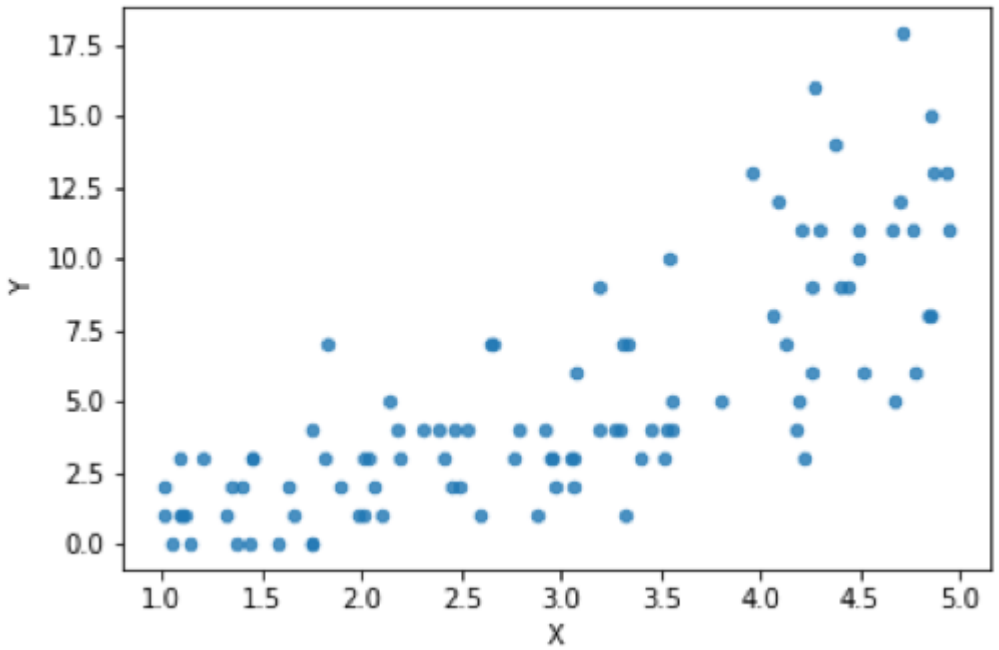
6.06 GLMs

Generalized Linear Models (GLMs)

- Generalized linear models are used to model the outcome variable as a linear combination of weights and predictor variable given the outcome variable and the error term follows the exponential family of distributions.

- This is different from the general linear models (linear regression) where response variable, Y, and the random error term (ϵ) must be based solely on the normal distribution.

Why use GLMs?



There are several problems if you try to apply linear regression for this kind of data.

1.The relationship between X and Y **does not look linear**. It's more likely to be exponential.

2.The variance of Y does not look constant with regard to X. Here, the variance of Y seems to increase when X increases.

3.As Y represents the number of products, it always has to be a positive integer. In other words, Y is a **discrete variable**. However, the normal distribution used for linear regression assumes continuous variables. This also means the prediction by linear regression can be negative. It's not appropriate for this kind of count data.

Here, the more proper model you can think of is the **Poisson regression** model. Poisson regression is an example of **generalized linear models (GLM)**.

GLM Components

- Random Component / Probability Distribution
  - The distributional assumption of our target variable.
  - Where does the randomness in our model come from?
  - What is the distribution of Y?
- Systematic Component / Linear Predictor
  - Which explanatory variables to include in the model?
- Link Function
  - connects the random and systematic component. It is the function of the expected value of the response variable which enables linearity in the parameters.
  - By its construction it allows the mean of the response variable to be nonlinearly related to the explanatory variables.

GLM Components – Poisson Example

- **Probability distribution** which generates the observed variable *y*. As we use Poisson distribution here, the model is called Poisson regression.
- **Linear predictor** is just a linear combination of parameter (*b*) and explanatory variable (*x*).
- **Link function** literally “links” the linear predictor and the parameter for probability distribution. In the case of Poisson regression, the typical link function is the log link function. This is because the parameter for Poisson regression must be positive (explained later).

Poisson GLM

- Count variables works well with Poisson distribution

- Examples
  - How many students will enroll in DSI next cohort?
  - How many puppies will be in the next litter?
  - How many lightbulbs will we go through in the bathroom next year?
  - How many sales will my website generate tomorrow?

Gamma GLM

- “Waiting-time” variables work well with Gamma distribution

- Examples
  - How long until this light bulb goes out?
  - How long will this rain last?
  - How long until the next economic downturn?
  - How long until this user unsubscribes from our service?