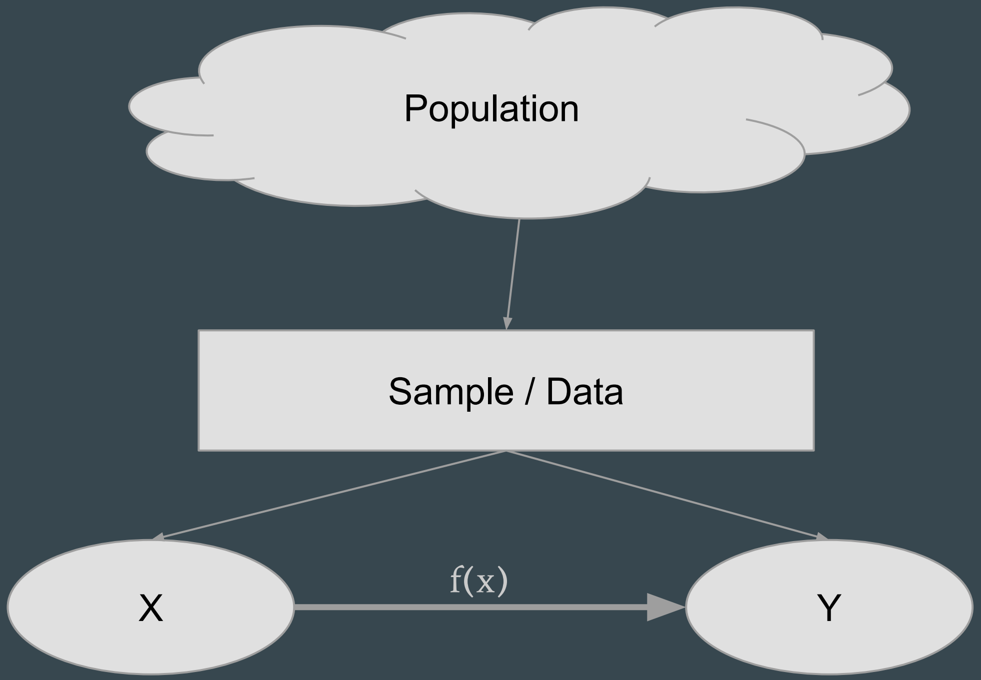
Generalized Linear Models (GLM)

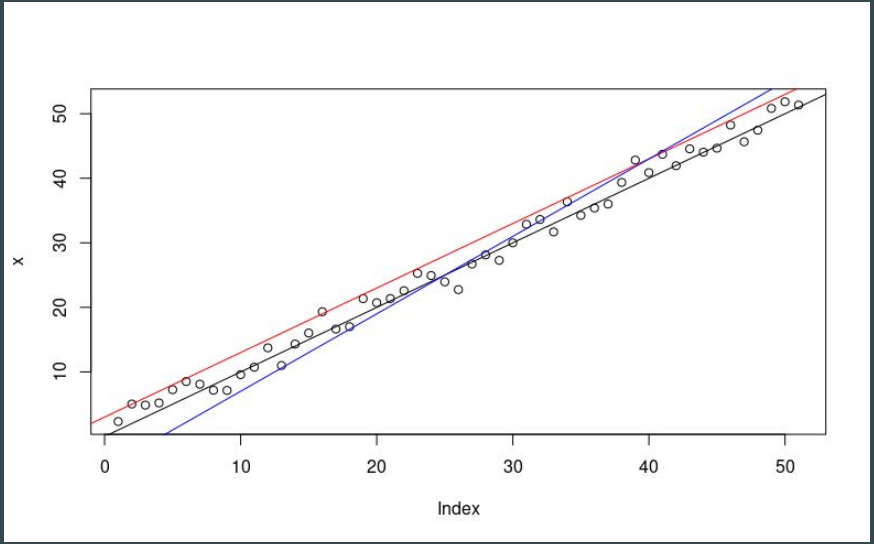
PLSC 309
8 April 2019

Review: statistical modelling



Review: Ordinary Least Squares

- Find the line (i.e slope and intercept) that minimizes the squared differences ($\hat{Y}$-Y)
- $\hat{Y}$-Y are known as errors or residuals



Review: MLE

In order to do MLE, we have to follow the following steps

1. Determine the probability distribution we think f(x) follows
2. Write down the probability distribution as a likelihood
3. Find the values of your parameters that maximizes the likelihood
4. Using the function you just produced, calculate the parameters with your data

Preview

- Provide an overview of the GLM framework
- Discuss distributions of Y
- The GLM equation and its parts
- Linear, Binomial, and Poisson GLM

