Introduction to Generalized Univariate Models: Models for Binary Outcomes

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #7

THE UNIVERSITY OF KANSAS

In This Lecture…

- A short review for maximum likelihood
- Expanding your linear models knowledge to models for outcomes that are **not** conditionally normally distributed
  - A class of models called Generalized Linear Models
- A furthering of our Maximum Likelihood discussion: how knowledge of distributions and likelihood functions makes virtually any type of model possible (in theory)
- An example of generalized models for binary data using logistic regression

REVIEWING MAXIMUM LIKELIHOOD

Properties of Maximum Likelihood Estimators

- Provided several assumptions ("regularity conditions") are met, maximum likelihood estimators have good statistical properties:

1. **Asymptotic Consistency:** as the sample size increases, the estimator converges in probability to its true value

2. **Asymptotic Normality:** as the sample size increases, the distribution of the estimator is normal (with variance given by "information" matrix)

3. **Efficiency:** No other estimator will have a smaller standard error

- Because they have such nice and well understood properties, MLEs are commonly used in statistical estimation

THE UNIVERSITY OF KANSAS

Things Involved in Maximum Likelihood Estimation

- **(Marginal) Likelihood/Probability Density Functions**:
  - The assumed distribution of one observation's data – following some type of probability density function that maps the *sample space* onto a likelihood
  - The outcome can come from any distribution

- **(Joint) Likelihood Function**:
  - The combination of the marginal likelihood functions (by a product when independence of observations is assumed)
  - Serves as the basis for finding the unknown parameters that find the maximum point

- **Log-Likelihood Function:**
  - The natural log of the joint likelihood function, used to make the function easier to work with statistically and computationally
  - Typically the function used to find the unknown parameters of the model

- **Function Optimization (finding the maximum):**
  - Initial values of the unknown parameters are selected and the log likelihood is calculated
  - New values are then found (typically using an efficient search mechanism like Newton Raphson) and the log likelihood is calculated again
  - If the change in log likelihoods is small, the algorithm stops (found the maximum); if not, the algorithm continues for another iteration of new parameter guesses

Once the Maximum Is Found…

- **Distribution of the Parameters:**
  - As sample size gets large, the parameters of the model follow a normal distribution (note, this is NOT the outcome)

- **Standard Errors of Parameters:**
  - The standard errors of parameters are found by calculating the ***information matrix***, which results from the matrix of second derivatives evaluated at the maximum value of the log likelihood function
  - The asymptotic covariance matrix of the parameters comes from -1 times the inverse of the information matrix (contains variances of parameters along the diagonal)
  - The standard error for each parameter is the square root of the variances
  - The variances and covariances of the parameters are used in calculating linear combinations of the parameters, as in the glht() function of the multcomp package in R

Once the Maximum is Found…

- **Likelihood Ratio/Deviance Tests:**
  - -2 times the log likelihood (at the maximum) provides what is often called a deviance statistic
  - Nested models are compared using differences in -2*log likelihood, which follows a Chi-Square distribution with DF = difference in number of parameters between models
  - Some software reports -2 log likelihood but some reports only the log likelihood
  - Sometimes the anova() function does this test for you

- **Wald Tests**:
  - (1 degree of freedom) Wald tests are typically formed by taking a parameter and dividing it by its standard error
  - Typically these are used to evaluate fixed effects for ML estimates of GLMs

- **Information Criteria**
  - The information criteria are used to select from non-nested models
  - The model with the lowest value on a given criterion (i.e., AIC, BIC) is the preferred model
  - This is not a hypothesis test: no p-values are given
  - These aren't used when models are nested (use likelihood ratio/deviance tests)

AN INTRODUCTION TO GENERALIZED MODELS

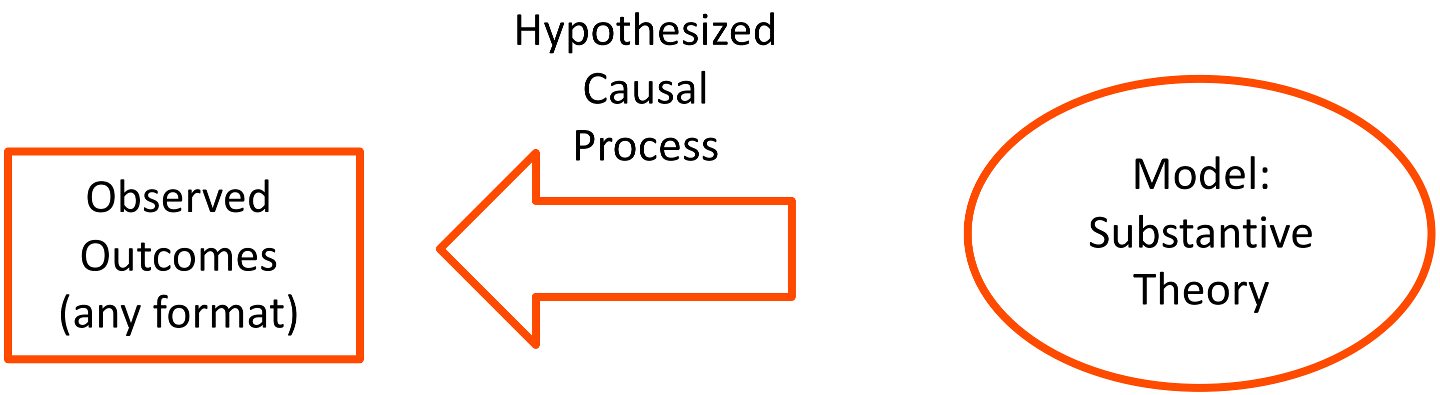
A World View of Models

- Statistical models can be broadly organized as:
  - General (normal outcome) vs. General*ized* (not normal outcome)
  - One dimension of sampling (one variance term per outcome) vs. multiple dimensions of sampling (multiple variance terms)
    - Fixed effects only vs. mixed (fixed and random effects = multilevel)

- All models have **fixed effects**, and then:
  - General Linear Models: conditionally normal distribution for data, fixed effects, no random effects
  - General Linear **Mixed** Models: conditionally normal distribution for data, fixed **and random effects**
  - General***ized*** Linear Models: **any conditional distribution for data**, fixed effects through **link functions**, no random effects
  - General***ized*** Linear **Mixed** Models: **any conditional distribution for data,** fixed **and random effects** through **link functions**

- **"Linear"** means the fixed effects predict the *link-transformed* DV in a linear combination of (effect*predictor) + (effect*predictor)…

Unpacking the Big Picture



- Substantive theory: what guides your study
- Hypothetical causal process: what the statistical model is testing (attempting to falsify) when estimated
- Observed outcomes: what you collect and evaluate based on your theory
  - Outcomes can take many forms:
    - Continuous variables (e.g., time, blood pressure, height)
    - Categorical variables (e.g., likert-type responses, ordered categories, nominal categories)
    - Combinations of continuous and categorical (e.g., either 0 or some other continuous number)

The Goal of Generalized Models

- Generalized models map the substantive theory onto the **sample space** of the observed outcomes
  - **Sample space** = type/range/outcomes that are possible

- The general idea is that the statistical model will not approximate the outcome well if the assumed distribution is not a good fit to the sample space of the outcome
  - If model does not fit the outcome, the findings cannot be believed

