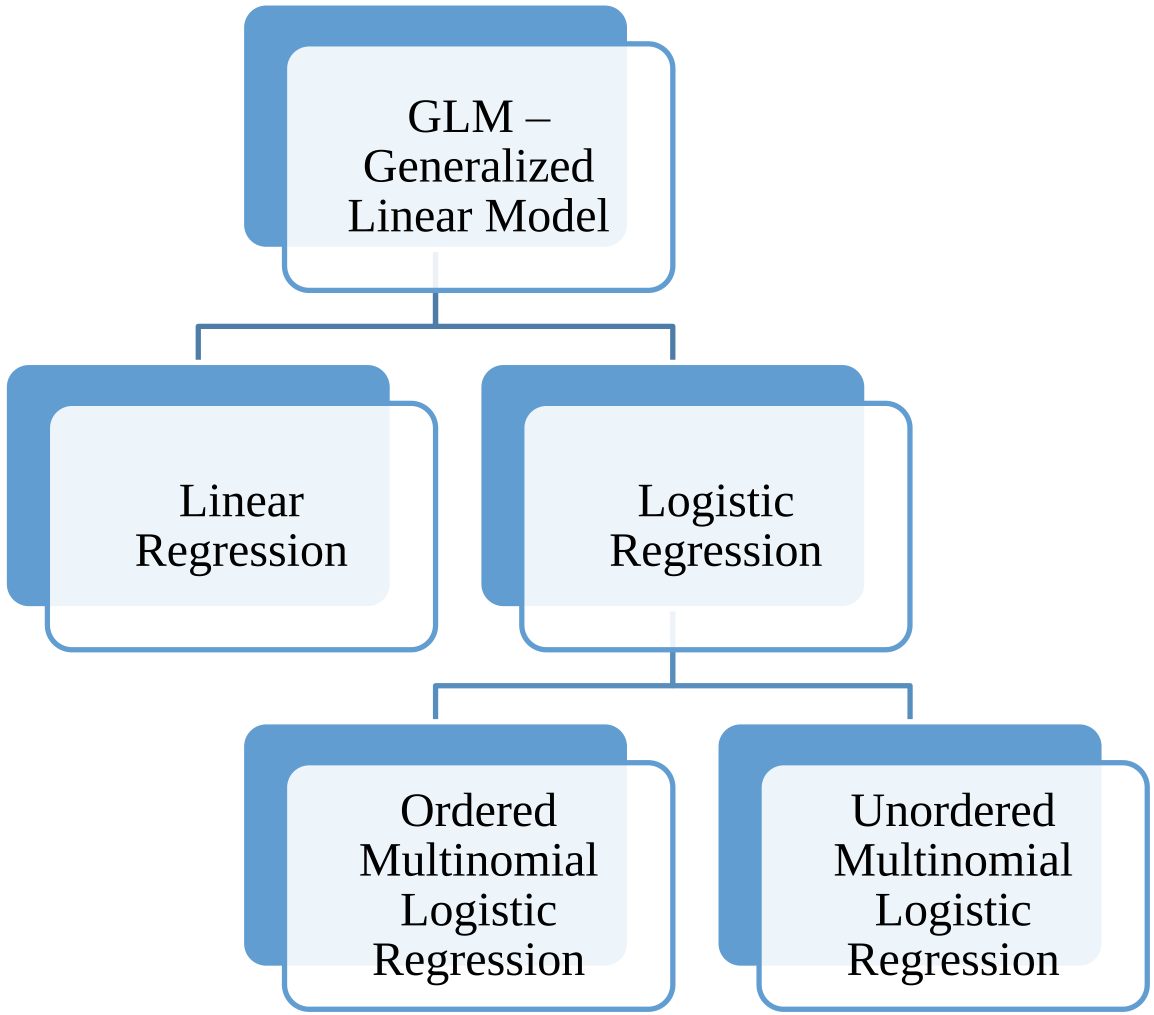
MULTINOMIAL MODELS

Ass.Prof. Andriy Stavytskyy

Agenda

- Ordered Logit/ Probit
- Multinomial Logit
- Event Count Models

Situating the Model



Four Types of Scales

- mutually exclusive categories with no logical order.
- mutually exclusive categories with logical rank order.
- ordered data with equal distance between each point (no absolute zero).
- ordered data with equal distance between each point (with a “true” zero).

Definition

- The ordered multinomial logistic model enables us to model ordinally scaled dependent variables with one or more independent variables.
- These IV(s) can take many different forms (ie. real numbers values, integers, categorical, binomial, etc.).

Does this Occur Much?

