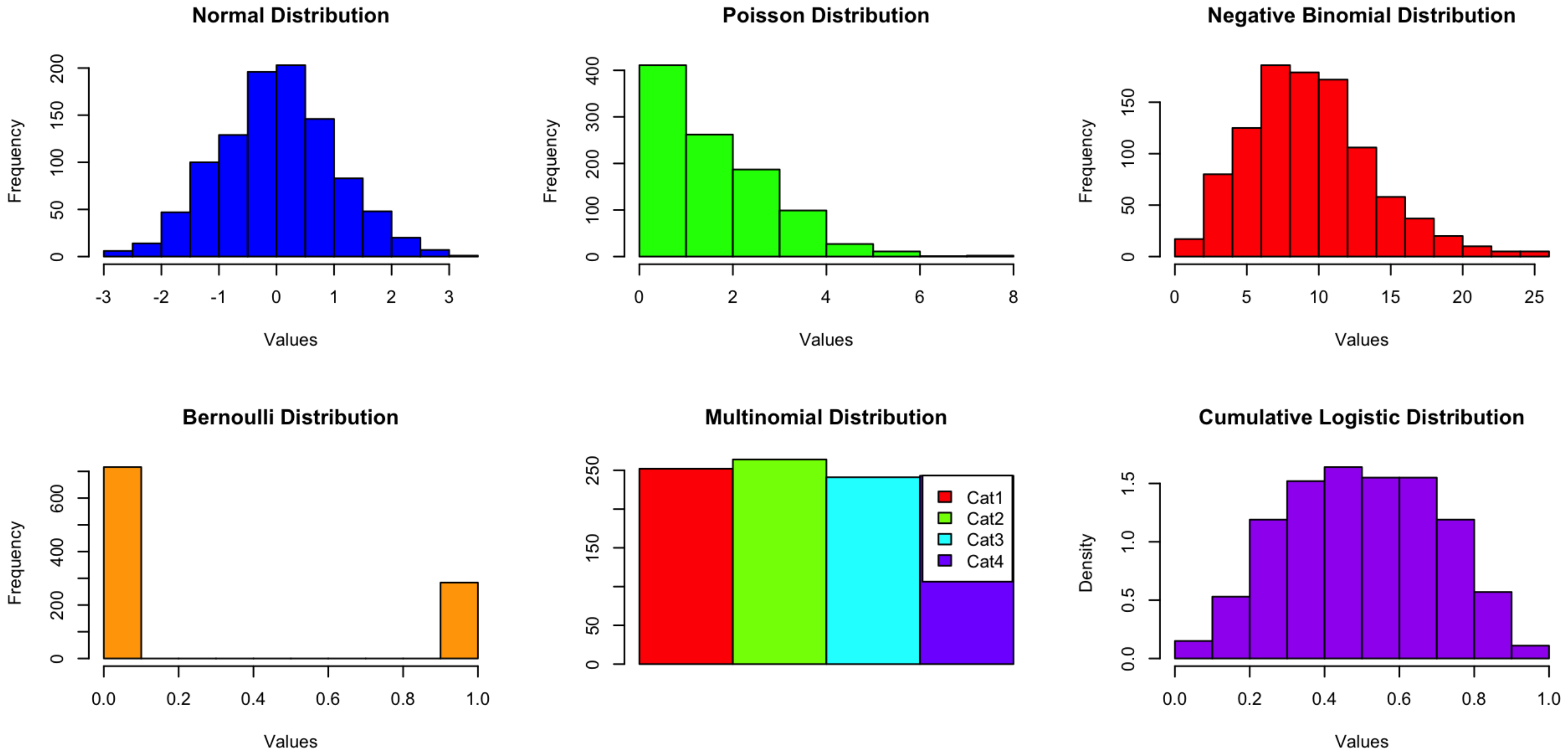
Choose the appropriate statistical distribution

- When modeling a variable Y we must choose an appropriate probability distribution f. This is the stochastic part of the data generating process and is used to model the stochastic component.

$$Y \sim f(y|\theta, \alpha),$$
$$\theta = g(X, \beta)$$

- For example, count data or dichotomous variables, cannot follow a normal distribution. Their distribution is generally, for count data, either the Poisson distribution or the Negative Binomial, and for dichotomous (binary) variables, the Bernoulli distribution.