Preview

Monday: GLM overview

Wednesday: Logistic regression deep-dive

Friday: Diagnostics for GLMs

PS 13 due Monday (April 15)

30,000 ft. view

Generalized Linear Models (GLM) allow us to measure all type of conditional probabilities of the type P(Y|X), but with all the good properties of OLS

How GLMs work

1. Determine distribution of outcome
2. Select an appropriate probability distribution that looks like the distribution of your outcome
3. Transform that probability distribution into a linear model

How GLMs work

There are two parts to the term Generalized Linear Model:

- *Generalized*: this refers to the fact that we can measure any type of relationship, not just straight lines
- *Linear Model*: we can transform that non-linear relationship to one expressed with linear parameters (β, α)

What we like about OLS (interpretability)

- Additivity is a major assumption of OLS
- While it makes it difficult to apply our results to the real-world, it makes those results very easy to read
- Each variable has its own β parameter, allowing us to directly interpret the effect of each variable

What we like about OLS (interpretability)

	Public posting about politics	
	β	*SE*
age	-.02	.01
gender	.11	.03
ethnicity	-.13*	.03
political interest	.45**	.01
political efficacy	.04	.01
F, Adj. R^2	*F* (5,225) = 15.32, .24**	

Note: Female =1, Male = 2, White = 1, non-White = 0. *p<.05. **p<.01.

OLS is interpretable due to additivity

- Additivity allows us to solve for each variable's coefficient separately
- This is a unique product of linear relationships
- Other types of relationships, e.g. Poisson or Exponential, cannot separate out the different effects of each X variable

MLE to the rescue!

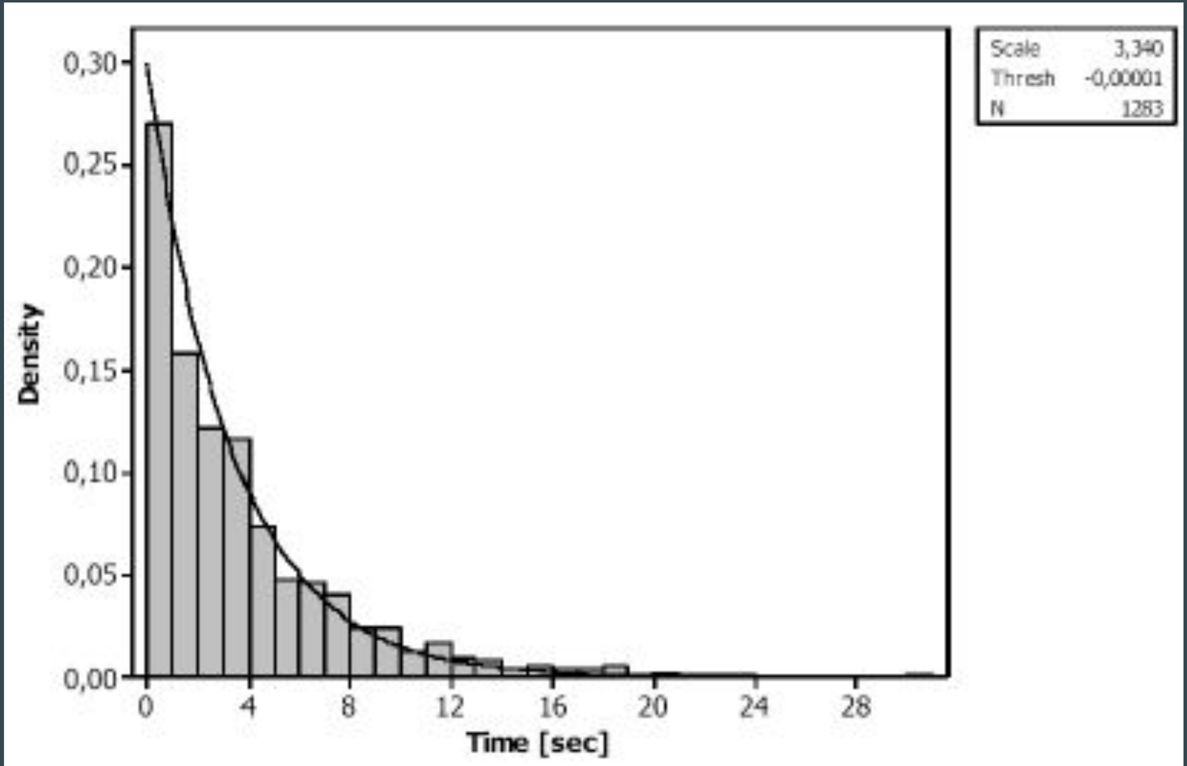
MLE gave us a way to estimate P(Y|X) for any distribution. We got this added power with two additional assumptions:

1. That P(Y|X) follows a certain distribution
2. That Y, conditional on X, is independent

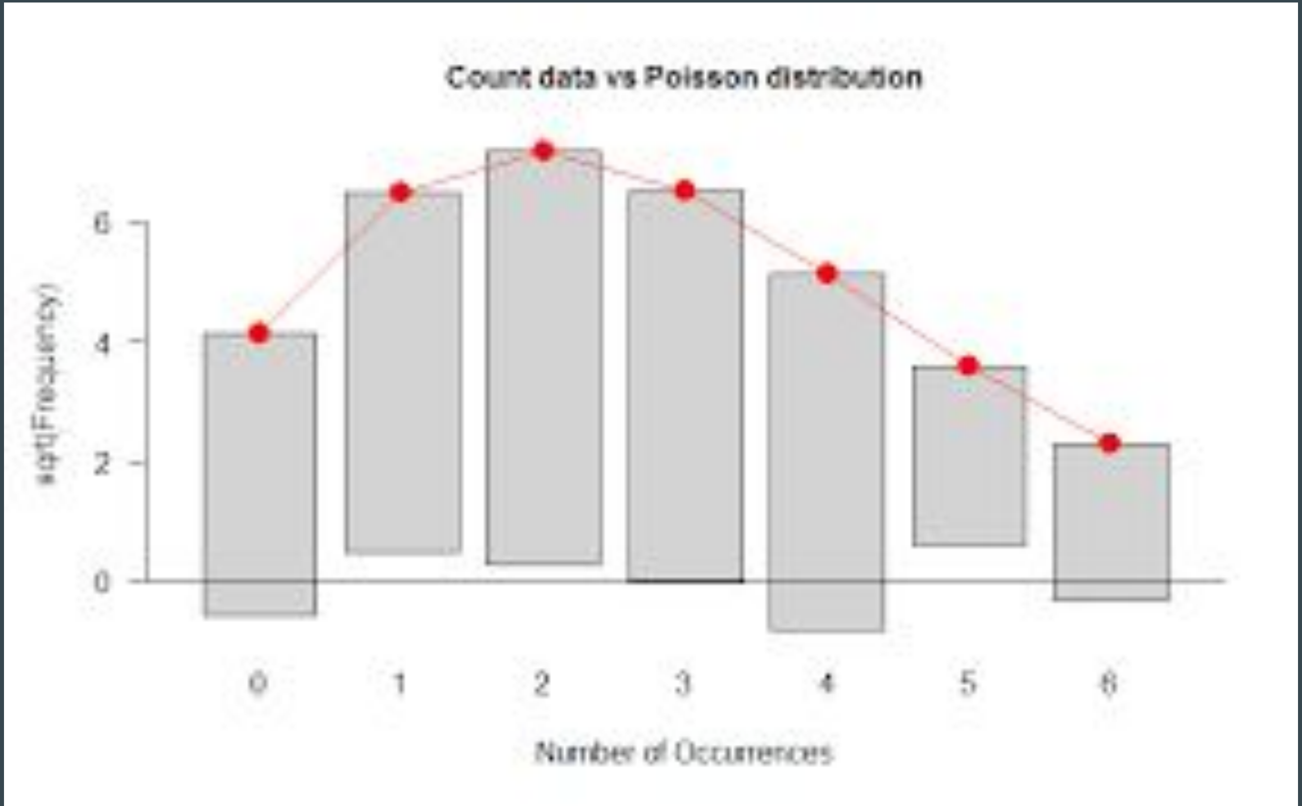
Distribution of Y

- MLE gives us a way to estimate P(Y|X)
- If Y is independent, conditional on X, to assume a distribution, we just need to look at Y
- Since we assume independence, we will just look at the distribution of Y to figure out the distribution we need to solve MLE for