- The key to making all of this work is the use of differing statistical distributions for the outcome

- Generalized models allow for different distributions for outcomes
  - The mean of the distribution is still modeled by the model for the means (the fixed effects)
  - The variance of the distribution may or may not be modeled (some distributions don't have variance terms)

What kind of outcome? *Generalized* vs. General

- **Genera*lized* Linear Models** → General Linear Models whose residuals follow some not-normal distribution and in which a link-transformed Y is predicted instead of Y

- Many kinds of non-normally distributed outcomes have some kind of generalized linear model to go with them:
  - Binary (dichotomous)
  - Unordered categorical (nominal)
  - Ordered categorical (ordinal)
  - Counts (discrete, positive values)
  - Censored (piled up and cut off at one end – left or right)
  - Zero-inflated (pile of 0's, then some distribution after)
  - Continuous but skewed data (pile on one end, long tail)

These two are often called "multinomial" inconsistently

THE UNIVERSITY OF KANSAS

Some Links/Distributions (from Wikipedia)

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	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	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian			Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $[0, +\infty)$	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: $[0, N]$	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			

THE UNIVERSITY OF KANSAS

3 Parts of a General*ized* Linear Model

- Link Function (main difference from GLM):
  - How a non-normal **outcome gets transformed** into something we can predict that is more continuous (unbounded)
  - For outcomes that are already normal, general linear models are just a special case with an "identity" link function (Y * 1)

- Model for the Means ("Structural Model"):
  - How predictors **linearly** relate to the link-transformed outcome
  - **New link-transformed $Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$**

- Model for the Variance ("Sampling/Stochastic Model"):
  - If the errors aren't normally distributed, then what are they?
  - Family of alternative distributions at our disposal that map onto what the distribution of errors could possibly look like

Link Functions: How Generalized Models Work

- Generalized models work by providing a mapping of the theoretical portion of the model (the right hand side of the equation) to the sample space of the outcome (the left hand side of the equation)
  - The mapping is done by a feature called a link function

- The link function is a non-linear function that takes the linear model predictors, random/latent terms, and constants and puts them onto the space of the outcome observed variables

- Link functions are typically expressed for the mean of the outcome variable (we will only focus on that)
  - In generalized models, the variance is often a function of the mean

Link Functions in Practice

- The link function expresses the conditional value of the mean of the outcome $E(Y_p) = \hat{Y}_p = \mu_y$ (E stands for expectation)…

- …through a (typically) non-linear **link function** $g(\cdot)$ (when used on conditional mean); or its inverse $g^{-1}(\cdot)$ when used on predictors…

- …of the observed predictors (and their regression weights):

$$\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$$

- Meaning:

$$E(Y_p) = \hat{Y}_p = \mu_y = g^{-1}(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p)$$

- The term $\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$ is called the **linear predictor**
  - Within the function, the values are linear combinations
  - Model for the means (fixed effects)

Normal GLMs in a Generalized Model Context

- Our familiar general linear model is actually a member of the generalized model family (it is **subsumed**)
  - The link function is called the identity, the linear predictor is unchanged

- The normal distribution has two parameters, a mean μ and a variance σ^2
  - Unlike most distributions, the normal distribution parameters are directly modeled by the GLM

- The expected value of an outcome from the GLM was

$$E(Y_p) = \hat{Y}_p = \mu_y = g^{-1}(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p)$$
$$= \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$$

- In conditionally normal GLMs, the inverse link function is called the identity:

$$g^{-1}(\cdot) = 1 * (\text{linear predictor})$$

  - The identity does not alter the predicted values – they can be any real number
  - This matches the sample space of the normal distribution – the mean can be any real number

And…About the Variance

- The other parameter of the normal distribution described the variance of an outcome – called the error variance

- We found that the model for the variance for the GLM was:

$$V(Y_p) = V(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p) = V(e_p) = \sigma_e^2$$

- Similarly, this term directly relates to the variance of the outcome in the normal distribution
    - We will quickly see distributions where this doesn't happen

GENERALIZED LINEAR MODELS FOR BINARY DATA

Today’s Data Example

- To help demonstrate generalized models for binary data, we borrow from an example listed on the UCLA ATS website:

  https://stats.idre.ucla.edu/stata/dae/ordered-logistic-regression/

- Data come from a survey of 400 college juniors looking at factors that influence the decision to apply to graduate school:
  - Y (outcome): student rating of likelihood he/she will apply to grad school – (0 = unlikely; 1 = somewhat likely; 2 = very likely)
    - We will first look at Y for two categories (0 = unlikely; 1 = somewhat or very likely) - this is to introduce the topic for you **Y is a binary outcome**
    - You wouldn’t do this in practice (use a different distribution for 3 categories)
  - ParentEd: indicator (0/1) if one or more parent has graduate degree
  - Public: indicator (0/1) if student attends a public university
  - GPA: grade point average on 4 point scale (4.0 = perfect)

Descriptive Statistics for Data

Analysis Variable : GPA				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
400	2.998925	0.3979409	1.9	4

Likelihood of Applying (1 = likely)				
Lapply	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	220	55	220	55
1	180	45	400	100

APPLY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	220	55	220	55
1	140	35	360	90
2	40	10	400	100

Parent Has Graduate Degree				
parentGD	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	337	84.25	337	84.25
1	63	15.75	400	100

Student Attends Public University				
PUBLIC	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	343	85.75	343	85.75
1	57	14.25	400	100

What If We Used a Normal GLM for Binary Outcomes?

- If Y_p is a binary (0 or 1) outcome…
  - Expected mean is proportion of people who have a 1 (or "p", the probability of $Y_p = 1$ in the sample)
  - The **probability of having a 1** is what we're trying to predict for each person, given the values of his/her predictors
  - General linear model: $\mathbf{Y_p = \beta_0 + \beta_1 x_p + \beta_2 z_p + e_p}$
    - β_0 = expected probability when all predictors are 0
    - βs = expected change in probability for a one-unit change in the predictor
    - e_p = difference between observed and predicted values
  - Model becomes Y_p = **(predicted probability of 1) +** $\mathbf{e_p}$

A General Linear Model Predicting Binary Outcomes?