Probability Distributions



Probability Distributions and DGP

- **Normal Distribution**: This distribution generally arises when a dataset is influenced by many small, random disturbances that are independent and identically distributed. A common example is measurement errors. In a normal distribution, most data points cluster around a central mean with a certain standard deviation, creating a symmetric, bell-shaped curve.
- **Example:** Imagine a study examining the time spent by individuals on social media platforms daily. The variation in time spent is influenced by numerous small factors like individual preferences, daily schedules, and content availability. The time spent by a large group of people is likely to form a normal distribution, with most individuals clustering around an average time, and fewer individuals showing extremely high or low usage times.

Probability Distributions and DGP

- **Bernoulli Distribution**: This is a simple distribution representing the outcome of a single experiment which can result in either success or failure (typically represented as 1 or 0, respectively). It is often used in situations where there is a binary outcome, like a coin toss.
- **Example:** Consider a study analyzing whether a news article is shared on social media by an individual. Here, the outcome for each individual is binary: they either share (success, 1) or do not share (failure, 0) the article. Each individual’s decision to share or not can be modeled as a Bernoulli trial.

Probability Distributions and DGP

- **Binomial Distribution**: The binomial distribution extends the Bernoulli distribution to encompass multiple trials. It represents the number of successes in a fixed number of independent trials, with each trial having the same probability of success. An example is the number of heads in a series of coin tosses.
- **Example:** Extend the previous example to a situation where researchers are interested in how many articles from a set of 10 are shared by individuals on social media. The number of articles shared by each individual follows a binomial distribution, where the number of trials is the total number of articles (10), and the probability of success is the likelihood of an individual sharing any given article.

Probability Distributions and DGP

- **Multinomial Distribution**: This is a generalization of the binomial distribution. It models the outcome of experiments where each trial results in one of several possible categories, rather than just success/failure. An example is the distribution of different outcomes (like rolling a 1, 2, 3, 4, 5, or 6) when rolling a dice multiple times.
- **Example:** Suppose a study is conducted to understand how viewers allocate their time across different types of media content (e.g., news, entertainment, educational, etc.) on a platform. Each viewing session can be categorized into one of these types. The distribution of the number of times each content type is viewed over multiple sessions by a viewer can be modeled using a multinomial distribution.