- "Ordinal data are the most frequently encountered type of data in the social sciences" (Johnson & Albert, 1999, p. 126).
- Examples
  - Yes, maybe, no
  - Likert scale (Strongly Agree – Strongly Disagree)
  - Always, frequently, sometimes, rarely, never
  - No hs diploma, hs diploma, some college, bachelor's degree, master's degree, doctoral degree
  - Free school lunch, reduced school lunch, full price lunch
  - 0-10k per year, 10-20K per year, 20-30K per year, 30 – 60K per year, > 60K per year
  - Low, medium, high
  - Basic math, regular math, pre-AP math, AP math
  - Nele's dancing ability, Meg's dancing ability, Saralyn's dancing ability, Jose's dancing ability, Kyle's dancing ability, Braden's dancing ability, a rock

ORDERED AND MULTINOMIAL LOGIT/ PROBIT

What is this for?

- extension of the logistic regresion model for binary response
- when your DV has multiple, ordered categories.

Examples:

- Bond ratings (AAA, AA, A, etc.),
- Grades (MVG, VG, G, etc.),
- opinion surveys (strongly agree, agree, disagree, strongly disagree)
- Some type of continuous outcome you might want to collapse - spending, ’performance’ (high, medium, low)
- Employment (fully, partial, unemploymed)

Assumptions of Ordered Logit Models – 1

- Maximum likelihood estimation – again, no 'sum of squares' estimation – this uses an iterative process that converges the model's log likelihood in comparison to an 'empty model' (Iteration 0)

Assumptions of Ordered Logit Models – 2

- Number of ordered responses <6. After the DV takes on 6+ values, the model can be run using OLS if distance between categories equal.

Assumptions of Ordered Logit Models – 3

- Proportional odds assumption (aka parallel regression): β's for one outcome group (low Bond rating countries) are the same as any other group (median, or high Bond rating states) – is an assumption to increase efficiancy in our estimates.

Note!

- we do NOT need to assume the distance between each interval in Y is the same! (as we would if using OLS)

Our algorithm

- we start with an observed, ordinal variable (Y)
- as in most models of estimation, Y is a function of a latent, unobserved variable Y*
- the variable Y* has ”threshold points” (’M’)– the value of Y depends on whether an observation has crossed these thresholds. If Y has 3 groups, then 2 cut-offs:

- $Y_i = 1$ if Y_i * is $\leq M_1$
- $Y_i = 2$ if M_1 is $\leq Y_i{*} \leq M_2$
- $Y_i = 3$ if Y_i * is $\geq M_2$

Estimating the model

- So, as in all statistical models we've covered, our latent variable Y* is a function of our right-hand side IV's plus some level of error:

$$Y^*{}_i = \sum_{k=1}^{k} \beta_k X_{ki} + \varepsilon_i = Z_i + \varepsilon_i$$

- Our model will estimate part of this:

$$Z_i = \sum_{k=1}^{k} \beta_k X_{ki} = E\left(Y^*{}_i\right)$$

- So Z, basically is Y* as a function of some disturbance (not a perfect measure of Y*). It is of a different scale than Y (e.g. continuous), but our estimates can give us Pr(Y=1, 2,..X) based on the value of Z.

- Like binary Logit, our link function is the log of the odds (logit), giving us odds/probability that an observation falls into a given Y category based on its levels of X's. Just like the probit and logit models, Z is continuous 0-1.

Important!

- There is no ’traditional’ intercept, just ’cut-off points’ (M) (like an intercept) & that they are different for each level of Y, but Beta’s do NOT vary for the levels of Y!

The point

- We want to estimate the probability that Y (observed variable) will take on a given value (in this case, 1, 2 or 3).
- Z helps us estimate the probability that a given observation will fall into a given Y category
- $P(Y = 1) = \frac{1}{1+exp(Z_i - M_1)}$
- $P(Y = 2) = \frac{1}{1+exp(Z_i - M_2)} - \frac{1}{1+exp(Z_i - M_1)}$
- $P(Y = 3) = 1 - \frac{1}{1+exp(Z_i - M_2)}$

Important!

- So with the estimate value of Z and the assumed logistic distribution of the error term, we can estimate the probability that an observation will fall into one of the categories of Y.

Example – 1

- The data set contains variables on 200 students. The outcome variable is prog, program type. The predictor variables are social economic status, ses, a three-level categorical variable and writing score, write, a continuous variable.

