Introduction to this Course on Generalized Linear Models
(and to Maximum Likelihood Estimation)

- Topics:
  - What to expect this semester
  - Quantitative methods: A Lego-inspired world view
  - A little bit about:
    - Maximum likelihood
    - Link functions
    - Outcome variable types and conditional distributions
    - Multivariate modeling and software
    - Why this "bridge" course will help you in future courses

What To Expect This Semester…

- I believe that **everyone is capable** and **can significantly benefit**** from learning more types of quantitative methods!
  - *The hard part is the working memory load, not the math!*

- **Philosophy:** Focus on accessibility + mastery learning
  - No anxiety-prone tasks (e.g., hand calculations, memorizing formulas)
  - No anxiety-prone methods of evaluation (e.g., timed tests)

- **Materials:** Unit = (wordy) lecture + example(s); 6 units planned
  - **Lecture** slides present concepts—the **what** and the **why**
  - **Example** documents: reinforce the concepts and demonstrate the **how using software**—STATA or R (your choice) first, then Mplus
  - All available at the course website (hosted outside of ICON)

** ***Benefits*** *include but are not limited to: Better research, more authorship opportunities, and actual money*

Course Requirements

- **All course requirements are take-home, open-note, and untimed**
- Late* work will be accepted (−2 for HW; −1 for FA)
  - **Extensions will be granted if requested at least 2 weeks in advance*
  - HW due dates **may be pushed later** (to ensure approximately 1 week after covering the material before HW is due), but never sooner
- **6 formative assessments (FA, 12 points) in ICON:** Top-of-head questions and story problems for structured review (will discuss answers in next class)
  - 2 points for an honest attempt to complete each FA (mostly without feedback)
  - An opportunity for you to request topics for further clarification and review
- **6 homework assignments (HW, 88 points):** Practice doing data analysis
  - Based directly on examples given (no googling or ChatGPT needed, ever!)
  - **Online HW 1–4, 6**: Unique canned dataset (made with a common story)
    - **Computation** sections: Instant feedback, infinite attempts
    - **Results** (interpretation) sections: Delayed feedback, single attempt
  - **HW 5: Individual data analysis + written results section**
    - Get my feedback for using a model of your choice on your data + optional **revision**

Our *Other* Responsibilities

- My job (besides providing materials and assignments):
  - **Answer questions** via email, in individual meetings, or in group-based zoom office hours—you can each work on homework during office hours and get (near) immediate assistance (and then keep working)
    - Email me first (but you can follow up with the TAs if they help you)

- Your job (in descending order of timely importance):
  - **Ask questions**—preferably in class, but any time is better than none
  - **Frequently review** the class material, focusing on mastering the vocabulary (words and symbols), logic, and procedural skills
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help if you get stuck** on one thing for more than 15 minutes
    - Please email me a screenshot of your code+error so I can respond easily
  - **Do the readings** for a broader perspective and additional examples (best after lecture; readings are for the whole unit, not just that day)
  - **Practice** using the software to implement the techniques you are learning **on data you care about**—this will help you so much more!

More About Your Experience in this Class

- **Attendance**: Strongly recommended but not required
  - **You choose each class**: In-person "roomer" or online "zoomer"
  - **Masks** are still welcome for in-person attendees
  - **Please do not attend in-person if you might be sick!**
  - You won't miss out: I will post YouTube-hosted **recordings** (audio + screenshare only) for each class at the course website
  - **Ask questions aloud or in the zoom chat window (+DM)** (even if you are attending class in-person as a "roomer")

- **Changes** will be sent via email by 9 am on class days
  - I will change to zoom-only if I am exposed to Covid!
  - I will change to zoom-only for dangerous weather (or no room)
  - NOTHING is more important than our health and safety…

Class-Sponsored Statistical Software

- To help address the needs of different Iowa degree programs, I will show examples primary using both **STATA and R** software
  - **STATA** (aka, Stata) = "Software for Statistics and Data Science"
  - **R** = free implementation of what was initially the "S" language
  - Some examples may (still) use **SAS** = "Statistical Analysis System"
  - **M*plus*** (SEM software) will be added for last unit on path analysis

- **Why not SPSS**? Because it doesn't have as much room to grow (and thus it isn't used in any other EMS advanced classes)
  - Just like SPSS, STATA has drop-down windows to generate syntax
  - SPSS *could* be used for some—but not all—of our content

- **My story:** After SPSS, I became a heavy-duty **SAS enthusiast** who:
  - Picked up enough STATA initially to teach workshops using it, and I am learning it better now that I teach it in my classes
  - Is (begrudgingly) learning enough (base) R to add it to my classes
  - So if you have **STATA or R tips**, *please* share them with me!

Which Program: STATA or R?