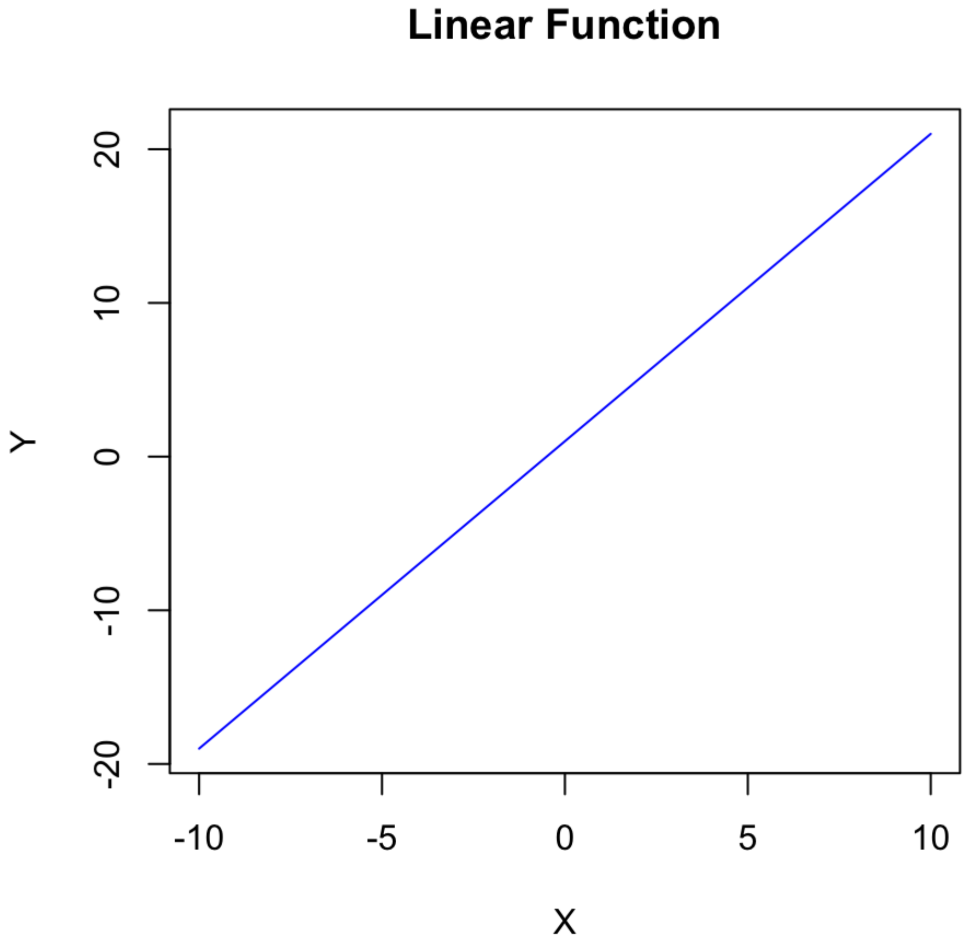
Choose the appropriate functional form

The proper functional form g should also be specified.

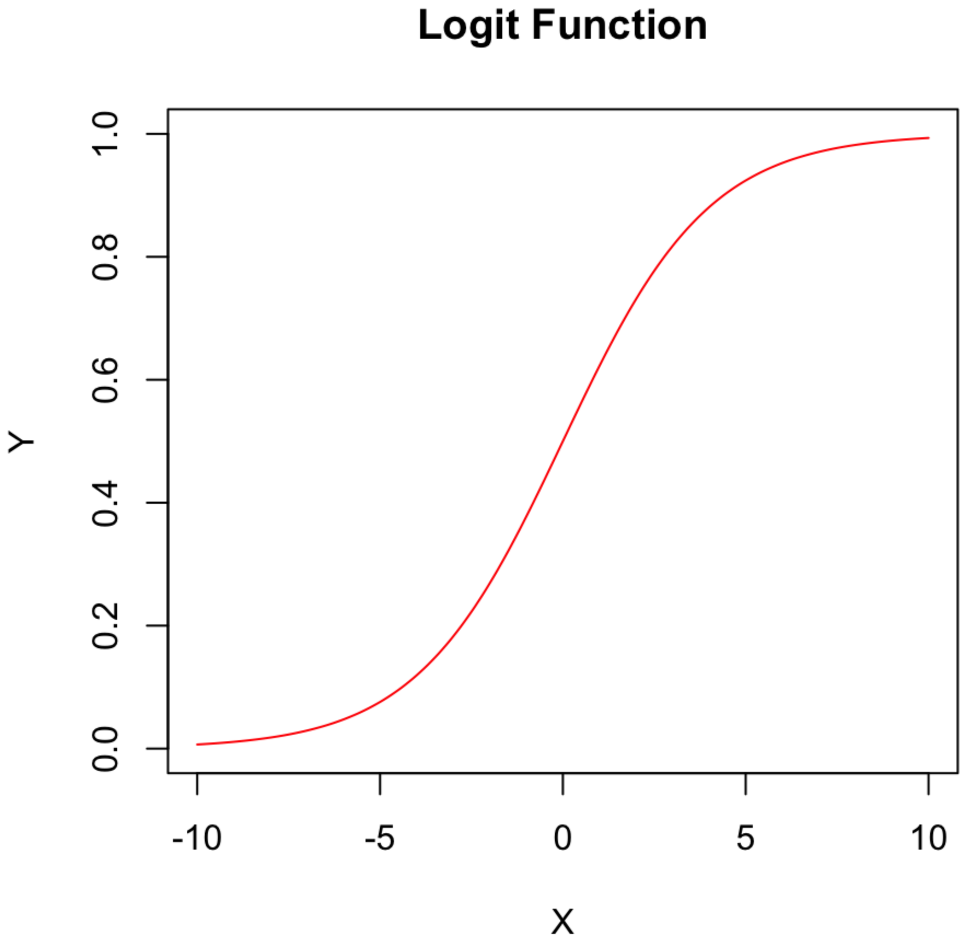
$$Y \sim f(y|\theta, \alpha),$$
$$\theta = g(X, \beta)$$

For example, the function that binds explanatory variables to a dichotomous variable (0-1), cannot be linear. In fact, a linear function extends indefinitely. The logistic function is more suitable, because it cannot extend beyond the values 0 and 1.

Choose the appropriate functional form



Linear Function



Logit Function