Example – 2

	id	female	ses	schtyp	prog	read	write	math	science	socst	honors	awards	cid
1	45	female	low	public	vocation	34	35	41	29	26	not enrolled	0	1
2	108	male	middle	public	general	34	33	41	36	36	not enrolled	0	1
3	15	male	high	public	vocation	39	39	44	26	42	not enrolled	0	1
4	67	male	low	public	vocation	37	37	42	33	32	not enrolled	0	1
5	153	male	middle	public	vocation	39	31	40	39	51	not enrolled	0	1
6	51	female	high	public	general	42	36	42	31	39	not enrolled	0	1
7	164	male	middle	public	vocation	31	36	46	39	46	not enrolled	0	1
8	133	male	middle	public	vocation	50	31	40	34	31	not enrolled	0	1
9	2	female	middle	public	vocation	39	41	33	42	41	not enrolled	0	1
10	53	male	middle	public	vocation	34	37	46	39	31	not enrolled	0	1

Example – 3

type of program	low	ses middle	high	Total
general	16	20	9	45
academic	19	44	42	105
vocation	12	31	7	50
Total	47	95	58	200

```
sum  ses science socst female
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ses	200	2.055	.7242914	1	3
science	200	51.85	9.900891	26	74
socst	200	52.405	10.73579	26	71
female	200	.545	.4992205	0	1

Example – 4

- Let’s say we want to estimate ’socio-economic stats’ (SES) as a function of test scores and gender

- $SES_i = \propto_{k-1} + \beta(science) + \beta(socialstudies) + \beta(female) + \epsilon_i$

- We have 200 obs in our data – let’s see how the summary stats look:

Example – 5

- We see that higher science & social science scores lead to higher SES & that females, on average, have lower SES

```
Ordered logistic regression                       Number of obs   =        200
                                                  LR chi2(3)      =      31.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -194.80235                       Pseudo R2       =     0.0749