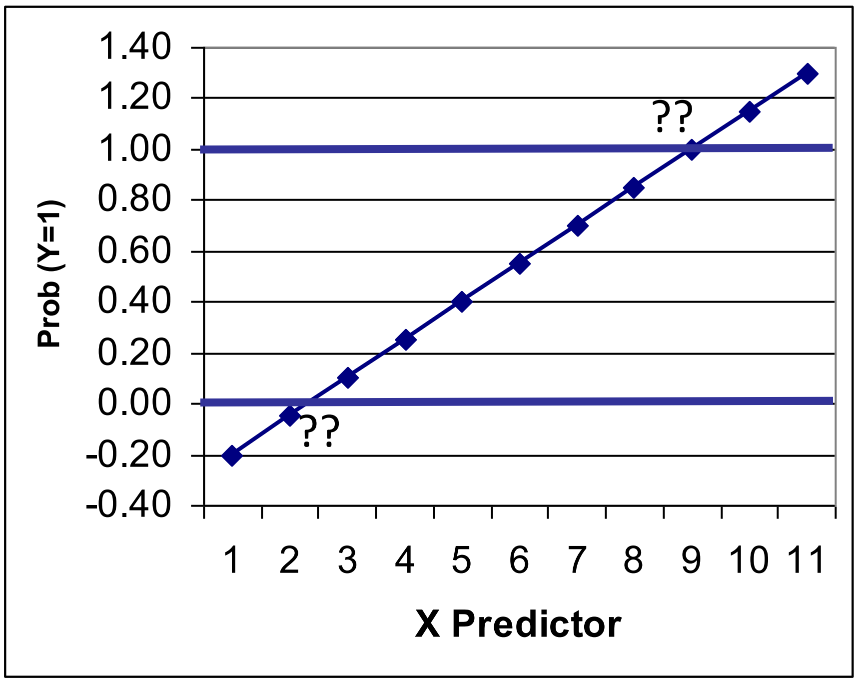
- But if Y_p is binary, then e_p can only be 2 things:
  - $e_p = Y_p - \hat{Y}_p$
    - If $Y_p = 0$ then e_p = (0 − predicted probability)
    - If $Y_p = 1$ then e_p = (1 − predicted probability)

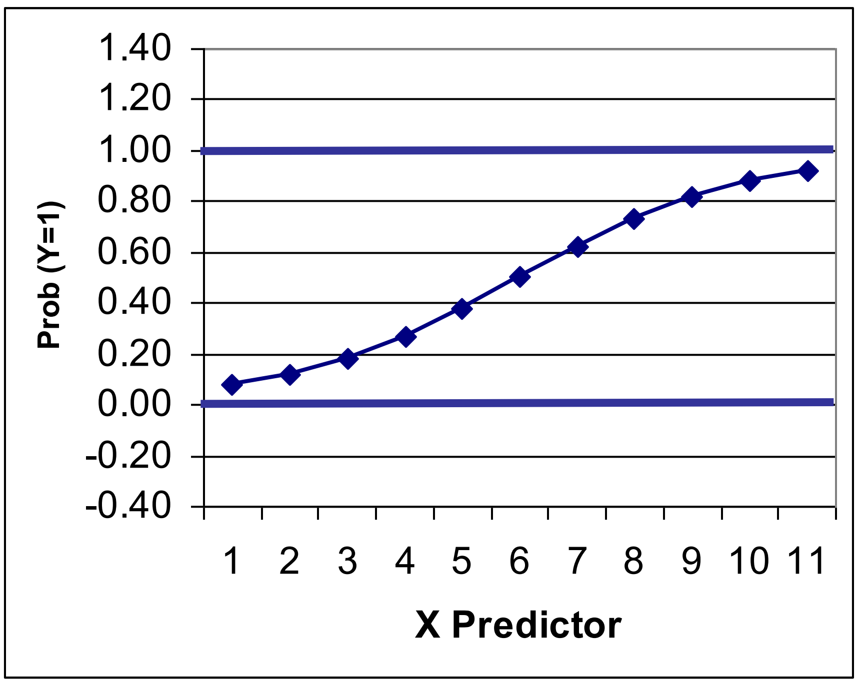
  - The mean of errors would still be 0…by definition

  - But variance of errors can't possibly be constant over levels of X like we assume in general linear models
    - The mean and variance of a binary outcome are **dependent**!
    - As shown shortly, mean = p and variance = p*(1-p), so they are tied together
    - This means that because the conditional mean of Y (p, the predicted probability Y= 1) is dependent on X, *then so is the error variance*

A General Linear Model With Binary Outcomes?

- How can we have a linear relationship between X & Y?
- Probability of a 1 is bounded between 0 and 1, but predicted probabilities from a linear model aren't bounded
  - Impossible values
- Linear relationship needs to 'shut off' somehow → made nonlinear





3 Problems with General* Linear Models Predicting Binary Outcomes

- *General = model for continuous, conditionally normal outcome

- Restricted range (e.g., 0 to 1 for binary item)
  - Predictors should not be linearly related to observed outcome
    → Effects of predictors need to be 'shut off' at some point to keep predicted values of binary outcome within range

- Variance is dependent on the mean, and not estimated
  - Fixed (→predicted value) and random (error) parts are related
    → So residuals can't have constant variance

- Further, residuals have a limited number of possible values
  - Predicted values can each only be off in two ways
    → So residuals can't be normally distributed

The Binary Case: Bernoulli Distribution

For items that are binary (dichotomous/two options), a frequent distribution chosen is the Bernoulli distribution (the Bernoulli distribution is also called a one-trial binomial distribution):

Notation: $Y_p \sim B(p_p)$ (where p is the conditional probability of a 1 for person p)

Sample Space: $Y_p \in \{0,1\}$ (Y_p can either be a 0 or a 1)

Probability Density Function (PDF):

$$f(Y_p) = (p_p)^{Y_p}(1 - p_p)^{1-Y_p}$$

Expected value (mean) of Y: $E(Y_p) = \mu_{Y_p} = p_p$

Variance of Y: $V(Y_p) = \sigma^2_{Y_p} = p_p(1 - p_p)$

Note: p_p is the only parameter – so we only need to provide a link function for it…

Generali*zed* Models for Binary Outcomes

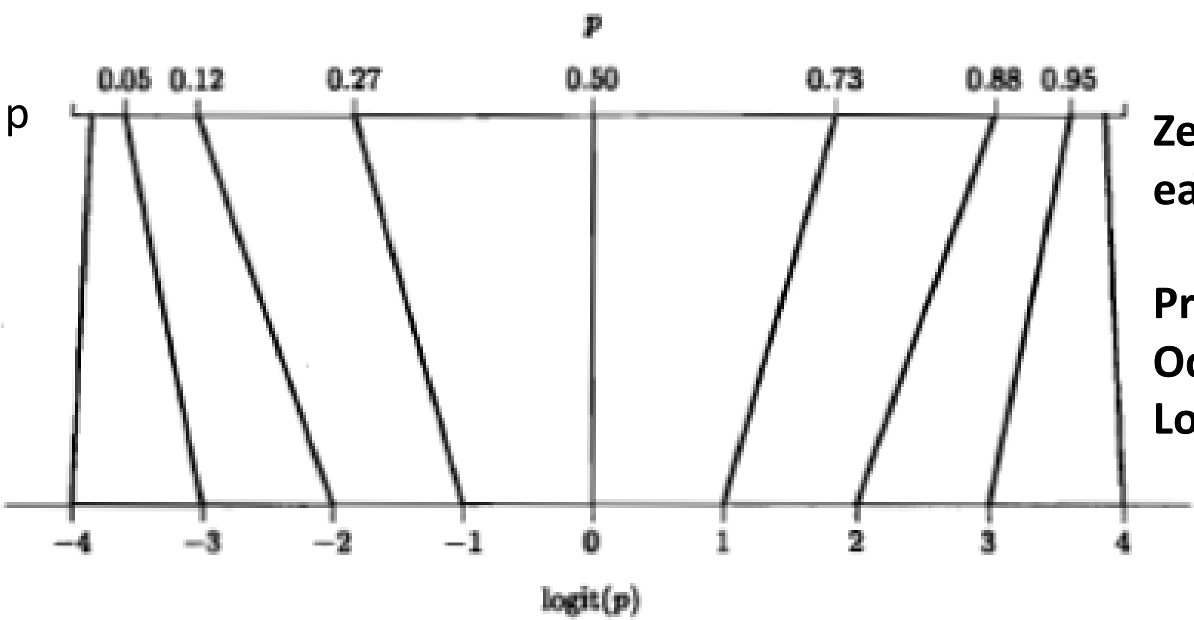
- Rather than modeling the probability of a 1 directly, we need to transform it into a more continuous variable with a **link function**, for example:
  - We could transform **probability** into an **odds ratio**:
    - Odds ratio: (p / 1-p) → prob(1) / prob(0)
    - If p = .7, then Odds(1) = 2.33; Odds(0) = .429
    - Odds scale is way skewed, asymmetric, and ranges from 0 to +∞
      - Nope, that's not helpful
  - **Take *natural log of odds ratio* → called "logit" link**
    - LN (p / 1-p) → Natural log of (prob(1) / prob(0))
    - If p = .7, then LN(Odds(1)) = .846; LN(Odds(0)) = -.846
    - Logit scale is now symmetric about 0 → DING
  - The logit link is one of many used for the Bernoulli distribution
    - Names of others: Probit, Log-Log, Complementary Log-Log