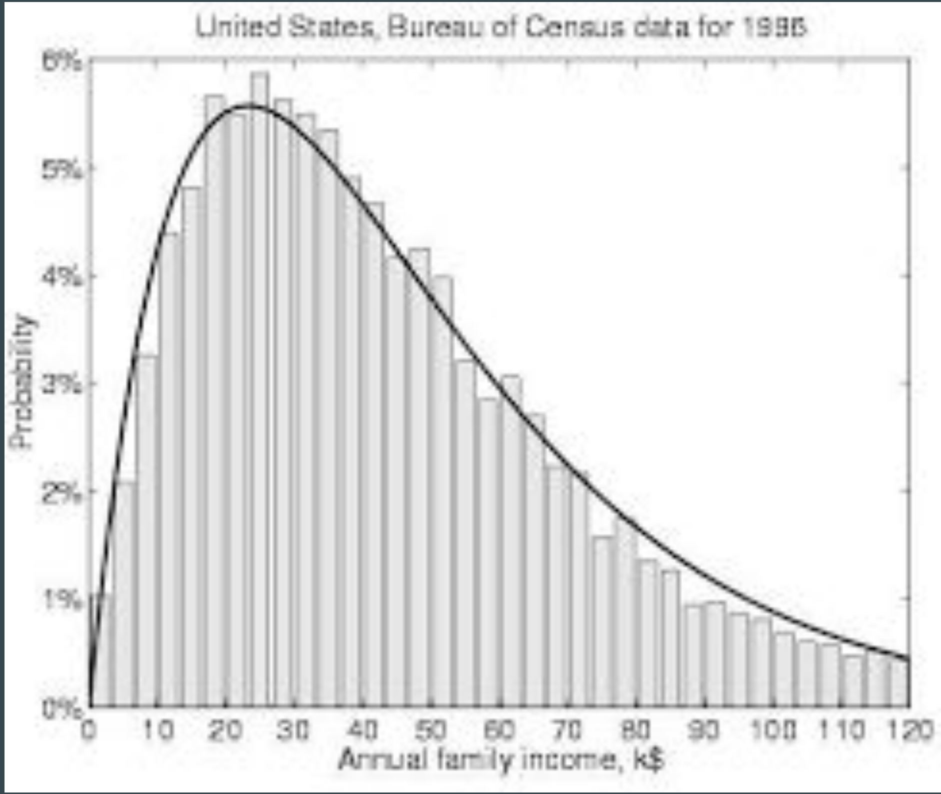
Distribution of Y (email wait times)



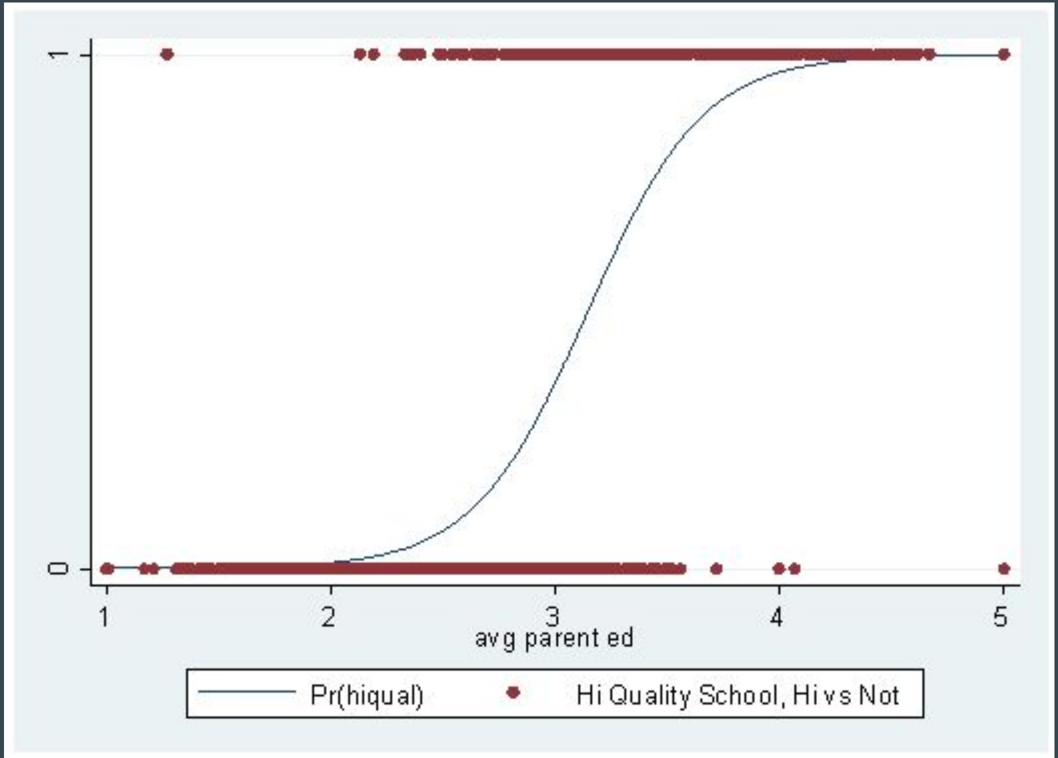
Distribution of Y (count data)