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     science |   .0300201   .0165862     1.81   0.070    -.0024882    .0625284
       socst |   .0531819   .0152711     3.48   0.000     .0232512    .0831127
      female |  -.4823977   .2796945    -1.72   0.085    -1.030589    .0657934
```

Example – 6

- Coefficients are pretty meaningless, so, let’s calculate the PR(Y=1, 2 and 3) for a female who got average test score on both tests.

Getting our ”thresholds”

- G1 (low SES): < 2.75
- >2.75 G2 (med. SES) <5.10
- G3 (high SES): >5.10

Example – 7

- Calculating 'Zi' for a female with average test scores (from 'sum') & our Beta estimates from the last slide:

- Zi = (0.03*51.85(science) + 0.0532*52.405(soc. Sci) – 0.4824*1(female)
- Zi = 3.86
- $P(Y = 1) = \frac{1}{1+exp(Z_i - M_1)} = \frac{1}{1+\exp(3.86-2.755)} = .249$

- $P(Y = 2) = \frac{1}{1+exp(Z_i - M_2)} - \frac{1}{1+exp(Z_i - M_1)} = \frac{1}{1+\exp(3.86-5.105)} - \frac{1}{1+\exp(3.86-2.755)} = .528$

- $P(Y = 3) = 1 - \frac{1}{1+exp(Z_i - M_2)} = 1 - \frac{1}{1+\exp(3.86-5.105)} = .223$

- **Total should add up to 1**

Example – 8

- So, a female with average test scores has a 24.9%, 52.8% and 22.3% probability of being in the low, medium and high levels of SES respectively!

Model diagnostics

- Just like with logit, here we have similar tests for ’goodness of fit
- Use the LR χ^2 statistic (& p-value) to test if all coefficients in the model $\neq 0$
- You can test nested models (omitted variables) with the LR test
- Can use a Chow test to check for structural breaks (sub-groups)

Note!

- In small samples, (say under 50 or so), you will often violate the Proportional/paralell odds assumption because outlying obesrvations will have a large impact on the model

- In this case, the estimates will be biased.

- To remedy this, you can use GENERALIZED LEAST SQUARES estimates

MULTINOMIAL LOGIT

Multinomial Logit

- Similar to ordered logit, when our DV takes on 2+ values, but still limited – 3, 4, 5 categories for example.
- Unlike ordered logit, the categories of the DV are ’not ordered’, but are nominal categories (aka ’categorical’).
- We are interested in the relative probability of these outcomes using a common set of parameters (IV’s)

- For example - given a set of IV’s (education, country/regional origin, parent’s income, rural/urban) we might want to know the following:

- Choice of a foreign language – English, Spanish, Chinese, Swedish
- Choice of drink: coffee, Coke, juice, wine
- Choice of occupation – police, teacher, or health care worker
- Mode of transportation – car, bus, tram, train
- Voting for a party or bloc – R-G, Alliansen or S.D.

Assumptions of ’mlogit’ models

- A common set of parameters (IV’s) can linearly predict probabilities of DV categorical outcomes, but do not assume error term is constant across Y outcomes.