Turning Probability into Logits

- **Logit is a nonlinear transformation of probability:**
  - Equal intervals in logits are NOT equal in probability
  - The logit goes from ±∞ and is symmetric about prob = .5 (logit = 0)
  - This solves the problem of using a linear model
    - The model will be **linear with respect to the logit**, which translates into nonlinear with respect to probability (i.e., it **shuts off as needed**)

Probability: p

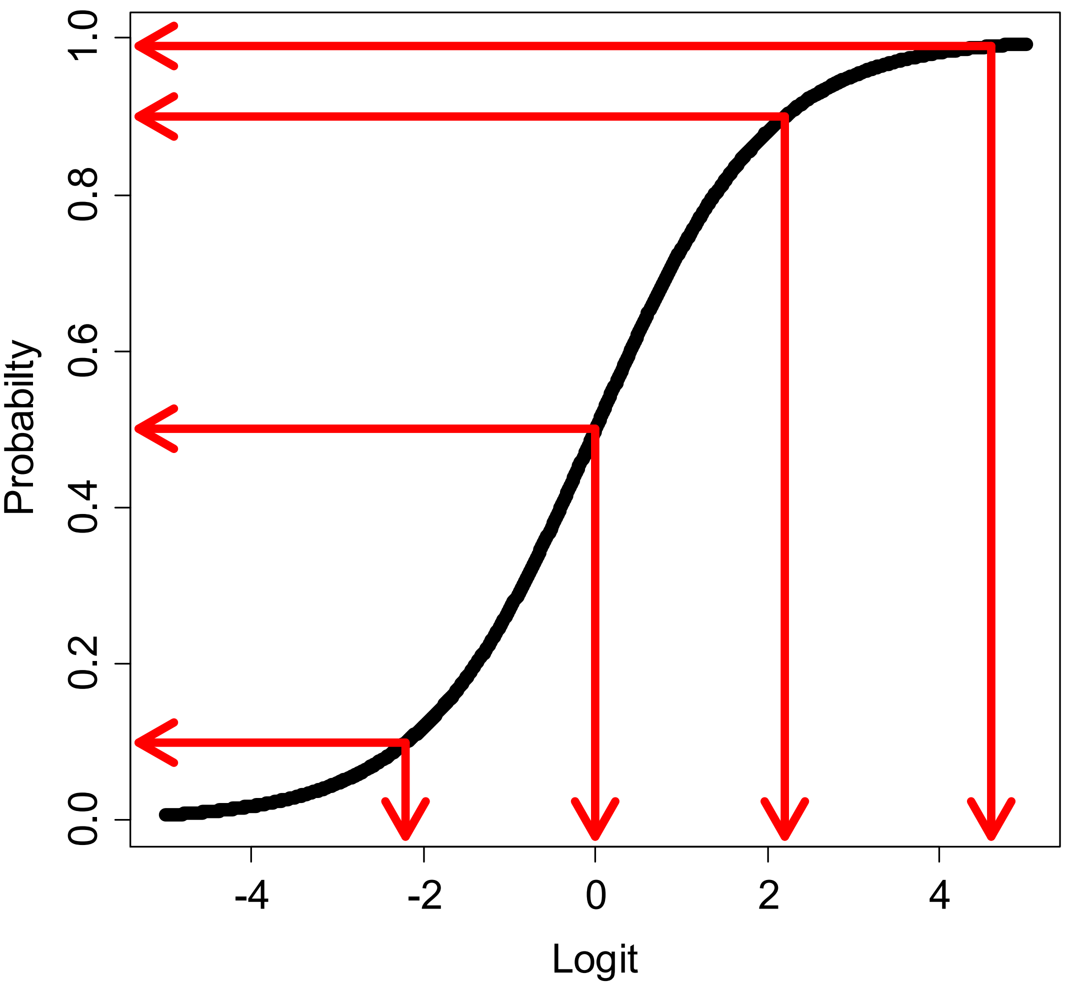
Logit:
LN (p / 1-p)



Zero-point on each scale:

Prob = .5
Odds = 1
Logit = 0

Transforming Probabilities to Logits



Probability	Logit
0.99	4.6
0.90	2.2
0.50	0.0
0.10	-2.2

Can you guess what a probability of .01 would be on the logit scale?

Transforming Logits to Probabilities: $g(\cdot)$ and $g^{-1}(\cdot)$

- In the terminology of generalized models, the link function for a logit is defined by (log = natural logarithm):

$$g\left(E\left(Y_p\right)\right) = \log\left(\frac{P\left(Y_p = 1\right)}{\left(1 - P\left(Y_p = 1\right)\right)}\right) = \underbrace{\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p}_{\boxed{\text{Linear Predictor}}}$$

- A logit can be translated to a probability with some algebra:

$$\exp\left[\log\left(\frac{P\left(Y_p = 1\right)}{\left(1 - P\left(Y_p = 1\right)\right)}\right)\right] = \exp\left[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right]$$

$$\leftrightarrow \left(1 - P\left(Y_p = 1\right)\right)\left[\frac{P\left(Y_p = 1\right)}{\left(1 - P\left(Y_p = 1\right)\right)}\right] = \left(\exp\left[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right]\right)\left(1 - P\left(Y_p = 1\right)\right)$$

Transforming Logits to Probabilities: $g(\cdot)$ and $g^{-1}(\cdot)$

- Continuing:

$$P(Y_p = 1) = \left(\exp[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p]\right) - \left(\left(\exp[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p]\right) P(Y_p = 1)\right)$$
$$P(Y_p = 1)\left(1 - \exp[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p]\right) = \exp[\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p]$$

- Which finally gives us:

$$P(Y_p = 1) = \frac{\exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}$$

Linear Predictor

- Therefore, the inverse logit (un-logit…or $g^{-1}(\cdot)$) is:

$$E(Y_p) = g^{-1}\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)$$
$$= \frac{\exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}$$

Written Another Way…

- The inverse logit $g^{-1}(\cdot)$ has another form that is sometimes used:

$$E\left(Y_p\right) = g^{-1}\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)$$