Distribution of Y (income)



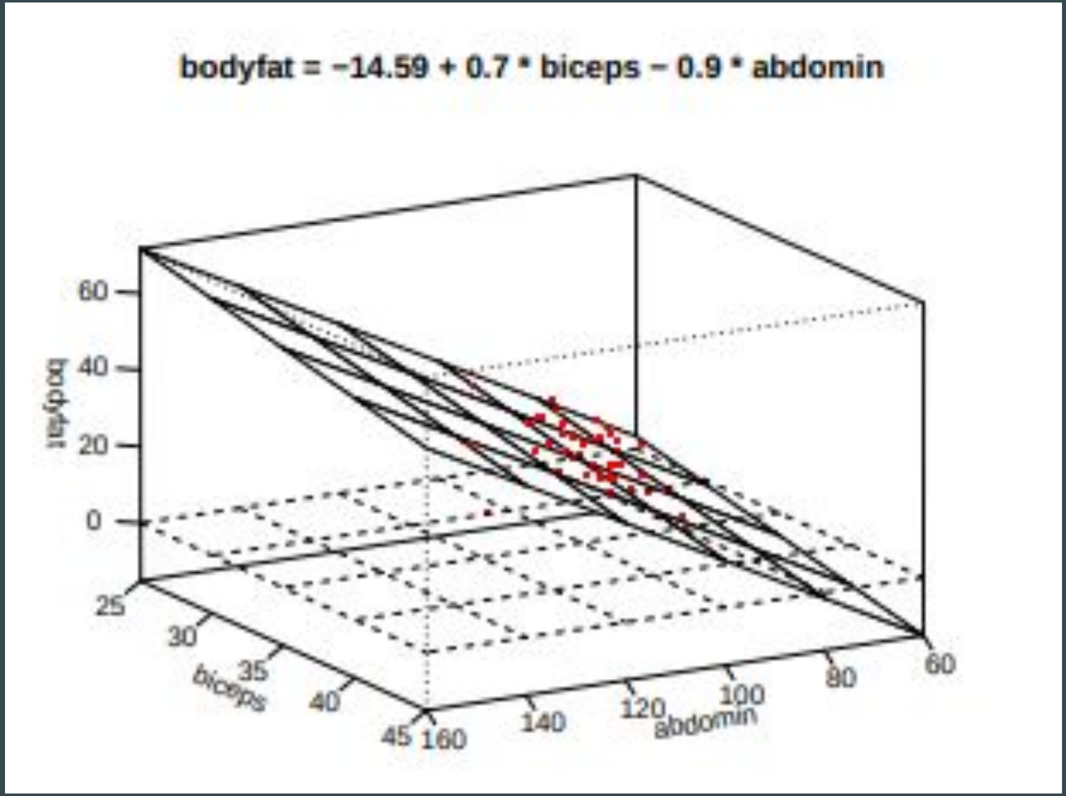
Distribution of Y (binary variable)



From general to linear model

- We have finished the first step of the GLM, detecting what type of probability distribution we want to estimate
- So how can we put this in a linear form?

How do we get a non-linear model in this form?



Anatomy of the GLM

GLM equation:

$$\eta = \alpha + \beta_1 X_1 \ldots + \beta_k X_k$$

With two additional pieces:

- **Link function:** $g(\mu)$
- **Variance function:** $V(\mu)$

So how is this different from OLS?

- The main GLM equation is equivalent to OLS with identical parameters
- What's different is the link function and variance function
  - These transform non-linear models into linear ones
  - They take a non-linear conditional probability and adjust it so it can be estimate in an additive way
- Link function address E(Y|X)
  - Aka the mean
- Variance addresses V(Y|X)

Linear GLM

GLM equation:

$$\eta = \alpha + \beta_1 X_1 \ldots + \beta_k X_k$$

With two additional pieces:

- **Link function:** $g(\mu) = \mu$
- **Variance function:** $V(\mu) = 1$
- For a linear GLM, the link function is the same as the mean, and the variance function is simply 1. Nothing needs to be changed!