- Unlike Ologit, these IV’s are CASE SPECIFIC – have independent effects on each category of the DV (e.g. different Betas across categories – no ’parallel odds assumption’).
- ”Independence of Irrelevant Alternatives” (IIA, from Arrow’s ’impossibility theorom) – the odds/probability of chosing one case of the DV over another does not depend on another’s presence or absence, ’irrelevant alternatives’ **strong assumption**
- **Multinomial logit is not appropriate if the assumption is violated.

Multinomial Logit Assumption 2 Examples

- IIA Example 1: Voting for certain parties
- **For ex., the probabilities of someone S, V, L, M, KD or, S.D. vs. M does not change if MP is added or taken away
  - Is IIA assumption likely met in this election model?
  - Probably not. If MP were removed, those voters would likely vote for V or S.
    - Removal of MP would increase likleyhood for S or V relative to M

- IIA Example 2: Consumer Preferences
  - Options: coffee, juice, wine, Coke
    - Might meet IIA assumption
  - Options: coffee, juice, Coke, Pepsi
    - Won't meet IIA assumption. Coke & Pepsi are very similar – substitutable.
    - Removal of Pepsi will drastically change odds ratios for coke vs. others.

Long and Freese (2006):

- "Multinomial and conditional logit models should only be used in cases where the alternatives "can plausibly be assumed to be distinct and weighed independently in the eyes of the decision-maker."
- Categories should be "distinct alternatives", not substitutes. Theory & argument very important
- Note: There are some formal tests for violation of IIA. But they don't always work well. Be cautious of them.

Diagnositics with MLogit

- Again, like logit (and ologit), we test the signficance of the full model with the χ^2 statistic, and 'improvements' (or omitted/ irellevant variables) with an LR test using the log likelihood ratios.
- Again, Pseudeo-R2 is meaningless by itself – only compared to other models with the same sample. BUT, the higher, the better.

EVENT COUNT MODELS

Description

- Again, we determine the use of an Event Count model by the structure of our DV
- So far, we've looked at variables that have normal and binary distributions (OLS, and Logit). We'll now consider a 3rd type, 'Gamma' distributions
- In this case, the DV is:
- a FIXED number of outcomes & NOT binary
- For ex., can be units of time (days, years, etc), units in fixed time (individual or geographic unit)
- Ordinal (more later if your DV is continuous)
- Positive (but can take '0')

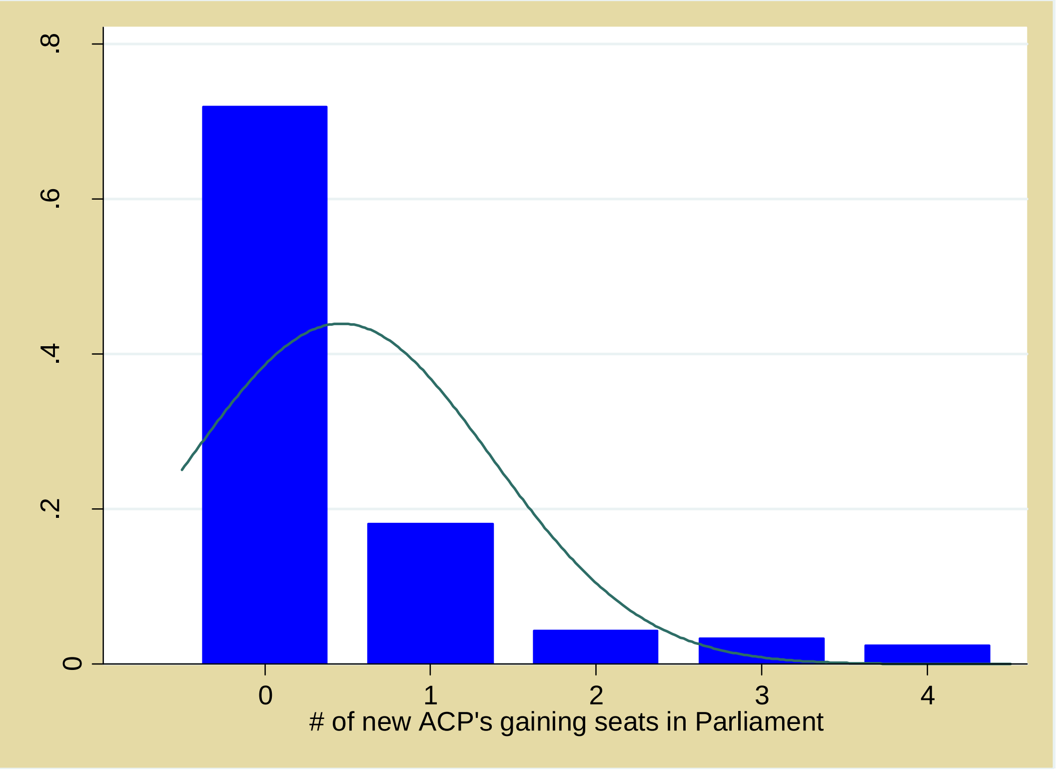
Some examples

- Number of new political parties entering parliament in a given election year
- The number of political protests or coup d’Etats in a country-year
- Number of presidential vetos in a year or mandate period
- Number of children in a household