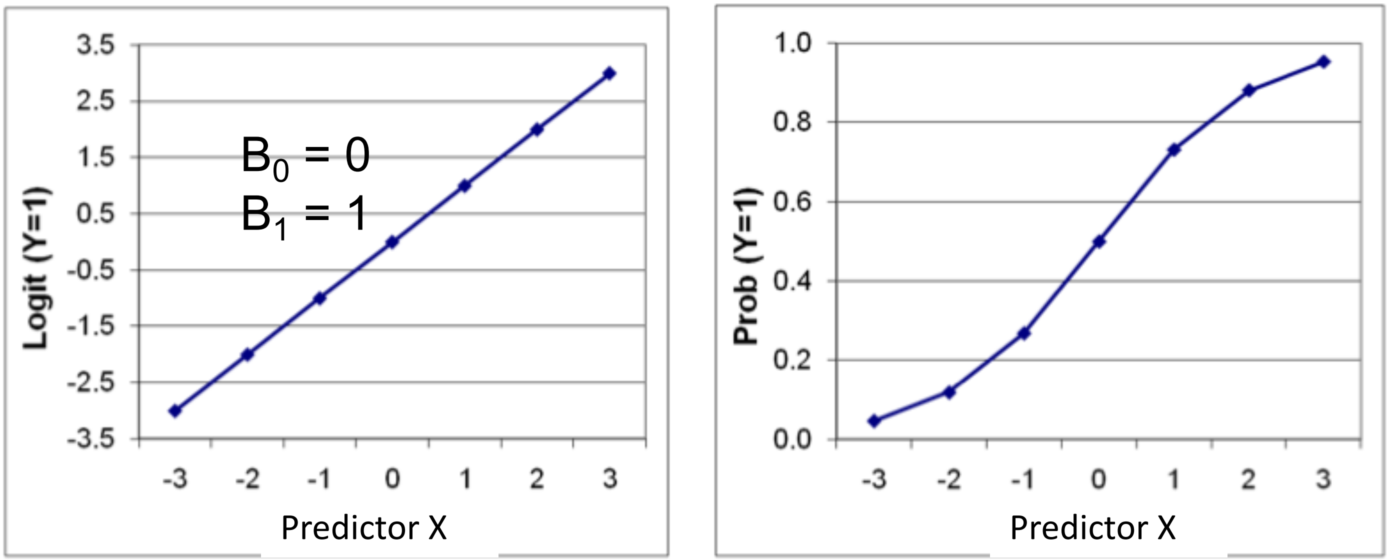
$$= \frac{\exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)}$$

$$= \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)\right)}$$

$$= \left(1 + \exp\left(-\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)\right)\right)^{-1}$$

Nonlinearity in Prediction

- The relationship between X and the probability of response=1 is **"nonlinear" → an s-shaped logistic curve** whose shape and location are dictated by the estimated fixed effects
  - **Linear** with respect to the **logit**, **nonlinear** with respect to **probability**



- The logit version of the model will be easier to explain; the probability version of the prediction will be easier to show

Putting it Together with Data: The Empty Model

- The empty model (under GLM):

$$Y_p = \beta_0 + e_p$$

where $e_p \sim N(0, \sigma_e^2)$ $E(Y_p) = \beta_0$ and $V(Y_p) = \sigma_e^2$

Linear Predictor

- The empty model for a Bernoulli distribution with a logit link:

$$g\left(E(Y_p)\right) = logit\left(P(Y_p = 1)\right) = logit(p_p) = \beta_0$$

$$p_p = P(Y_p = 1) = E(Y_p) = g^{-1}(\beta_0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$$V(Y_p) = p_p(1 - p_p)$$

- Note: many generalized LMs don't list an error term in the linear predictor – is for the expected value and error usually has a 0 mean so it disappears

- We could have listed e_p for the logit function
  - e_p would have a logistic distribution with a zero mean and variance $\frac{\pi^2}{3} = 3.29$
  - Variance is fixed – cannot modify variance of Bernoulli distribution after modeling the mean

THE UNIVERSITY OF KANSAS

LOGISTIC REGRESSION IN R

The Ordinal Package

- The ordinal package is useful for modeling categorical dependent variables

- We will use the clm() function
  - clm stands for cumulative linear models

Unpacking clm() Function Syntax

- Example syntax below for empty model differs only slightly from lm() syntax we have already seen

```
# response variable must be a factor:
data01$Lapply = factor(data01$Lapply)

# EMPTY MODEL PREDICTING DICHOTOMOUS (0/1): Likely To Apply; Modeling Prob of 1
model01 = clm(formula = Lapply ~ 1, data = data01, control = clm.control(trace = 1))
summary(model01)
```

- The dependent variable must be stored as a factor

- The formula and data arguments are identical to lm()

- The control argument is only used here to show iteration history of the ML algorithm

Empty Model Output

- The empty model is estimating one parameter: β_0
- However, for this package, the logistic regression is formed using a threshold (τ_0) rather than intercept rather
  - Here $\beta_0 = -\tau_0$

```
> summary(model01)
formula: Lapply ~ 1
data:    data01

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  400  -275.26 552.51 3(0)  3.31e-14 1.0e+00

Threshold coefficients:
    Estimate Std. Error z value
0|1   0.2007     0.1005   1.997
```


THE UNIVERSITY OF KANSAS

Interpretation of summary() Output

- $\tau_0 = 0.2007, \text{so}\ldots$
- $\beta_0 = -0.2007\ (0.1005)$: interpreted as the predicted **logit** of y_p =1 for an individual when all predictors are zero
  - Because of the empty model, this becomes average **logit** for sample
  - Note: exp(-.2007)/(1+exp(-.2007)) = .55 – the sample mean proportion
- The log-likelihood is -256.26
  - Used for nested model comparisons
- The AIC is 552.51
  - Used for non-nested model comparisons

Predicting Logits, Odds, & Probabilities:

- <u>Coefficients for each form of the model:</u>
  - Logit: $\text{Log}(p_p/1-p_p) = \boldsymbol{\beta_0}$
    - Predictor effects are **linear and additive** like in regression, but what does a 'change in the logit' mean anyway?
    - Here, we are saying the average logit is -.2007
  - Odds: $(p_p/1-p_p) = \textbf{exp}(\boldsymbol{\beta_0})$
    - A compromise: effects of predictors are **multiplicative**
    - Here, we are saying the average odds of a applying to grad school is exp(-.2007) = .819
  - Prob: $P(y_p=1) = \dfrac{\textbf{exp}(\boldsymbol{\beta_0})}{\textbf{1+ exp}(\boldsymbol{\beta_0})}$
    - Effects of predictors on probability are **nonlinear and non-additive** (no "one-unit change" language allowed)
    - Here, we are saying the average probability of applying to grad school is .550

THE UNIVERSITY OF KANSAS

MAXIMUM LIKELIHOOD ESTIMATION OF GENERALIZED MODELS

Maximum Likelihood Estimation of Generalized Models

- The process of ML estimation in Generalized Models is similar to that from the GLM, with two exceptions:
  - The error variance is not estimated
  - The fixed effects do not have closed form equations (so are now part of the log likelihood function search)