- **Yes, you will need to learn to use at least one of these!**
  - Each is available (with VPN) in the free U Iowa Virtual Desktop
  - More programs = more "technical skills" for your CV; easier collaboration with colleagues (who only know one program)

- **To consider** when choosing which program to focus on:
  - Future use: R can be freely installed on your own machine, whereas STATA install = $$$ ($48 for 6-month student license)
  - **STATA** is popular in fields that use **large, weighted survey data** (e.g., sociology, political science, public health, EPLS at Iowa)
    - STATA syntax and documentation is easier and more standardized
  - **R** will be used exclusively in classes by Drs. Aloe, LeBeau, or Templin, and it has become increasingly mainstream (replacing SAS), **but**:
    - R packages are only as good as their authors (so little quality control)
    - Syntax and capabilities are idiosyncratic to the packages (so grrrrrr)

Working with Programs Through Syntax

- For help getting started with each program, please see the videos for my other class, PSQF 6243 (posted for 2/7)
- Don't worry: I DO NOT need you to memorize syntax, ever!

- Instead, you can do exactly what I (still) do:
  - **Find the example source file** for what you need to do
  - Figure out how to **modify it** to work for your homework
  - **Copy** (control+C), **paste** (control+V), and **find and replace** (control+H) are your friends (Mac: swap control for command)

- Don't hesitate to ask for help (i.e., email me a screenshot)
- It will get easier with practice, I promise!!!

What You Are Supposed To Know Already

- Recommended prerequisite: PSQF 6243 or equivalent

- **Working prerequisites** are familiarity with:
  - Descriptive statistics and bivariate associations (e.g., correlation)
  - Statistical concepts (e.g., null hypothesis testing)
  - **General linear models** (i.e., regression, ANOVA)
    - With **moderation—interaction terms** of all kinds!
  - Use of some (non-excel) software for all of the above

- We will review these concepts in unit 1 (and reiterate throughout)
  - For a more thorough treatment of general linear models (and interaction terms), please review units 4–7 of my version of PSQF 6243

- This class will focus on **general*ized* linear models**… so what is that?

What We Will Cover This Semester

- **Units 2–4: General*ized* linear models for univariate outcomes**
  - **General*ized*** → regression predicting non-normal conditional outcomes
  - Unit 2: binary, ordinal, and nominal categorical outcomes
  - Unit 3: count and "if-and-how-much" outcomes
  - Unit 4: other non-normal (binomial, skewed) outcomes; quantile regression
- **Unit 5: Multivariate models via univariate software**
  - e.g., for repeated measures, dyadic/family data, and difference scores
  - Examples mostly for normal conditional outcomes (for good reason)
- **Unit 6: Multivariate models via path analysis**
  - Unit 5 examples done via path analysis instead (to help translation)
  - Mediation using normal conditional outcomes
  - Mediation using non-normal conditional outcomes
- But first, the bigger picture and some background…

Quant Methods: A Lego-Based Approach

My goal today:

a) describe these **4 Legos**

b) use them to provide the "big picture" of this course

Big Picture Idea:
If you understand the elemental building blocks of statistical models, then you can build **anything**!

The Origins of These Legos

- Problem: The **giant canyon** between two types of classes
  → To cross it, students need **2 kinds** of training
  - Become conversant in **traditional** methods (and the terms that go with them) still commonly used in many research areas
  - Recognize the **building blocks** of modern analytic techniques (current and future) to build a pathway to fluency with them
    - Recognizing the building blocks of traditional methods helps, too

- Solution: Build a **bridge course** that crosses this canyon
  - In specific: PSQF 6270, Generalized Linear Models
  - In general: A Lego-based **philosophy** for learning quantitative methods (developed in cahoots with Jonathan Templin)

The 4 Lego Building Blocks

The Legos we will cover in this course…

1. **Linear models** (for **answering questions** of prediction)
2. **Estimation** (for iterative ways of **finding the answers**)
3. **Link functions** (for predicting **any type of outcome**)

…will better prepare you to learn models that ALSO have:

4. (a) **Random effects** / (b) **Latent variables**

   (a) for modeling multivariate **"correlation/dependency"** using multilevel (*aka*, mixed-effects) models (MLM)

   (b) for modeling relations of **"unobserved constructs"** using confirmatory factor analysis (FA), item response theory (IRT), and structural equation models (SEM)

1. Linear Models Run the World

- **Linear models are the mechanism** by which the vast majority of all research questions will be answered
  - *Is there an effect of a given predictor? Is that effect the same for everyone? Is the effect still there after considering something else?*

- A linear-models world view entails starting with the most **traditional models**, but from a **different perspective**
  - More intuitive: linear regression models
    - *Because the focus is on the fixed effects in the model equation*
  - Less intuitive: analysis of variance in group-based designs
    - *Because the focus is on cell and marginal mean differences (which are indirectly provided by the model fixed effects)*
  - Both of these are flavors of the **General Linear Model**