Binomial GLM

GLM equation:

$$\eta = \alpha + \beta_1 X_1 \ldots + \beta_k X_k$$

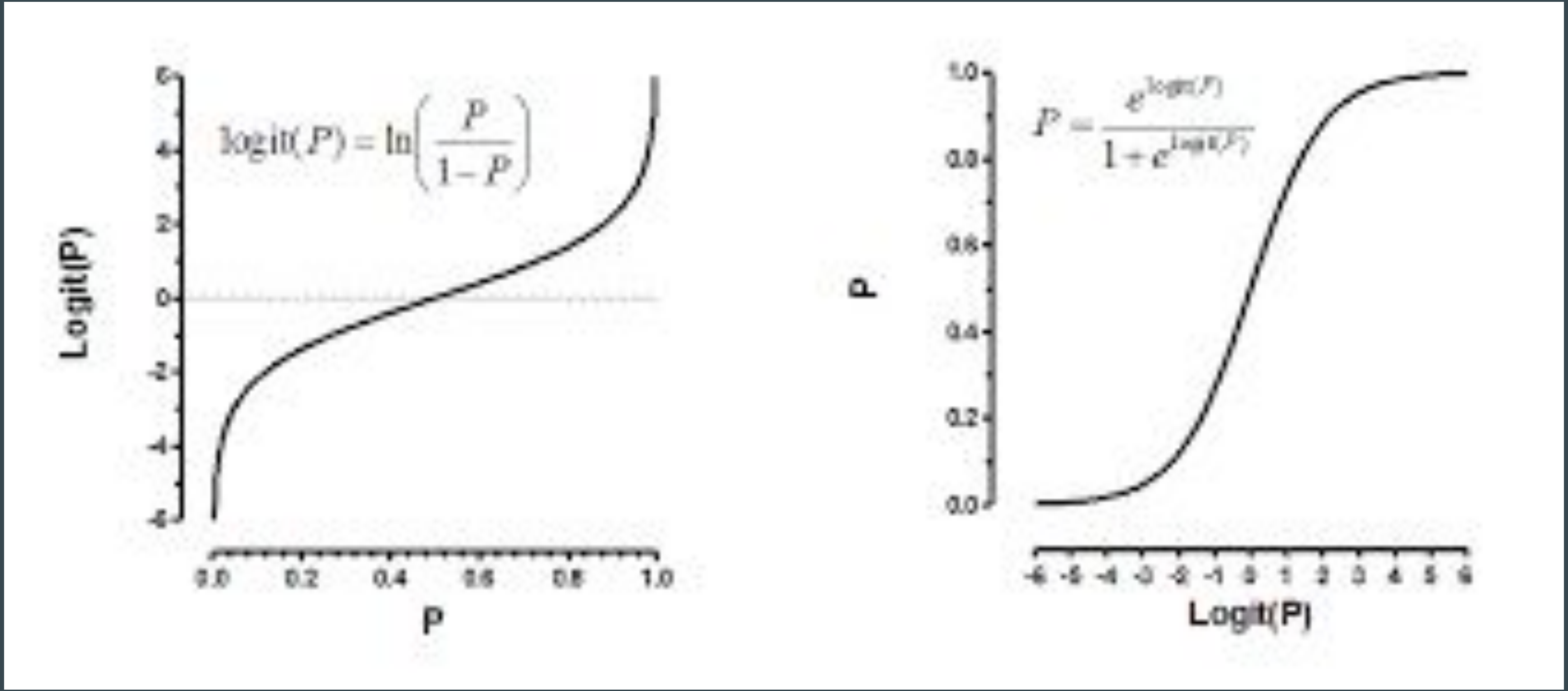
With two additional pieces:

- **Link function:** $g(\mu) = \text{logit}(\mu)$
- **Variance function:** $V(\mu) = \mu(1\text{-}\mu)$