- We will describe this process for the previous analysis, using our grid search

- Here, each observation has a Bernoulli distribution where the "height" of the curve is given by the PDF:

$$f(Y_p) = (p_p)^{Y_p}(1 - p_p)^{1-Y_p}$$

- The generalized linear model then models

$$E(Y_p) = p_p = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

THE UNIVERSITY OF KANSAS

From One Observation…To The Sample

- The likelihood function shown previously was for one observation, but we will be working with a sample
  - Assuming the sample observations are independent and identically distributed, we can form the joint distribution of the sample

Multiplication comes from independence assumption:
Here, $L(\beta_0|Y_p)$ is the Bernoulli PDF for Y_p using a logit link for β_0

$$L(\beta_0|Y_1, \ldots, Y_N) = L(\beta_0|Y_1) \times L(\beta_0|Y_2) \times \cdots \times L(\beta_0|Y_N)$$

$$= \prod_{p=1}^{N} f(Y_p) = \prod_{p=1}^{N} p_p^{Y_p}(1 - p_p)^{1-Y_p}$$

$$= \prod_{p=1}^{N} \left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)^{Y_p} \left(1 - \left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)\right)^{1-Y_p}$$

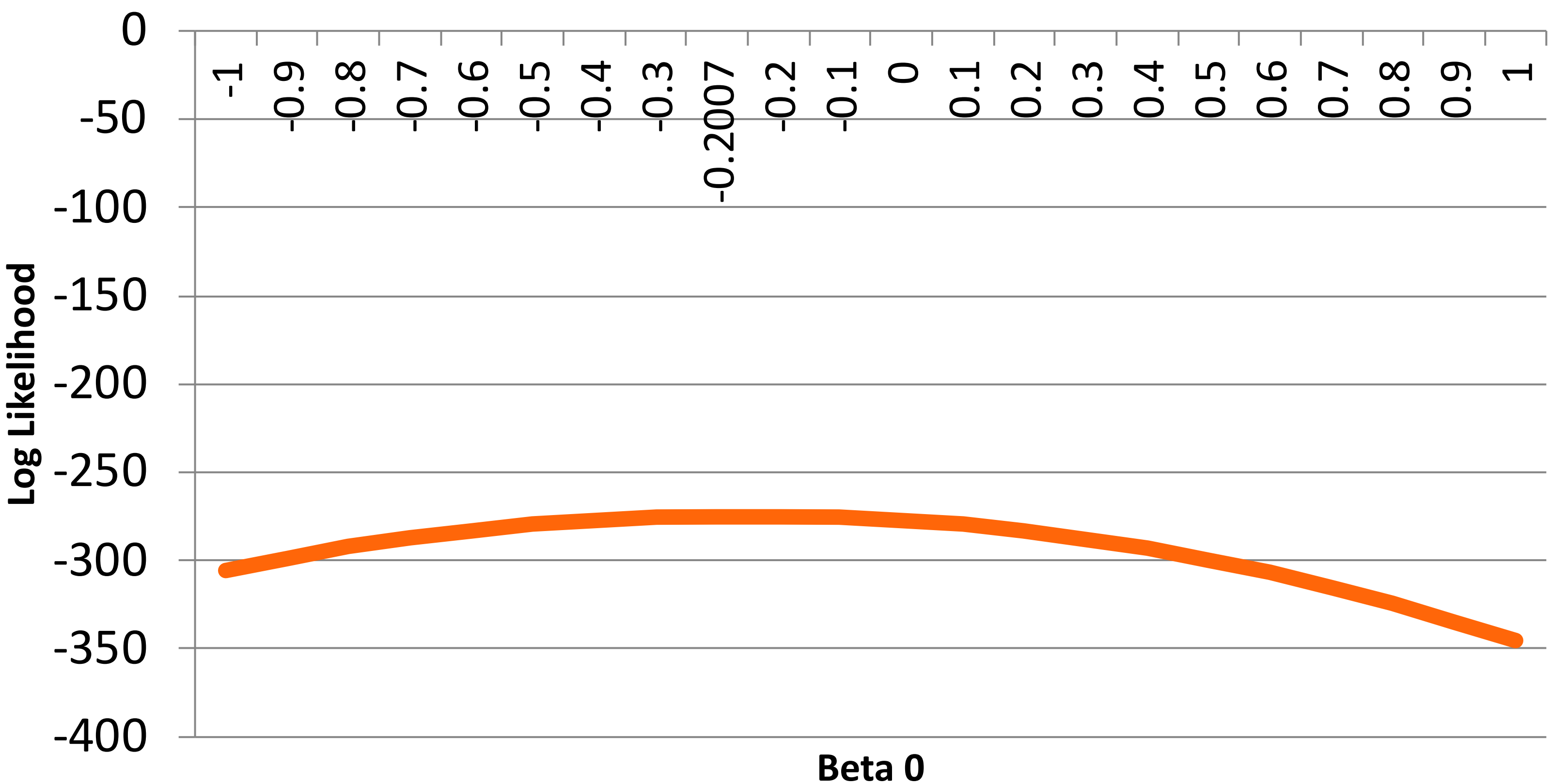
The Log Likelihood Function

- The log likelihood function is found by taking the natural log of the likelihood function:

$$\log L(\beta_0|Y_1,\dots,Y_N) = \log\left(L(\beta_0|Y_1)\times L(\beta_0|Y_2)\times\cdots\times L(\beta_0|Y_N)\right)$$
$$= \sum_{p=1}^{N} \log\left(L(\beta_0|Y_p)\right) = \sum_{p=1}^{N} \log\left[p_p^{Y_p}\left(1-p_p\right)^{1-Y_p}\right]$$
$$= \sum_{p=1}^{N} Y_p \log\left(p_p\right) + \left(1-Y_p\right)\log\left(1-p_p\right)$$
$$= \sum_{p=1}^{N} Y_p \log\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right) + \left(1-Y_p\right)\log\left(1-\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)$$

Grid Search of the Log Likelihood Function

- Just like we did for the normal distribution, we can plot the log likelihood function for all possible values of β_0



Iteration History from clm()

- We can show the history of iterations, where the "value" column is -1 times the log-likelihood

```
> model01 = clm(formula = Lapply ~ 1, data = data01, control = clm.control(trace = 1))
iter:  step factor:     Value:     max|grad|:   Parameters:
   0:  1.000000e+00:    277.259:   2.000e+01:   0
nll reduction:  2.00332e+00
   1:  1.000000e+00:    275.256:   6.640e-02:   0.2
nll reduction:  2.22672e-05
   2:  1.000000e+00:    275.256:   2.222e-06:   0.2007
nll reduction:  -5.68434e-14
   3:  1.000000e+00:    275.256:   3.308e-14:   0.2007
```

At the Maximum…

- At the maximum ($\beta_0 = -0.2007$) we now assume that the parameter β_0 has a normal distribution
  - Only the **data** Y have a Bernoulli distribution

- Putting this into statistical context:

$$\beta_0 \sim N\left(\hat{\beta}_0, se\left(\hat{\beta}_0\right)^2\right)$$

- This says that the true parameter β_0 has a mean at our estimate and has a variance equal to the square of the standard error of our estimate