Flavors of General Linear Models

- Unlike any other family of statistical models, **the same General Linear Model is called different names** (often taught in different classes) based on **what kind of predictor variables** are included:
  - One quantitative predictor? "Simple (linear) regression"
  - 2+ quantitative predictors? "Multiple (linear) regression"
  - One categorical predictor with two groups? "Independent *t*-test"
  - One categorical predictor with 3+ groups? "One-way ANOVA"
  - 2+ categorical predictors (with interactions between them)? "Two-way (or more-way) ANOVA"
  - 2+ categorical predictors (with interactions between them) and 1+ quantitative predictors (without interactions with the categorical predictors)? "Two (or more)-way ANCOVA"
  - Whatever combination is necessary? "Multiple regression"
- These distinctions only serve to confuse people and obfuscate what is **just one model**, the General Linear Model… **here is why**:

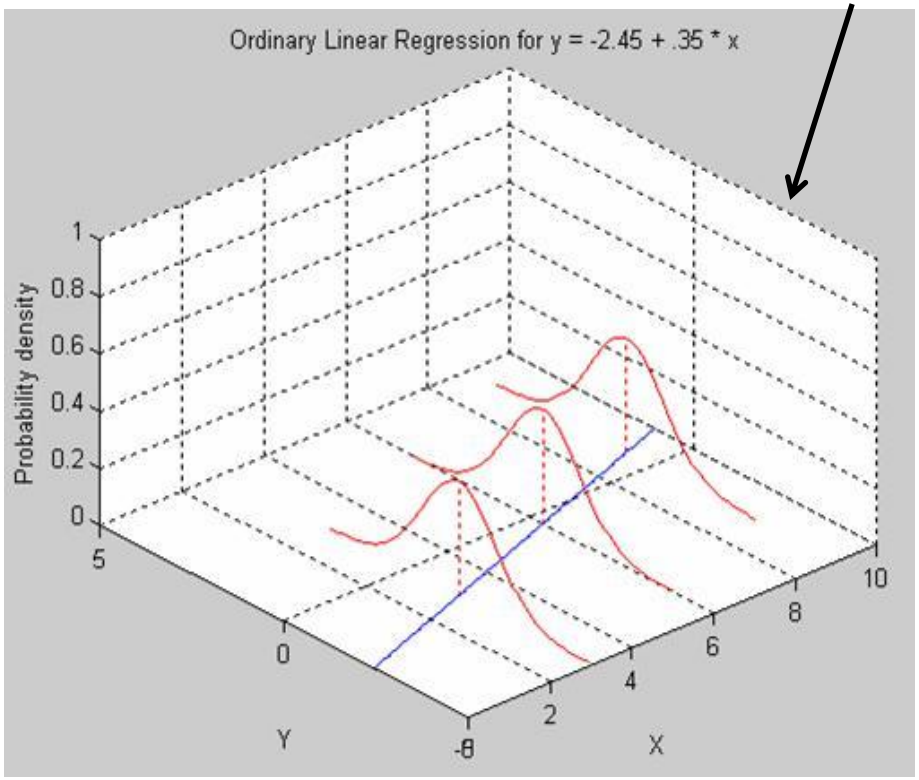
General Linear Model Residuals

- GLM for actual $\boldsymbol{y_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i) + e_i}$
- GLM for predicted $\boldsymbol{\hat{y}_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i)}$
  - Btw, we will practice interpreting fixed effects in GLMs in unit 1

- No matter what kind of predictors (and whether their interactions) are included, the term "**General**" in GLM refers to the use of a **conditional normal distribution** for the $\boldsymbol{e_i}$ residuals, in which $\boldsymbol{e_i}$ = actual $\boldsymbol{y_i}$ − predicted $\boldsymbol{\hat{y}_i}$
  - This "general" idea is written formally like this $\boldsymbol{y_i \sim \mathbf{N}(\hat{y}_i, \sigma_e^2)}$: $\boldsymbol{y_i}$ is *Normally* distributed with $Conditional\ Mean = \boldsymbol{\hat{y}_i}$ and $Variance = \boldsymbol{\sigma_e^2}$
  - In addition, in the GLM, the $\boldsymbol{e_i}$ **residuals are assumed independent**, (although in many types of research designs this cannot be true)
  - Further, everyone with the same combination of x_i predictor values would have the same $\boldsymbol{\hat{y}_i}$, and the **model predicts equally well** for everyone (because there is **only one residual variance**, $\boldsymbol{\sigma_e^2}$)

General Linear Model Residual Variance

- The GLM assumes equal (constant) residual variability across all predictor values: "**homoscedasticity**" = "**homogeneity of variance**"



Otherwise, "**heteroscedasticity** " = "**heterogeneity of variance**" → model predicts differentially well across x_i (SEs will need adjusted)

