GLM

Rahul Telang

Linear model

- A linear model is an assumption about the nature of the relationship between (y) such as income and (x) such as education
- It describes how much income changes on average for a unit increase in education. It also describes how much of the variation of income is not explained by education

$$y_i = X_i \beta + e_i$$

- The systematic part is the average of Y given a value of X

  $\mu = E(y \mid X) = X\beta$

- And the stochastic part is what is left unexplained

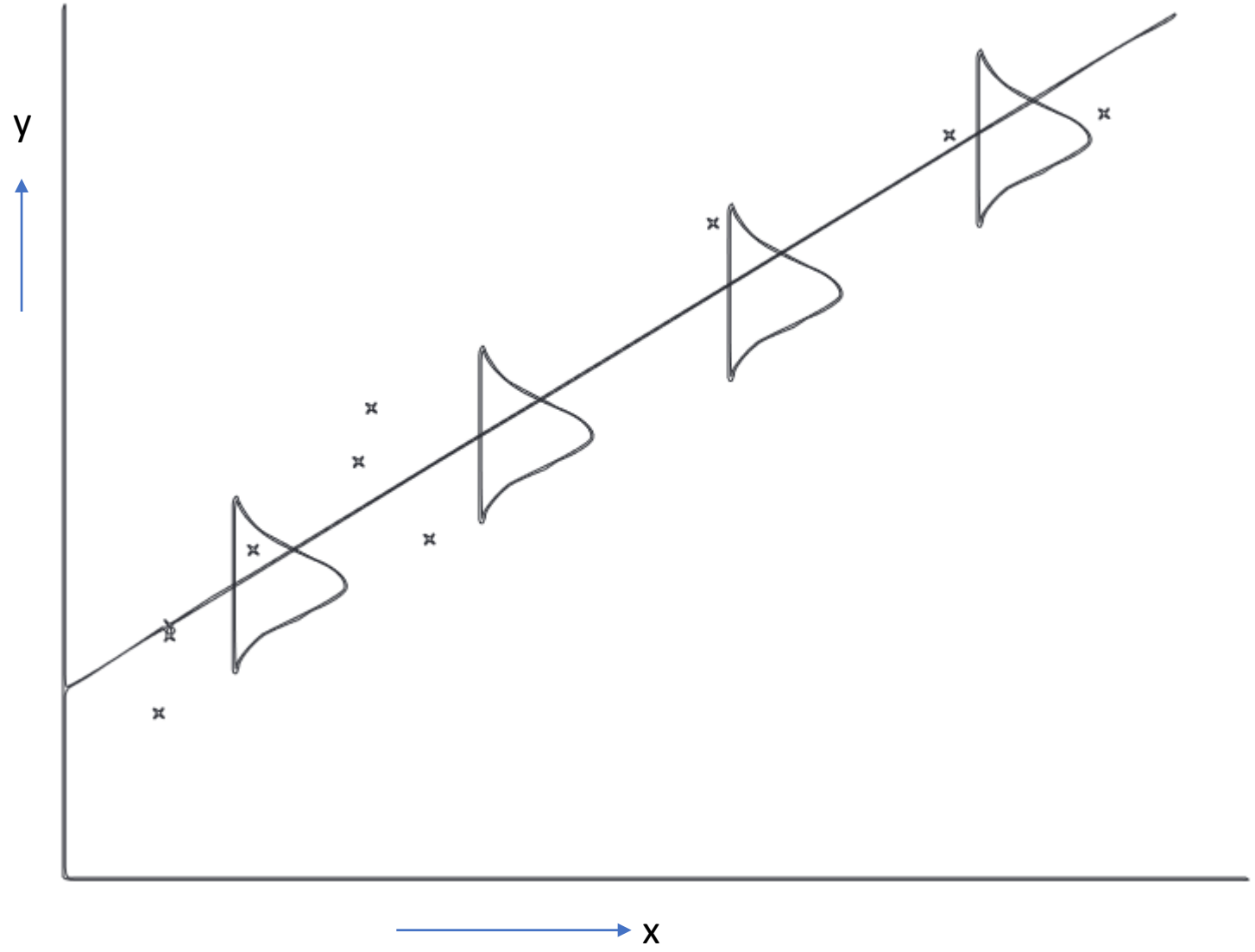
  $ei = yi - Xi\beta$

Linear model

- The stochastic component defines the distribution of Y
- It describes the variation of income when we have no predictors (i.e. when we do not know anything about education), all the variation of income is stochastic. We specify this component by making assumptions about the statistical process that generated the values of income. In linear models it is assumed to be “normal”
- We have seen that regression

Linear Model

- Responses are independent of each other;
- The mean response changes in a linear way with the conditions, but the functional shape of the distribution remains fundamentally unchanged.



GLM

- While the interpretability and robust theory has made regressions widely popular, the assumptions are too restrictive for many real-world prediction problems.
- The linear regression model assumes that the outcome given the input features follows a Gaussian distribution. Or $y \sim N (\mu = \beta_0 + \beta_1 x_{1i} + ..., \sigma^2)$
- However, this assumption excludes many realistic cases:
  - The outcome can be a category (cancer vs. healthy),
  - count (number of children),
  - The time to the occurrence of an event (time to failure of a machine) or,
  - a skewed outcome with a few very high values (household income).
- This linear model can be extended to model all these types of outcomes. This extension is called **Generalized Linear Models** or **GLMs** for short.