- Number of vaccinations a child gets in a year, or doctor visits an adult makes
- Number of civic organizations an individual joins or is a member of in a given year.

Key characteristics of ’Event Data’ – 1

- The count of events is non-negative
- are independent of one another
- Counts must be integers (e.g. discrete) – cannot be 2.2, 3.7 but 2 or 4.
- Can have 1-parameter (λ) distribution (mean=VAR)
- Using a histogram, we see that the distribution of Yi outcomes is usually large in 0 or 1, and diminishes rapidly from the 2nd or 3rd outcome on
- The distribution is thus NOT normal (’Gausian’)– it is a ’gamma distribution: for count data we use these models:
- 1.Poisson
- 2. negativel binomal

Key characteristics of ’Event Data’ – 2



Poisson Models: Assumptions & workings

- Like logit, estimates with Maximum Likelihood estimation (MLE), which finds the value of the parameter that fits the model 'best' (log likelihood)

- Our "link function" in this case is Lambda – λ

- Goals are to:

- 1) estimate the increase Pr(Y=n) for a unit change in X. In Poisson regression, the model expresses the log outcome rate as a linear function of a set of predictors. (like Logit, β's need to be transformed for interpretation)

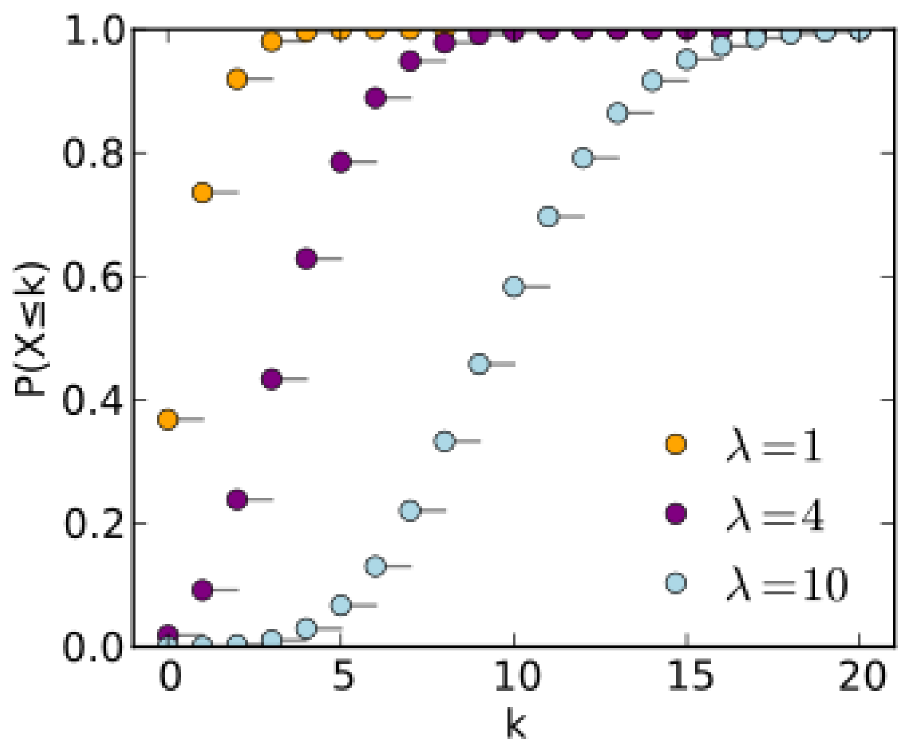
- 2) predict the expected count-outcome (group) for an observation (like ologit). But because of our DV distribution, the normal/logit curve can't be used, thus the Gamma distribution fills this gap.

Why better than OLS??

- OLS will produce a linear estimate of the relationship between βX and Y that will be less than 0 and greater than our highest count (unrealistic predictions).
- OLS assumes the difference is the same between all counts in Y (0 to 1 is the same as 3 to 4), like Ologit, Poisson does not.
- we will almost always have heteroskadasticity (as there will probably be more VAR in Y-outcomes with more observations)
- error term is not normally distributed

The Poisson distribution

- $\Pr(Y_i = \text{k}) = \frac{\lambda^k e^{-\lambda}}{k!}$
- λ is calculated as the mean of Yi
- $e^{-\lambda}$ is equal to the exponent inverse of Lambda
- K is the number of outcomes in Y
- K! is the factorial of K (ex. 4! = 4 × 3 × 2 × 1 = 24)
- $\boldsymbol{\lambda}$ is the expected value of Yi (mean of DV) and also its variance:



So:
λ= E(Y) = Var(Y)
Notice when λ =1 the CDF is highly concentrated between 0 and 10, as Lamda increases, what does the CDF look like?

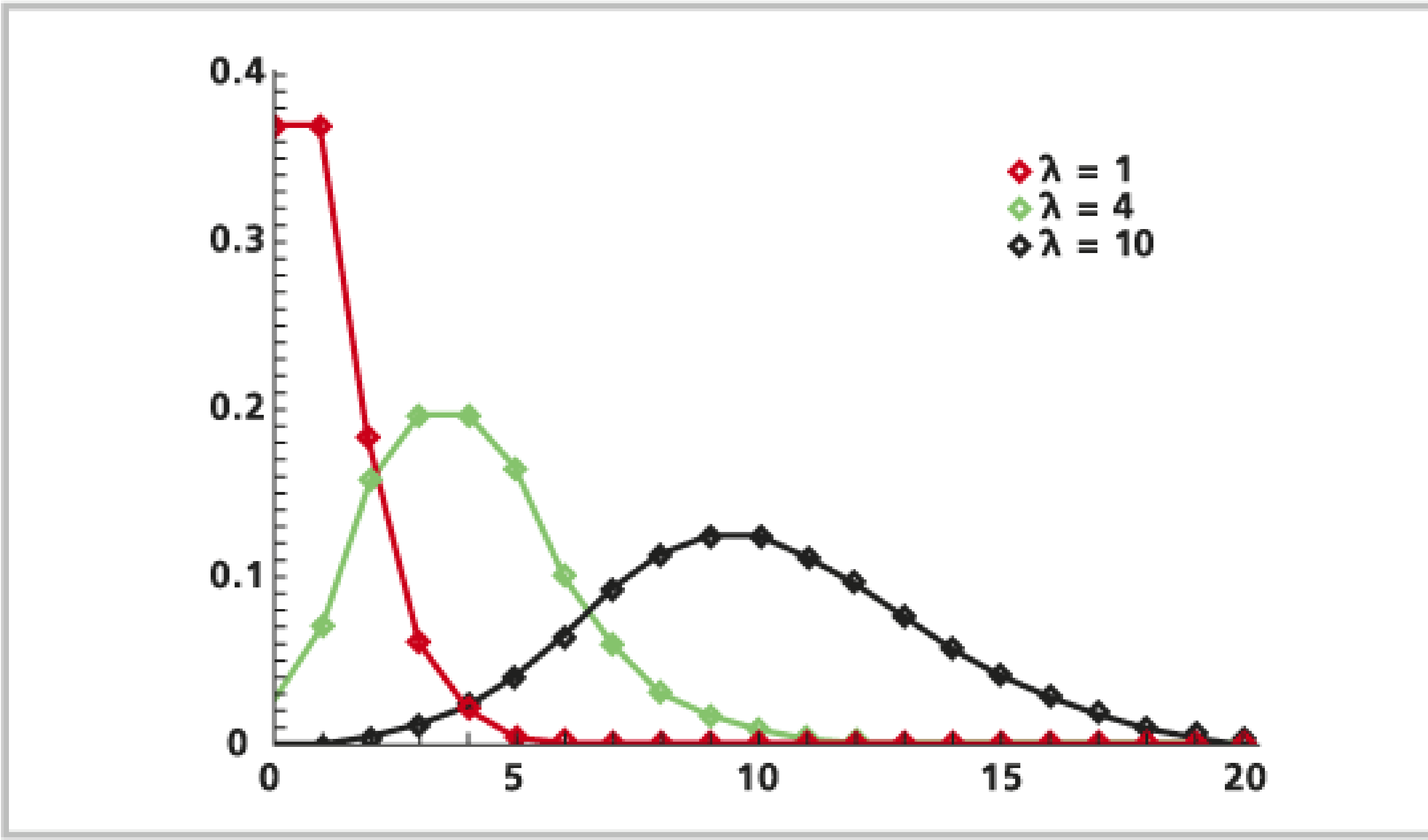
Poisson distributions at different levels of Lambda – 1

- λ is equal to rate of the event (DV)

- So, if the mean of the distribution (λ) is high enough, than OLS is ok. So we can generate Pr(Y=n|Xi) in a similar way as a normal curve – e.g. Mean approaches 10

- BUT the data we will discuss will have a mean closer to about 1 or less
- 3 examples with K=20 & λ=1, 4 & 10