"Not good" → $\boldsymbol{\sigma_e^2}$ increases as the x_i predictor increases (→ fan out)

Solution: Find a new model in which variance can differ (this **leaves GLM**)

Left image borrowed from: http://www.omidrouhani.com/research/logisticregression/html/logisticregression.htm
Right image borrowed from: https://ajh1143.github.io/HomVar/

How the Lego Blocks Fit Together

1. **Linear models** answer research questions, and are the first building block of every more complex analysis
   - *Is there an effect? Is this effect the same for everyone? Is the effect still there after considering something else?*

What other blocks you will need is determined by:

3. How your outcome is measured → **link functions**
4. Your dimensions of sampling → **random/latent effects**

- How can we add these Legos? → **2. new estimation**
  - **Least squares** is taught first, but is greatly limited in practice
  - **Maximum likelihood** picks up where least squares leaves off
  - **Bayesian** picks up where maximum likelihood gives up

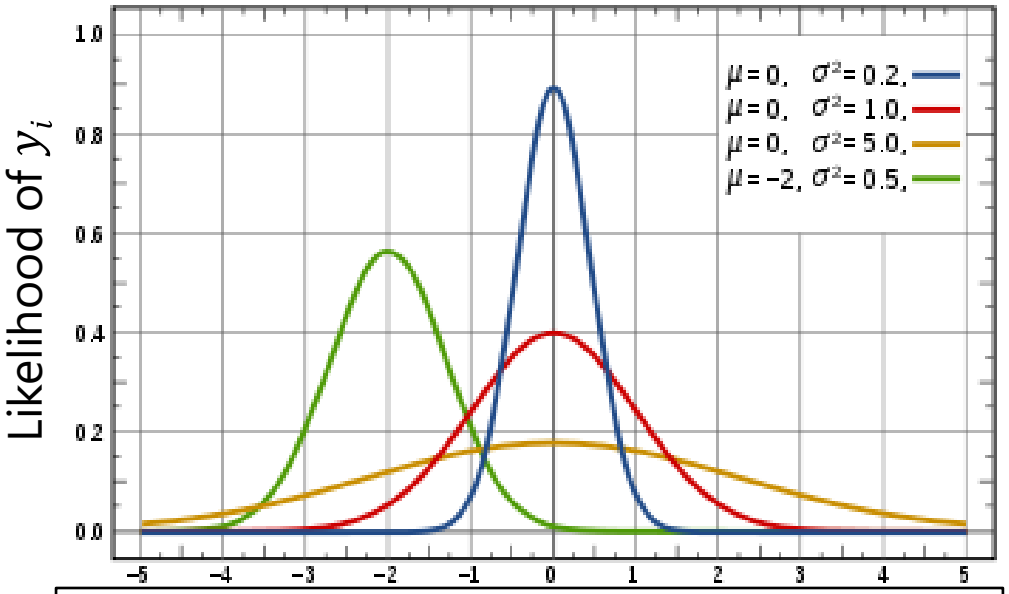
2. Estimation via Maximum Likelihood

- Ordinary Least Squares (OLS) can find answers in **some** kinds of data
  - "Best" fixed effects are those that minimize the sum of squared errors
  - How? Calculate sums of squares → mean squares → F-ratios…

- The good news: **Maximum Likelihood (ML) can find the answers** with more flexibility in **many more kinds of data**
  - Non-normal, multivariate, clustered, or incomplete data… in fact, an ML variant called *residual ML* (or *REML*) simplifies to OLS
  - Minimizing sum of squared errors = maximizing likelihood of the data
  - OLS calculations are computational shortcuts to REML (see Enders ch. 3)

- **The even better news**: If you understand **this**, then you understand the basics of ML
  - Can still work with some calculations for pedagogical purposes, though, like this…

Univariate Normal **P**robability **D**istribution **F**unction



Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}} * \exp\left[-\frac{1}{2} * \frac{(y_i - y_i)^2}{\sigma_e^2} \right]$$

Sum over persons of log of $f(y_i)$= Model Log-Likelihood (LL) → Fit Index

- This PDF tells us how **likely** (i.e., **tall**) any value of $\boldsymbol{y_i}$ is given two things:
  - Conditional mean $\boldsymbol{\hat{y}_i}$
  - Residual variance $\boldsymbol{\sigma_e^2}$
- We can see this work using the NORMDIST function in excel!
  - Easiest for **empty** model:

$$\boldsymbol{y_i = \beta_0 + e_i}$$

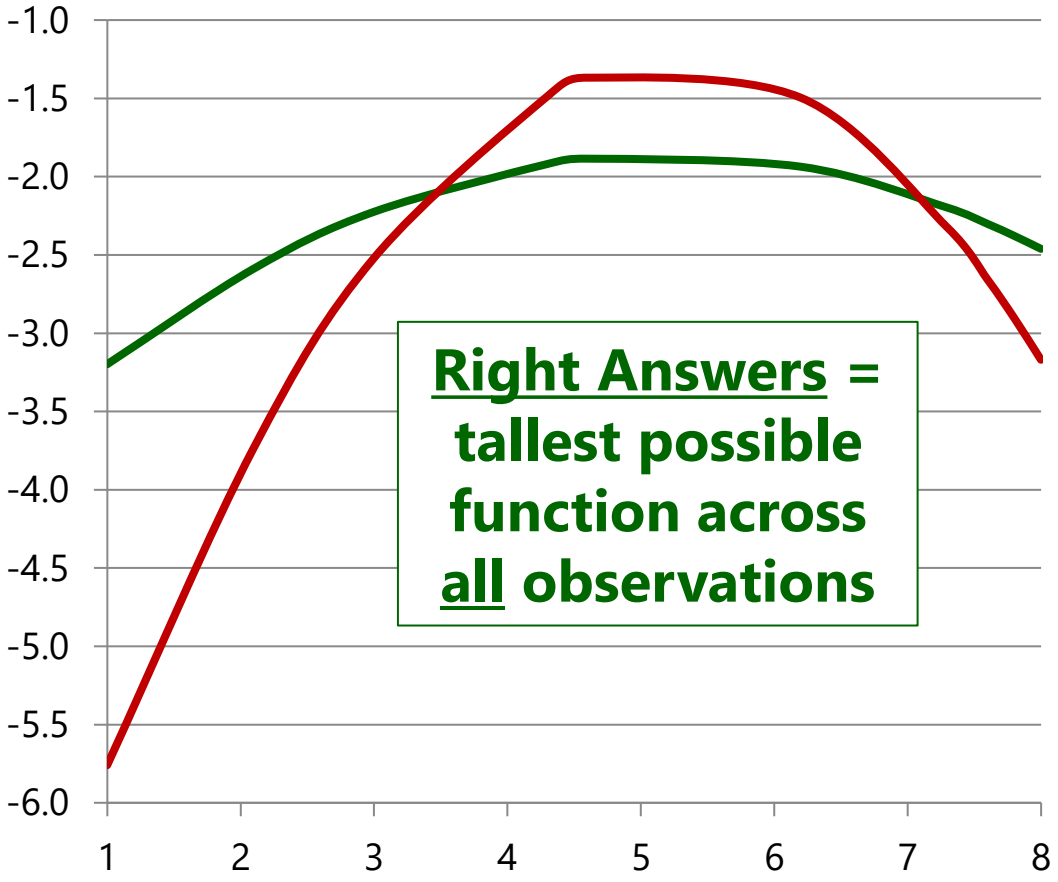
- We can check our math via software using ML!

Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

ML via Excel NORMDIST

Key idea: Normal Distribution formula → data height

Mean	**5.19**	**5.24**
Variance	**6.56**	**2.00**
	Right	**Wrong**
Outcome	**Log(Height)**	**Log(Height)**
1.0	**-3.20**	**-5.76**
2.1	**-2.59**	**-3.73**
3.0	**-2.22**	**-2.52**
4.3	**-1.92**	**-1.49**
4.6	**-1.89**	**-1.37**
6.2	**-1.94**	**-1.50**
7.3	**-2.20**	**-2.33**
7.6	**-2.30**	**-2.66**
7.8	**-2.38**	**-2.90**
8.0	**-2.46**	**-3.17**
SUM = Model LL = taller is better		
	-23.09	**-27.42**



What's so great about "normal"?

- Why must we assume "**normality**, **independence**, and **constant variance**" of residuals in General Linear Model? Because those are **required by the formula it uses** to calculate each outcome's height!
  - The normal distribution only has one variance that is shared over people
  - Summing the log-likelihood over persons implies independent values

- **The magic of ML:** if your residuals aren't normally distributed, then you can just **pick a different formula for height**, such as one that:
  - Has a better-suited probability distribution for non-normal outcomes
  - Includes a linear model for heterogeneity of variance across people
  - And/or uses a multivariate version instead for dependent outcomes

3. Then, link functions to the rescue!

- Linear models + ML + link functions = general*ized* linear models
- But first, *what other types of outcomes (and distributions) are there???*