Logit link

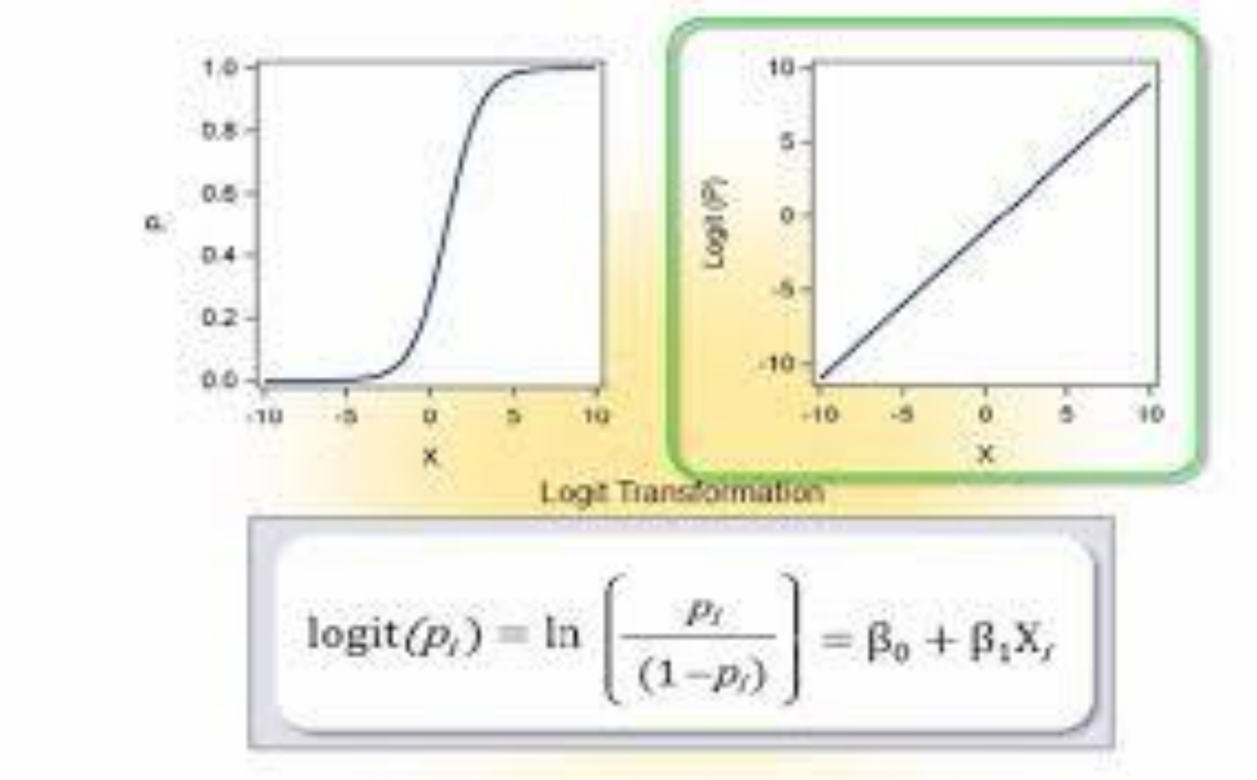
- The output of a binomial is limited from (0, 1), whereas linear regression is (-inf, inf)
- The link function transforms our conditional probability from a (0, 1) space to a (-inf, inf)

Logit link



$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$
Logit(P)
P
0.0
0.2
0.4
0.6
0.8
1.0
$P = \frac{e^{\text{logit}(P)}}{1+e^{\text{logit}(P)}}$
Logit(P)
P

Logit link



1.0
0.8
0.6
0.4
0.2
0.0
-10
-5
0
5
10
X
10
5
0
-5
-10
-10
-5
0
5
10
X
Logit Transformation
$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 X_i$

Binomial GLM

GLM equation:

$$\eta = \alpha + \beta_1 X_1 \ldots + \beta_k X_k$$

With two additional pieces:

- **Link function:** $g(\mu) = \log(\mu/1\text{-}\mu)$
- **Variance function:** $V(\mu) = \mu(1\text{-}\mu)$

Binomial GLM (example)

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	Model				
	(1) Civil War	(2) "Ethnic" War	(3) Civil War	(4) Civil War (Plus Empires)	(5) Civil War (COW)
Prior war	−0.954**	−0.849*	−0.916**	−0.688**	−0.551
	(0.314)	(0.388)	(0.312)	(0.264)	(0.374)
Per capita income[a,b]	−0.344***	−0.379***	−0.318***	−0.305***	−0.309***
	(0.072)	(0.100)	(0.071)	(0.063)	(0.079)
log(population)[a,b]	0.263***	0.389***	0.272***	0.267***	0.223**
	(0.073)	(0.110)	(0.074)	(0.069)	(0.079)
log(% mountainous)	0.219**	0.120	0.199*	0.192*	0.418***
	(0.085)	(0.106)	(0.085)	(0.082)	(0.103)
Noncontiguous state	0.443	0.481	0.426	0.798**	−0.171
	(0.274)	(0.398)	(0.272)	(0.241)	(0.328)
Oil exporter	0.858**	0.809*	0.751**	0.548*	1.269***
	(0.279)	(0.352)	(0.278)	(0.262)	(0.297)
New state	1.709***	1.777***	1.658***	1.523***	1.147**
	(0.339)	(0.415)	(0.342)	(0.332)	(0.413)
Instability[a]	0.618**	0.385	0.513*	0.548*	0.584*
	(0.235)	(0.316)	(0.242)	(0.225)	(0.268)
Democracy[a,c]	0.021	0.013			
	(0.017)	(0.022)			
Ethnic fractionalization	0.166	0.146	0.164	0.490	−0.119
	(0.373)	(0.584)	(0.368)	(0.345)	(0.396)
Religious fractionalization	0.285	1.533*	0.326		1.176*
	(0.509)	(0.724)	(0.506)		(0.563)
Anocracy[a]			0.521*		0.597*
			(0.237)		(0.261)
Democracy[a,d]			0.127		0.219
			(0.304)		(0.334)
Constant	−6.731***	−8.450***	−7.019***	−6.801***	−7.503***
	(0.736)	(1.092)	(0.751)	(0.681)	(0.854)
N	6327	5186	6327	6360	5378