Poisson distributions at different levels of Lambda – 2



Important assumptions of a Poisson Model

- The observations are assumed to be independent of one another
- Logarithm of rate changes in the DV are expressed linearly with equal increment increases in the IV's
- ”Equidispersion” – e.g., the mean of the DV = the Variance (although this does not happen that very often).
- Breaking this is called ”overdispearsion” – when VAR in our data is greater than the model assumes. If violated, we can't use Poisson for hypothesis testing.
- **If outcome cases of Y are not independent, then we will mostly likely see ”overdispersion” – which if large enough, will lead us to use a Negative Binomial model (more later…)

Overdispersion: Causes & Consequences

- Possible causes:

1. a poorly fitted model
   - Omited variables
   - Outliers
   - Wrong functional form of 1+ of our IV’s in the model
   - Unaccounted heteroskadescticity from structural breaks.

2. $VAR(Y_i) > \mu_i$ (variance of our data greater than the mean)

   -very common with individual level data!

- Consequences:
  - Underestimated SE’s (think opposite effect of multicollinearity)
  - Overstimated p-values & poor prediections

Important extra model test in Poisson

- Before going on to interpret the model's Betas, we need to know whether we've 'chosen correctly' with Poisson – does the Poisson estimation form fit our data?? E.g. is the Gamma distribution appropriate?

- Otherwise, we might consider ologit
- A 'goodness of fit' test (χ^2) will let us know if we have a problem from – the H_0 is the the model's form DOES fit our data, a rejection of H_0 means that Poisson might be the WRONG estimation.
- Other reasons for rejection would be omitted IV's or incorrect functional forms

Time to interpret

- Like logit, the Betas are basically meaningless, but - Poisson can give us Odds ratio (IRR), or 'incident rate ratio' = exponentiated Betas (like logit)
- $\frac{\lambda|X_{program=academic}}{\lambda|X_{program=general}} = \exp(\beta X_{program})$
  $= \exp(1.08) = 2.95$
- Ex., holding math score constant, a student in an academic program (compared with general) has 2.95 times the incident rate
- Also, we see that for every increase in one unit in a math score (e.g. '1'), the percent change in the incident rate increases by 7%, holding program constant

obs	200
wald Chi2	80.15
pr>Chi2	0.000
Psuedo R2	0.2118

no. of Awards	Beta	robsut s.e.	IRR
program (comparison=general)			
academic	1.08	0.32	2.956
vocational	0.369	0.401	1.447
math score	0.07	0.01	1.07
const.	-5.24	0.65	

Negative Binomial Models (NBM) – 1

- Are also ”count” models for limited DV’s, very similar to Poisson in both assumptions and interpretation
- Uses a version of Lambda as a link function to estimate Pr(Y) as well
- Key difference from Poisson is that the Var(Y) is assumed to be larger than the Mean(Y) (e.g. ’overdispersion’).