Other Types of Outcome Variables

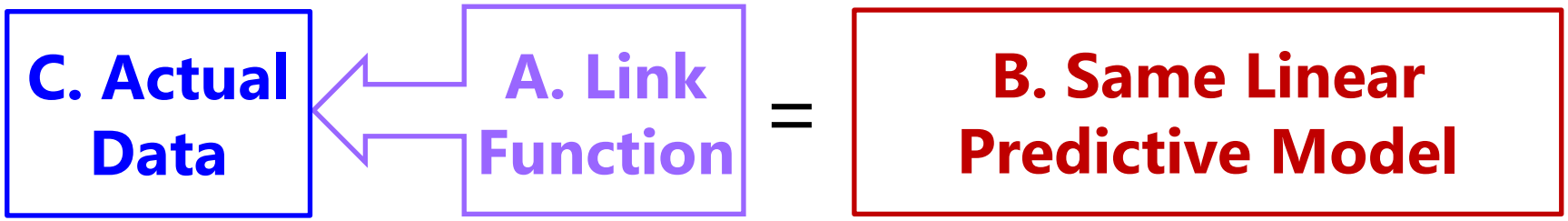
* Note: this is related to traditional levels of measurement, but I am approaching it from more of a "how-to-model them" perspective

- First, **categorical variables: *where the numbers are labels***
  - **Binary** (dichotomous) = 2 choices ("binary" means coded as 0 or 1)
    - e.g., dead or alive; finished or unfinished dissertation
  - **Nominal** = 3+ unordered choices
    - e.g., favorite type of pet, likely reaction to a situation
  - **Ordinal** = 3+ choices with some natural (undeniable) order, but the distances between the values used don't really mean anything
    - e.g., 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree
    - Equally ordinal values: 1, 20, 300, 4000
  - Synonyms for a "**categorical**" variable: discrete variable, qualitative variable, grouping variable, factor variable in R, i. variable in STATA)

Some Other Types of Outcome Variables

- Next, **quantitative variables** where the **numbers are really numbers** (interval measurement → equal distances between all possible values), but that have **one or more natural boundaries**
  - **Binomial** = # of occurrences out of known possible (**2 boundaries**)
    - e.g., # correct on a test, which is bounded by 0 and total possible
    - Correcting for different totals possible by computing proportion correct (or rate of occurrence) is still binomial (just bounded by 0 and 1 instead)
    - Scale sums with observed boundaries may also look binomial-ish
  - **Count** = # of occurrences out of unknown possible (**1 boundary**)
    - e.g., # of cigarettes smoked each day (only whole numbers used = discrete)
    - Minimum = 0, but maximum could be any positive number
    - No zeros possible? → zero-*truncated* count
    - More zeros than expected? → zero-*inflated* count ("if and how much")
  - **Censored** = Floor and/or ceiling pile-ups due to measurement limitations:
    - e.g., length of time until relapse (where some people haven't by study end)
    - Model tries to predict what would have happened without artificial boundaries

3 Parts of Generalized Linear Models



A. **Link Function**: Transformation of *conditional mean* to keep predicted outcomes within the bounds of the outcome

B. **Same Linear Model**: How the model linearly predicts the *link-transformed* conditional mean of the outcome

C. **Conditional Distribution**: How the outcome residuals could be distributed given the possible values of the outcome

Generalized linear models work for many kinds of outcomes…

Quick Example for Binary Outcomes

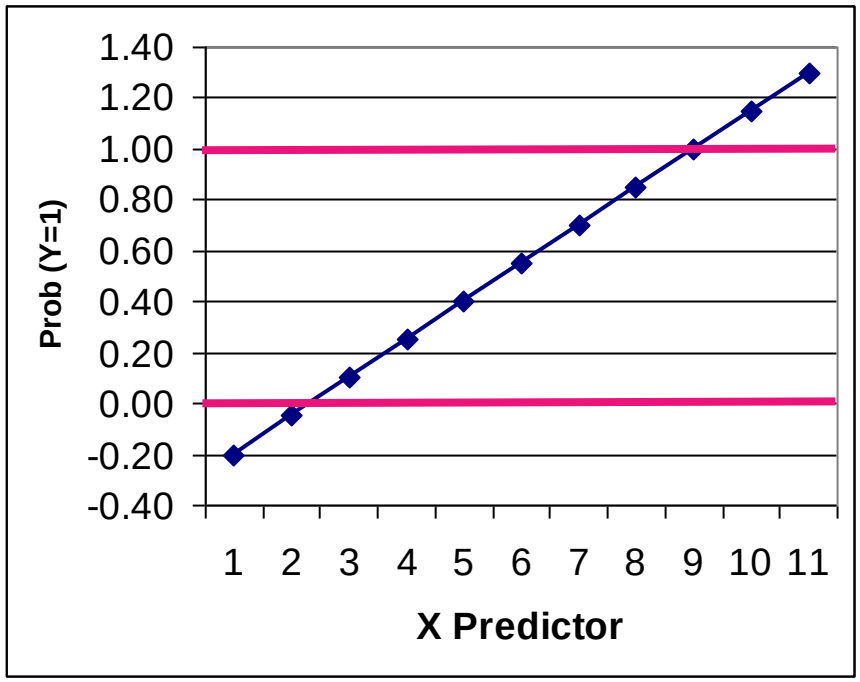
We need to go from this ***unbounded* linear model** for predicting probability…