ADDING PREDICTORS TO THE EMPTY MODEL

Adding Predictors to the Empty Model

- Having examined how the logistic link function works and how estimation works, we can now add predictor variables to our model:

$$g\left(E\left(Y_p\right)\right) = logit\left(P\left(Y_p = 0\right)\right) = logit\left(p_p\right)$$
$$= \beta_0 + \beta_1 PARED_p + \beta_2\left(GPA_p - 3\right) + \beta_3 PUBLIC_p$$

$$p_p = E\left(Y_p\right) = g^{-1}\left(\beta_0 + \beta_1 PARED_p + \beta_2\left(GPA_p - 3\right) + \beta_3 PUBLIC_p\right)$$
$$= \frac{\exp\left(\beta_0 + \beta_1 PARED_p + \beta_2\left(GPA_p - 3\right) + \beta_3 PUBLIC_p\right)}{1 + \exp\left(\beta_0 + \beta_1 PARED_p + \beta_2\left(GPA_p - 3\right) + \beta_3 PUBLIC_p\right)}$$

$$V\left(Y_p\right) = p_p\left(1 - p_p\right)$$

- Here PARED is Parent Education, PUBLIC is Public University, and GPA is Grade Point Average (centered at a value of 3)
- For now, we will omit any interactions (to simplify interpretation)
- We will also use the default parameterization (modeling Y = 0)

Understanding R Input and Output

- First…the syntax

```
# MODEL 02: ADDING PREDICTORS TO THE EMPTY MODEL
model02 = clm(formula = Lapply ~ 1 + PARED + PUBLIC + GPA3,
              data = data01, control = clm.control(trace = 1))
```

- The algorithm iteration history:

```
> # MODEL 02: ADDING PREDICTORS TO THE EMPTY MODEL
> model02 = clm(formula = Lapply ~ 1 + PARED + PUBLIC + GPA3,
+               data = data01, control = clm.control(trace = 1))
iter:  step factor:     Value:     max|grad|:   Parameters:
   0:  1.000000e+00:    277.259:   2.000e+01:   0 0 0 0
nll reduction:  1.22751e+01
   1:  1.000000e+00:    264.984:   5.723e-01:   0.3322 1.014 -0.1885 0.5169
nll reduction:  2.13685e-02
   2:  1.000000e+00:    264.962:   4.991e-03:   0.3382 1.059 -0.2005 0.5481
nll reduction:  1.17396e-06
   3:  1.000000e+00:    264.962:   3.705e-07:   0.3382 1.06 -0.2006 0.5482
```


THE UNIVERSITY OF KANSAS

Question #1: Does Conditional Model Fit Better than Empty Model

- Question #1: does this model fit better than the empty model?

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1: \text{At least one not equal to zero}$$

```
anova(model01, model02)
```

- Likelihood Ratio Test Statistic = Deviance = -2*(-275.26- -264.96) = 20.586
  - -275.26 is log likelihood from empty model
  - -264.96 is log likelihood from conditional model
- DF = 4 – 1 =3
  - Parameters from empty model = 1
  - Parameters from this model = 4
- P-value: $p = .0001283$
- Conclusion: reject H_0; this model is preferred to empty model

```
> anova(model01, model02)
Likelihood ratio tests of cumulative link models:

        formula:                            link: threshold:
model01 Lapply ~ 1                          logit flexible
model02 Lapply ~ 1 + PARED + PUBLIC + GPA3 logit flexible

        no.par    AIC  logLik LR.stat df Pr(>Chisq)
model01      1 552.51 -275.26
model02      4 537.92 -264.96  20.586  3  0.0001283 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting Model Parameters from summary()

- Parameter Estimates:

```
> summary(model02)
formula: Lapply ~ 1 + PARED + PUBLIC + GPA3
data:    data01

 link  threshold nobs logLik  AIC    niter max.grad cond.H
 logit flexible  400  -264.96 537.92 3(0)  3.71e-07 1.0e+01