Note: The dependent variable is coded "1" for country years in which a civil war began and "0" in all others. Standard errors are in parentheses. Estimations performed using Stata 7.0. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

[a] Lagged one year.
[b] In 1000's.
[c] Polity IV; varies from −10 to 10.
[d] Dichotomous.

Poisson GLM

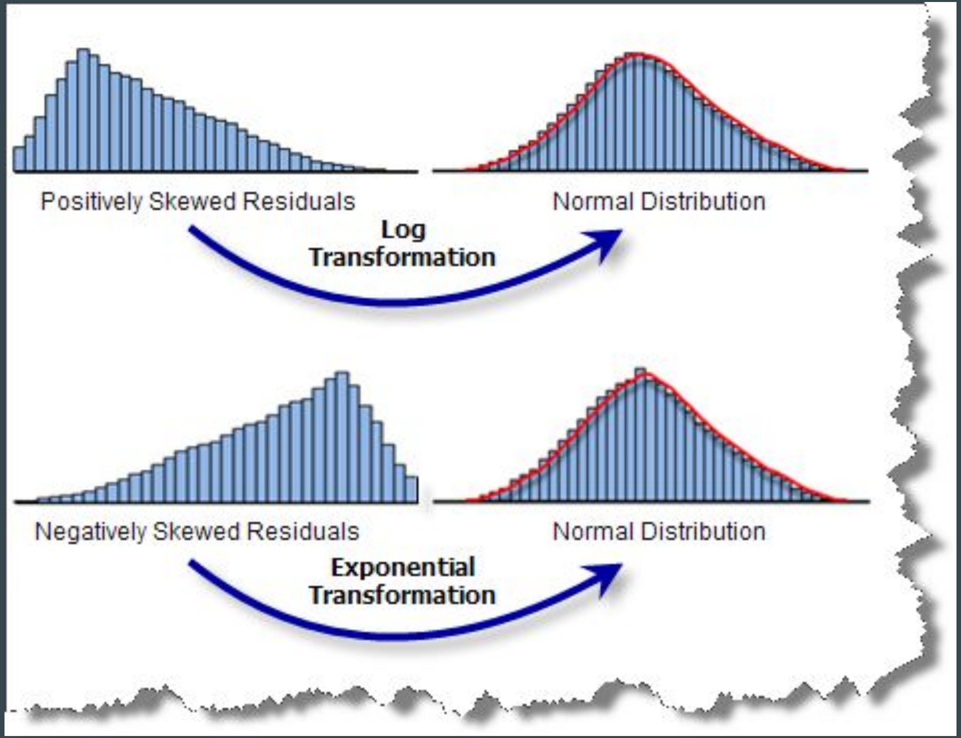
GLM equation:

$$\eta = \alpha + \beta_1 X_1 \dots + \beta_k X_k$$

With two additional pieces:

- **Link function:** $g(\mu) = \log(\mu)$
- **Variance function:** $V(\mu) = \mu$

Log link function



Poisson GLM (example)

Variable	*Parameter Estimate*	*Standard Error*	*t Ratio*	*Significance Level*
Structured versus Fair Sentencing	.2413	.0326	7.40	<.0001
Prior jail and infractions versus no prior jail	.5501	.0403	13.65	<.0001
Prior jail and no infractions versus no prior jail	.0413	.0341	1.21	<.2259
Prisoner age	−.0831	.0022	−37.77	<.0001

GLM permutations

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	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
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			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

GLM parts

We have seen that a GLM allows us to use a linear framework for non-linear relationships. There are three additional components:

1. Assume a probability distribution (sometimes called the random family)
2. Adjust conditional means (link function) to linear
3. Adjust variance to linear (variance function)

Review

- Generalized linear model allows us to use our convenient linear regression framework for non-linear relationships
- Use link functions and variance functions to alter our original distribution to one that behaves linearly
- Use MLE to estimate new parameters