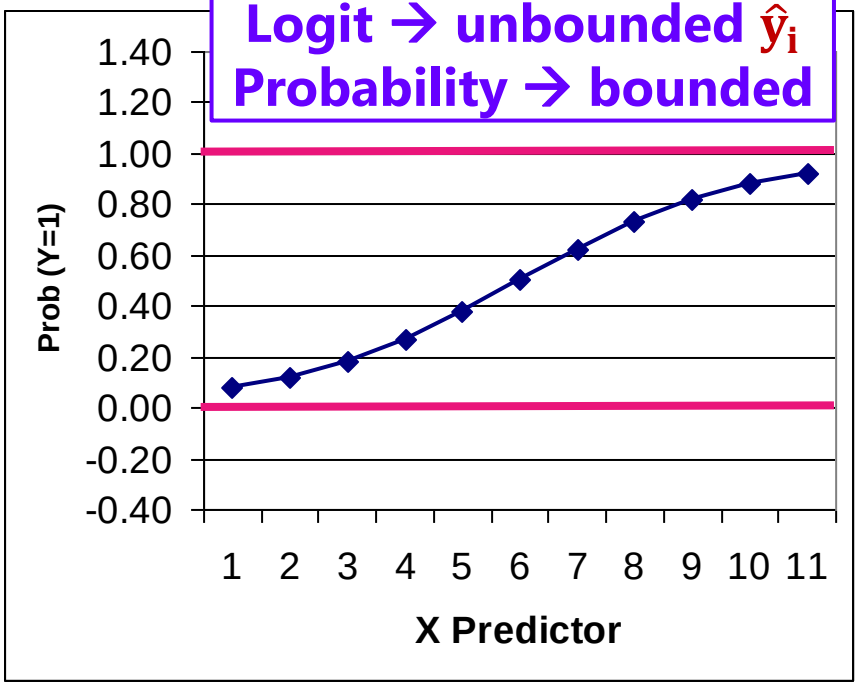
$$p(y_i = 1) = \beta_0 + \beta_1(x_i)$$

To this…

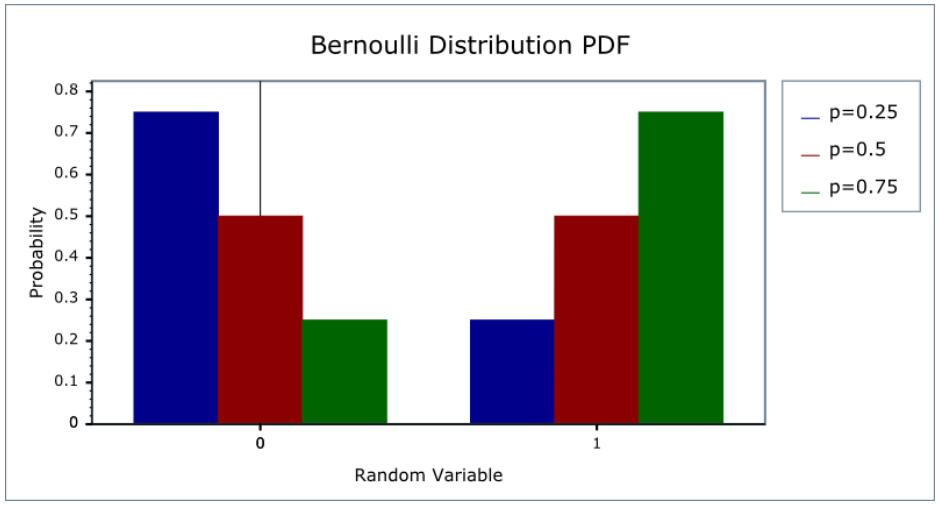
Logit Link

$$\text{Log}\left(\frac{p(y_i = 1)}{p(y_i = 0)}\right) = \beta_0 + \beta_1(x_i)$$





Bernoulli Distribution: Binary Variables



Bernoulli PDF: $f(y_i) = (p_i)^{y_i}(1-p_i)^{1-y_i}$

= $p(1)$ if 1, → p
$p(0)$ if 0 → q

= New way of getting height!

So now we assume **Bernoulli and non-constant variance** (instead of normal and homogeneous) because…

The Bernoulli distribution has only one parameter, called p, which is the mean: the proportion of 1 values (and $1 - p = q$).

The mean determines variance = $p * q$ (and skewness = $\frac{1-2p}{\sqrt{p*q}}$)

Mean and Variance of a Binary Variable

Mean (p)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

Image borrowed from: https://www.boost.org/doc/libs/1_70_0/libs/math/doc/html/math_toolkit/dist_ref/dists/bernoulli_dist.html

There's (Probably) a Model for That!