GLM

- The advantage of GML is that it follows the same structure as linear regression

$$Y = X\beta + \varepsilon$$

But in GLM we do not have to make assumption that Y is normally distributed. In short, we can be more flexible in choosing distribution of ε and hence, in turn, Y.

GLM

- How does GLM accomplish it?
- To be able to allow for non-gaussian term, in GLM
  - The distribution of the response variable, hence error term, given x is a member of the ***exponential family*** of distributions.
- Discussion of "exponential family" is beyond the scope of this course. (see https://en.wikipedia.org/wiki/Exponential_family). But exponential family incorporates many distribution that are of use like – normal, Poisson, logistics and so on..
- Any distribution that can be written as

  $P(y,\eta) = b(y)\exp(\eta^T T(y) - \alpha(\eta))$

  will be from exponential family. The parameter η is called a natural parameter.

GLM

- In GLM, this natural parameter (η) is **linear** weighted sum of covariates. Hence,
  $\eta = x'\beta$
- In regression, dependent variable is weighted sum of covariates
  - $E[y] = b0 + b1\ X$
- We now need to connect η with $E(Y)$. Since mean $E(y) = \mu$, we need one-to-one continuously differentiable transformation $g(\cdot)$ such that
  $\eta = g(\mu)$,

where $g(\cdot)$ is called the **link function**

- Since the link function is invertible and one-to-one, we have
  $\mu = g^{-1}(\eta) = g^{-1}(x'\beta)$.

Use GLM

- To operationalize GLM, one to has to define Link function g(.). Depending on the error distribution, one can define appropriate link function.
- For example, take traditional regression,

  $Y = x'\beta + e$
- Since gaussian error term fits in exponential family, GLM is applicable
- In OLS, we know that $E(Y)= \mu = x'\beta$
- In GLM, $\eta = x'\beta$,
- A link function links η with $g(\mu)$. Notice that $\eta = \mu$. Therefore, g(.) is an identity function. In short there is no transformation needed to run regression in GLM form.
- As we go forward, we will go over different link functions which are widely used.

Link functions

- Example of link functions

(Wikipedia)

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

Cost/Benefits of GLM

- To estimate the parameters, Log likelihood is a way to move forwards .
- GLM also does likelihood. Since it follows the same structure as regression, interpreting and using it is easy.
- It does likelihood maximization, but the technical details are hidden from us. It ensure that the likehood function is convex and there is one global maxima (as opposed to several local maxima). This is a major challenges of using maximum log likelihood for parameter estimation.
- All GLMs use the same method when finding solution. This makes the outcome clear.
- But
  - One is restricted to the exponential family.
  - One estimates the impact of covariates on the mean of Y. It does not directly measure how covariates change the distribution.

GLM

- To summarize in regression we have,

$$Y = \beta X + \varepsilon$$
$$E[Y] = \beta X$$
$$E[Y] = \mu = \eta$$
$$\varepsilon \sim N(0, \sigma_e^2)$$

- In GLM, we have a link function that transforms μ into $\eta = \beta X$

$$g(\mu) = \eta = \beta X$$
$$E[Y] = \mu = g^{-1}(\eta)$$
$$Y = g^{-1}(\eta) + \varepsilon$$
$$\varepsilon \sim \text{exponential family}$$

Example

- Using Python to estimate linear regression using GLM.
- Python offers Statsmodel library to perform various statistical model. It allows various function to do the analysis using GLM or maximum likelihood.
- https://www.statsmodels.org/stable/index.html provides details about Statsmodel
- There are two key libraries one should download

  import statsmodels.api as sm

  from patsy import dmatrices
- To fit most of the models in statsmodels, we create two design matrices. The first one is a matrix of endogenous variable(s) (dependent variable). The second is a matrix of exogenous variable(s) (i.e. independent, regressor, etc.)

example

- The data which is stored in Statsmodel library is already organized in endog and exog format.
- But for any other data we use, we use Dmatrics library from patsy. DMatrics is an internal data structure which is optimized for both memory efficiency and training speed. Dmatrices will generate two metrices (endog and exog) in a model Y=b0+b1X where Y (dependent variable) falls under endog metrics and X (independent variables) fall under exog metrics. Dmatrices will return a tuple structure, which we need to unpack and store the dependent variable and the independent variables separately.
- Example where a data set can be converted in endo and exog metrices to perform GLM operations

Python Example

Output

- Link function - In Regression, the link function used is $E(y) = \mu = \eta$ Or identity
- Method - iteratively reweighted least squares (IRLS) is the algorithm used for Maximum likelihood estimation.
- Deviance (D)- Deviance is a measure of error; lower deviance means better fit to data. If we use dummy as a covariate, we get null deviance. If we have full predictability, deviance will zero (fully saturated model). Deviance will lie between null deviance and fully saturated model).

Python example…

- Using statsmodel data
- Using csv file