- Also, if we cannot assume that the outcomes of Y are independent from one another, than a NBM might be more appropriate
- A matter of efficiency: we prefer Poisson becasue of greater efficiency, but there is a clear solution when we violate key model assumptions, so we take NBM instead.

Negative Binomial Models (NBM) – 2

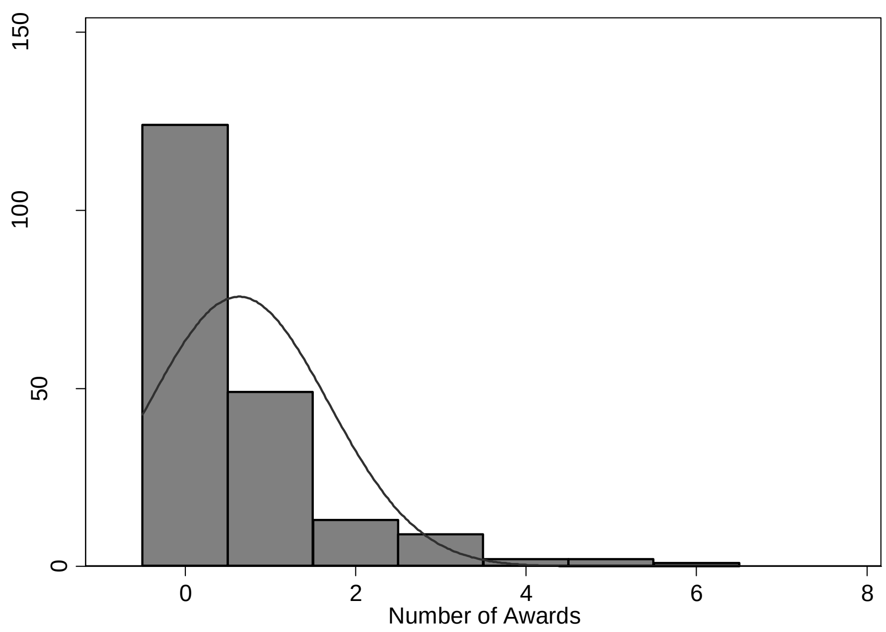
- Like Poisson, the NBM assumes constant variance in Y, which is estimated by maximum likelihood as:
- Var(Y) = $\lambda + \lambda^{2}/\boldsymbol{\alpha}$
- $\boldsymbol{\alpha} =$ the 'dispearsion parameter' (set at '0' in Poisson), so instead of one parameter being estimated, there are 2 (which is why less 'efficient')
- Uses logged Betas, so like logit (& Poisson) can use Odds ratios

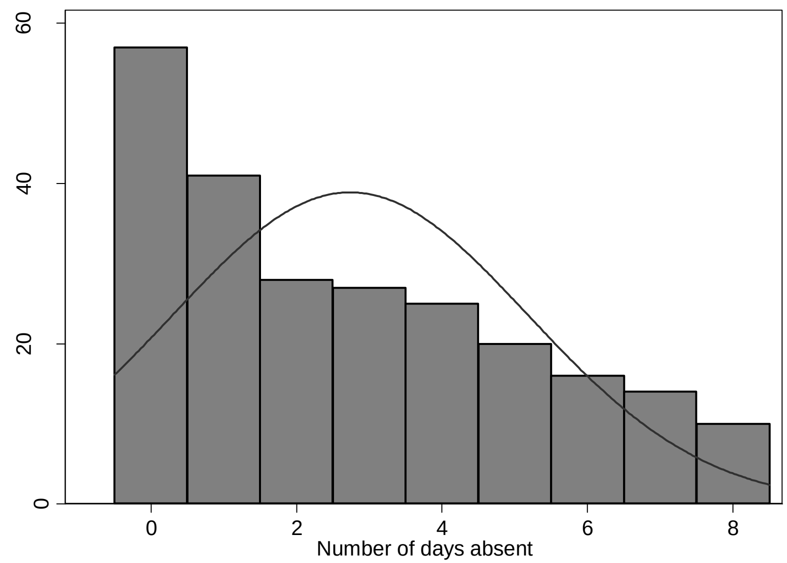
- So, NBM's are basically a more general type of Poisson model.

Key differences

- Because of the quadradic function in the assumed Var(Y), they are LESS EFFICIENT – Poisson will produce SMALLER s.e.'s for beta estimates, in med-large samples, the estimates are consistant (not-biased) however.
- Following, NBM's will result in larger expected probabilities for smaller counts (e.g. # of Yi outcomes) compared with Poisson
- NBM's will have slightly larger probabilities for larger counts

Example: common Poisson vs. Negative binomial distributions





NBM vs. Poisson for our example

DV=Absences	Negative Binomial beta	Negative Binomial s.e.	Poisson beta	Poisson s.e.
math	-0.0045	0.0025	-0.0049	0.0016
Baseline=general				
Academic	-0.558	0.192	-0.554	0.109
Vocational	-0.956	0.199	-0.958	0.120
constant	1.85	0.212	1.87	0.121

See how close the Betas are?

This shows that Poisson is still a **consistent estimator,** dispite overdispersion

However, what is the difference here?

Yes, s.e.'s considerably larger in NBM, leads to higher Z-scores in Poisson and maybe greater type-1 error

REVIEW

Summary review

- Sometimes, our DV's will have a limited distribution: 0/1, 0-4, 1-5, categorical responses, etc.

- This results in many problems for OLS, such as heterogeneity of the error term, which gives biased and and unrealistic estimation for our betas.

- Like in OLS, we want to make predictions about Pr(Y) given values of Xi, etc., but we need to transform our Y's to probabilities, odds, etc. using LINK FUNCTIONS.

- For binary variables, our link functions can be logit or probit. Same for ordinal or categorical data.

- For count data, we take advantage of gamma distributions, and use Lamba as our link function (for Poission and NBM)

- Remember, none of the betas produced make intuative sense, and thus they need to be transformed (odds, pr, etc.) margins.

- Also, the choice of any of these models is based on your Dep. Variable!!

QUESTIONS?

THANK YOU FOR YOUR ATTENTION!