- Many kinds of **non-normal outcomes** can be analyzed with generalized models through the **magic of ML**

- **Two parts: Link function + other conditional distribution**
  - **Binary** → **Logit** + **Bernoulli**
  - **Ordinal** or **Nominal** → **Logit** + **Multinomial**
  - **Proportion** → **Logit** + **Binomial/Beta-Binomial**
  - **Count** → **Log** + **Poisson/Negative Binomial**
  - **Censored** → **Tobit** + **Normal/Bernoulli**
  - **Skewed Continuous** → **Log** + **Log-Normal/Gamma**
  - **Bimodal Continuous** → **Logit** + **Beta**
  - **Zero-Inflated** (if and how much) → **Logit/Log** + **Bernoulli/other**

B. Same Linear Predictive Model

- Your **outcome type** will often guide you towards the most useful **link function** and **conditional distribution**
- Then you can include **whatever fixed effects of predictors** best address your study design and research questions, just as in GLMs estimated using ordinary least squares, with a few small differences:
  - The **specific model names** that distinguish categorical from quantitative predictors ("ANOVA" vs "regression") **are gone**
  - Interpreted as **predicting the link-transformed** conditional mean (e.g., the logit of the probability; the log of the expected count)
  - $\boldsymbol{F}$**-tests** will show up without sums of squares and mean squares, but they are **interpreted the same way** (significance of multiple fixed slopes at once; weighted ratio of *known* to *unknown* info)
  - All parameters (fixed effects and variance-related terms) and their SEs will result from maximum likelihood estimation, but whether they are tested **using denominator DF (t vs z; F vs χ^2) will vary by software**
  - The use of conditional distributions without a separately estimated residual variance means that an **unambiguous R^2 will not be possible**

History of Generalized Linear Models

- Before ML estimation was widely available, other approaches were used to "handle" non-normality and non-constant variance; these should now all be considered as last resorts!
  - **Data transformations** (i.e., data cleaning, *shudder*)
    - e.g., positively skewed outcomes could be transformed via the square root or natural log to better approximate normality
    - e.g., an arc-sine transformation "stabilizes variance" (makes variance more constant) for proportions
    - e.g., a logit-transform creates an S-shaped curve to respect boundaries of predicted proportions in linear models
  - **Nonparametric statistics**: most are less flexible than generalized models because they still require some kind of least squares estimation
    - e.g., they may require rank transformations first (as in Spearman correlation) that throw away information about absolute distances
    - e.g., the same type of non-normal distribution must hold across groups

From Univariate to Multivariate Models

- This course will begin with prediction using generalized linear models of all kinds of outcomes, one at a time, but many types of data and/or research questions require **multivariate models**:
- **When y_i is still a single outcome conceptually, but:**
  - You have more than one outcome per person created by multiple conditions (e.g., longitudinal or repeated measures designs)
  - When your outcome is measured multiple times for a pair or group with distinguishable members (e.g., dyadic or family data)
- **When your hypotheses involve more than one y_i outcome:**
  - To compare predictor effect sizes across outcomes (e.g., is the treatment effect bigger on outcome A than outcome B?)
  - You want to test indirect effects among them (i.e., mediation), such that a single variable is both a predictor and an outcome

From Univariate to Multivariate

- Ordinary least squares (OLS) has a "closed form" solution (its "sums of squares" formulae) when used for GLM for single outcomes
- For GLM for multiple outcomes, **OLS quickly becomes useless**…
  - Does not allow any missing outcomes (listwise-deletes entire person)
  - Only two options for modeling residual correlation between outcomes
  - Requires balanced data (same number of outcomes per higher unit)

- We will continue using maximum likelihood (ML) estimation for **multivariate models** to solve these problems, but some multivariate model variants will **require a switch in software**
  - Models in which all variables are either predictors or outcomes can be done by tricking univariate (regression-type) ML software (e.g., MIXED)
  - Otherwise, models must be estimated in ML using "truly" multivariate software (such as is used in path analysis or latent variable modeling)
    - Fewer choices for generalized linear models across software packages

This Course and Beyond: Lego #4

- **This course** will help you understand how to **combine**:
  - (1) **linear models** and (3) **link functions** to predict any kind of outcome, which is possible through (2) the use of **ML estimation** (btw, you will also learn how ML is used to assess model fit in multivariate models)

- Conquering this material serves two distinct purposes:
  - Being able to **predict any kind of outcome** in order to test univariate or multivariate hypotheses **is useful** in and of itself!
  - In addition, these are all **essential pre-requisite** skills for classes that also include **Lego #4: random effects and/or latent variables**
    - Multilevel (*aka,* mixed-effects) models (MLM), factor analysis (FA), structural equation models (SEM), item response theory (IRT)…
    - PSQF 6271 (longitudinal) and 6272 (clustered) MLM; PSQF 6262 (IRT) and PSQF 6249 (FA and SEM)

- So please stick with us—I hope you won't regret it! ☺