Coefficients:
       Estimate Std. Error z value Pr(>|z|)
PARED    1.0596     0.2974   3.563 0.000367 ***
PUBLIC  -0.2006     0.3053  -0.657 0.511283
GPA3     0.5482     0.2724   2.012 0.044178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
    Estimate Std. Error z value
0|1   0.3382     0.1187   2.849
```

- Intercept $\beta_0 = -0.3382\ (0.1187)$: this is the predicted value for the **logit of y_p = 1** for a person with: 3.0 GPA, parents without a graduate degree, and at a private university
  - Converted to a probability: .417 – probability a student with 3.0 GPA, parents without a graduate degree, and at a private university is likely to apply to grad school (y_p = 1)

THE UNIVERSITY OF KANSAS

Interpreting Model Parameters

parentGD: $\beta_1 = 1.0596\ (0.2974); p = .0004$

The change in the **logit of y_p = 1** for every one-unit change in parentGD…or, the difference in the **logit of y_p = 1** for students who have parents with a graduate degree

Because logit of y_p = 1 means a rating of "likely to apply" this means that students who have a parent with a graduate degree are more likely to rate the item with a "likely to apply"

More on Slopes

- The quantification of **how much** less likely a student is to respond with "unlikely to apply" can be done using odds ratios or probabilities:

Odds Ratios:

- Odds of "likely to apply" (Y=1) for student **with** parental graduate degree: $\exp(\beta_0 + \beta_1) = 2.05$
- Odds of "likely to apply" (Y=1) for student **without** parental graduate degree: $\exp(\beta_0) = .713$
- Ratio of odds = $2.88525 = \exp(\beta_1)$ - meaning, a student **with** parental graduate degree has almost 3x the odds of rating "likely to apply"

Probabilities:

- Probability of "likely to apply" for student **with** parental graduate degree: $\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} = .673$
- Probability of "likely to apply" for student **without** parental graduate degree: $\frac{\exp(\beta_0)}{1+\exp(\beta_0)} = .416$

Interpreting Model Parameters

PUBLIC: $\beta_2 = -0.2006\ (0.3053); p = .5113$:

The change in the **logit of y_p = 1** for every one-unit change in GPA…

But, PUBLIC is a coded variable where 0 represents a student in a private university, so this is the difference in logits of the **logit of y_p = 1** for students in public vs private universities

Because logit of 1 means a rating of "likely to apply" this means that students who are at a public university are more unlikely to rate "likely to apply"

More on Slopes

- The quantification of **how much** more likely a student is to respond with "likely to apply" can be done using odds ratios or probabilities:

Public	Logit	Odds of 1	Prob = 1
1	-0.539	0.583	0.368
0	-0.338	0.713	0.416

- The odds are found by: $\exp\left(\beta_0 + \beta_3 PUB_p\right)$

- The probability is found by: $\frac{\exp\left(\beta_0+\beta_3 PUB_p\right)}{1+\exp\left(\beta_0+\beta_3 PUB_p\right)}$

Interpreting Model Parameters

GPA3: $\beta_2 = 0.5482\ (0.2724); p = .0442$:

The change in the **logit of $\mathbf{y_p}$ = 1** for one-unit change in GPA

Because logit of y_p = 1 means a rating of “likely to apply” this means that students who have a higher GPA are more likely to rate “likely to apply”

More on Slopes

- The quantification of **how much** more likely a student is to respond with "likely to apply" can be done using odds ratios or probabilities:

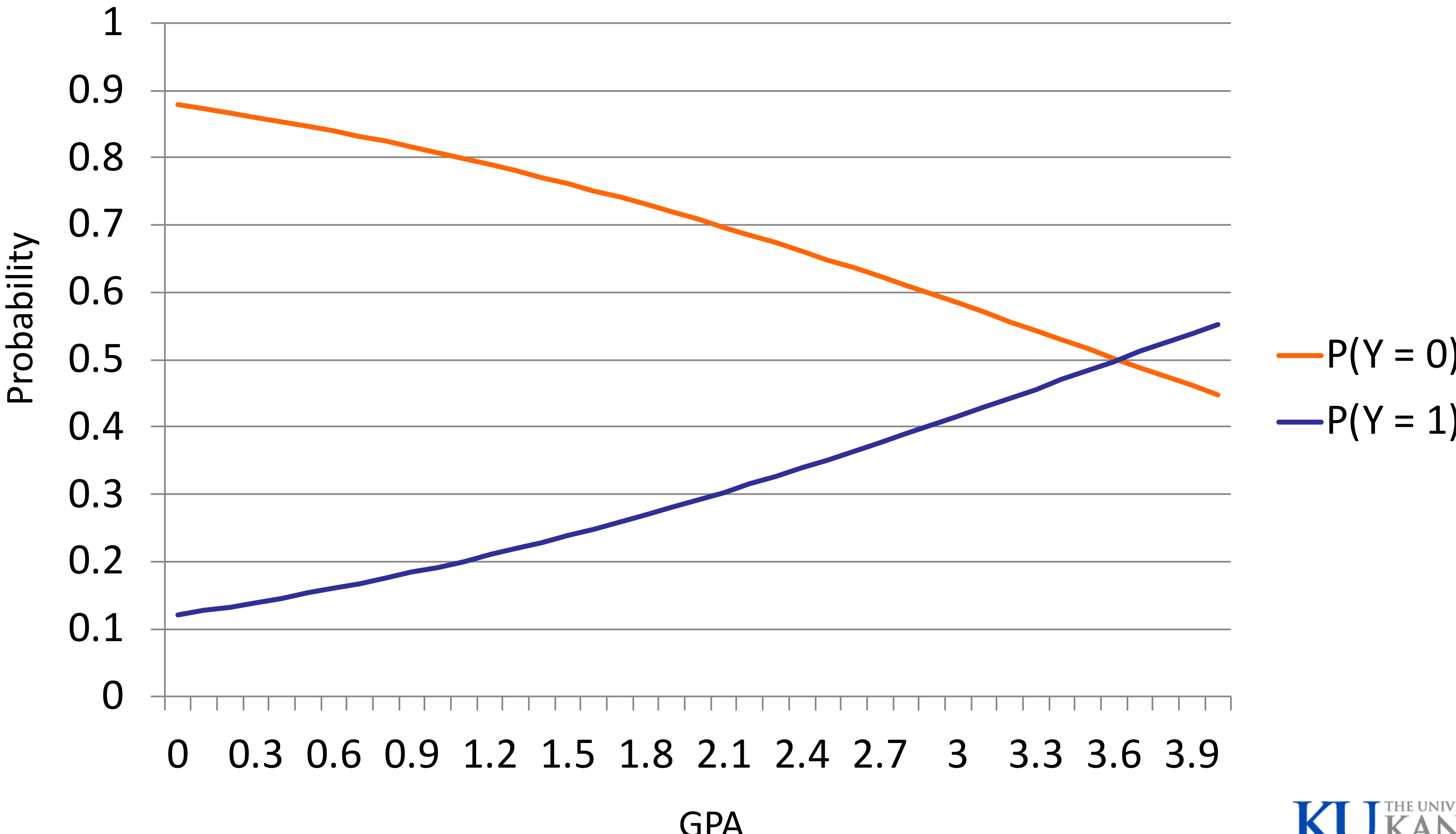
GPA3	Logit	Odds of 1	Prob = 1
1	0.210	1.234	0.552
0	-0.338	0.713	0.416
-1	-0.886	0.412	0.292
-2	-1.435	0.238	0.192

- The odds are found by: $\exp\left(\beta_0 + \beta_2\left(GPA_p - 3\right)\right)$

- The probability is found by: $\frac{\exp\left(\beta_0+\beta_2\left(GPA_p-3\right)\right)}{1+\exp\left(\beta_0+\beta_2\left(GPA_p-3\right)\right)}$

Plotting GPA

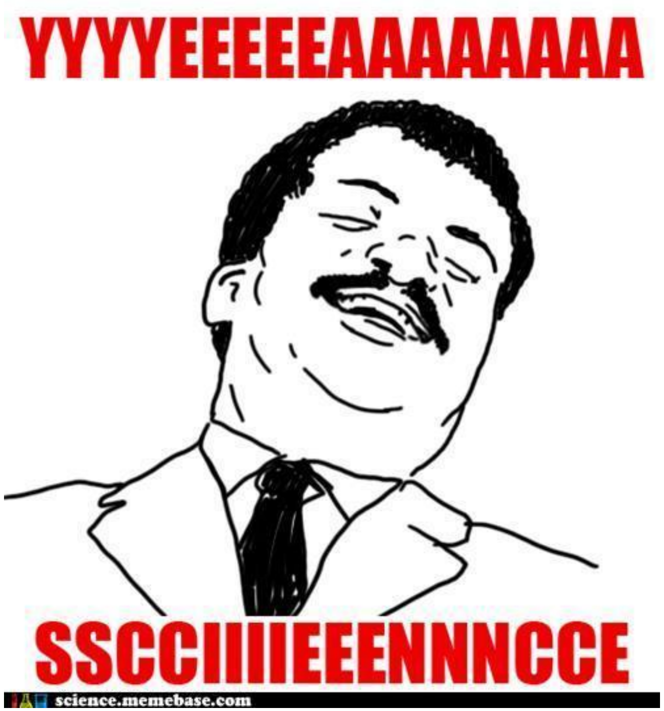
- Because GPA is an **unconditional** main effect, we can plot values of it versus probabilities of rating "likely to apply"



Interpretation In General

- In general, the linear model interpretation that you have worked on to this point still applies for generalized models, with some nuances

- For logistic models with two responses:
  - Regression weights are now for LOGITS
  - The direction of what is being modeled has to be understood (Y = 0 or = 1)
  - The change in odds and probability is not linear per unit change in the IV, but instead is linear with respect to the logit
    - Hence the term “linear predictor”
  - Interactions will still
    - Will still modify the conditional main effects
    - Simple main effects are effects when interacting variables = 0


YYYYEEEEEAAAAAAAA
SSCCIIIIIEEENNNCCE
science.memebase.com

WRAPPING UP

Wrapping Up

- Generalized linear models are models for outcomes with distributions that are not necessarily normal

- The estimation process is largely the same: maximum likelihood is still the gold standard as it provides estimates with understandable properties

- Learning about each type of distribution and link takes time:
  - They all are unique and all have slightly different ways of mapping outcome data onto your model

- Logistic regression is one of the more frequently used generalized models – binary outcomes are common

THE UNIVERSITY OF KANSAS